

frontiers of social psychology

Measurement in Social Psychology

Edited by
Hart Blanton, Jessica M. LaCroix,
and Gregory D. Webster

A **Psychology Press** Book



MEASUREMENT IN SOCIAL PSYCHOLOGY

Although best known for experimental methods, social psychology also has a strong tradition of measurement. This volume seeks to highlight this tradition by introducing readers to measurement strategies that help drive social psychological research and theory development.

The book opens with an analysis of the measurement technique that dominates most of the social sciences, self-report. Chapter 1 presents a conceptual framework for interpreting the data generated from self-report, which it uses to provide practical advice on writing strong and structured self-report items. From there, attention is drawn to the many other innovative measurement and data-collection techniques that have helped expand the range of theories social psychologists test. Chapters 2 through 6 introduce techniques designed to measure the internal psychological states of individual respondents, with strategies that can stand alone or complement anything obtained via self-report. Included are chapters on implicit, elicitation, and diary approaches to collecting response data from participants, as well as neurological and psychobiological approaches to inferring underlying mechanisms. The remaining chapters introduce creative data-collection techniques, with particular attention given to the rich forms of data humans often leave behind. Also included are chapters on textual analysis, archival analysis, geocoding, and social media harvesting.

The many methods covered in this book complement one another, such that the full volume provides researchers with a powerful toolset to help them better explore what is “social” about human behavior. This is fascinating reading for students and researchers in social psychology.

Hart Blanton is Professor of Communication at Texas A&M University. He conducts research in the areas of social influence, health communication, and research methodology.

Jessica M. LaCroix is Research Assistant Professor at the Uniformed Services University of the Health Sciences and specializes in health psychology, research methodology, and military suicide prevention.

Gregory D. Webster is Associate Professor of Social Psychology at University of Florida with graduate degrees from the College of William & Mary and the University of Colorado.

Frontiers of Social Psychology

Series Editors:

Arie W. Kruglanski

University of Maryland at College Park

Joseph P. Forgas

University of New South Wales

Frontiers of Social Psychology is a series of domain-specific handbooks. Each volume provides readers with an overview of the most recent theoretical, methodological, and practical developments in a substantive area of social psychology, in greater depth than is possible in general social psychology handbooks. The editors and contributors are all internationally renowned scholars whose work is at the cutting edge of research.

Scholarly, yet accessible, the volumes in the *Frontiers* series are an essential resource for senior undergraduates, postgraduates, researchers, and practitioners and are suitable as texts in advanced courses in specific subareas of social psychology.

Published Titles

Intergroup Conflicts and their Resolution

Bar-Tal

Social Motivation

Dunning

Social Cognition

Strack & Förster

Social Psychology of Consumer Behavior

Wänke

For continually updated information about published and forthcoming titles in the *Frontiers of Social Psychology* series, please visit: <https://www.routledge.com/psychology/series/FSP>

MEASUREMENT IN SOCIAL PSYCHOLOGY

*Edited by Hart Blanton, Jessica M. LaCroix,
and Gregory D. Webster*

First published 2019
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2019 Taylor & Francis

The right of Hart Blanton, Jessica M. LaCroix, and Gregory D. Webster to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
A catalog record for this title has been requested

ISBN: 978-1-138-91323-3 (hbk)
ISBN: 978-1-138-91324-0 (pbk)
ISBN: 978-0-429-45292-5 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

CONTENTS

1	From Principles to Measurement: Theory-Based Tips on Writing Better Questions	1
	<i>Hart Blanton and James Jaccard</i>	
2	Implicit Measures: Procedures, Use, and Interpretation	29
	<i>Bertram Gawronski and Adam Hahn</i>	
3	Elicitation Research	56
	<i>William A. Fisher, Jeffrey D. Fisher, and Katrina Aberizk</i>	
4	Psychobiological Measurement	75
	<i>Peggy M. Zoccola</i>	
5	It's About Time: Event-Related Brain Potentials and the Temporal Parameters of Mental Events	102
	<i>Meredith P. Levens, Hannah I. Volpert-Esmond, and Bruce D. Bartholow</i>	
6	Using Daily Diary Methods to Inform and Enrich Social Psychological Research	127
	<i>Marcella H. Boynton and Ross E. O'Hara</i>	
7	Textual Analysis	153
	<i>Cindy K. Chung and James W. Pennebaker</i>	

vi Contents

8	Data to Die For: Archival Research <i>Brett W. Pelham</i>	174
9	Geocoding: Using Space to Enhance Social Psychological Research <i>Natasza Marrouch and Blair T. Johnson</i>	201
10	Social Media Harvesting <i>Man-pui Sally Chan, Alex Morales, Mohsen Farhadloo, Ryan Joseph Palmer, and Dolores Albarracín</i>	228
	<i>Index</i>	265

1

FROM PRINCIPLES TO MEASUREMENT

Theory-Based Tips on Writing Better Questions

Hart Blanton and James Jaccard

Self-reports are the dominant assessment method in the social sciences and a large part of their appeal is the ease with which questions can be generated and administered. In our view, however, this apparent ease obscures the care that is needed to produce questions that generate meaningful data. In this chapter, we introduce and review basic principles of measurement, which we then use as a foundation to offer specific advice (“tips”) on how to write more effective questions.

Principles of Measurement

A Measurement Model

Suppose a researcher wanted to measure consumers’ judgments of the quality of a product. Perceptions of product quality cannot be observed directly—perceived quality is a latent, theoretical psychological construct, assumed to be continuous in character, such that it can only be inferred indirectly through observable actions. One such action can be ratings a consumer makes on a rating scale. Suppose consumers are asked to rate the perceived quality of a product on a scale that ranges from 0 (“very low quality”) to 6 (“very high quality”). By seeking to quantify product perceptions in this manner—and whether the researcher has realized it or not—a formal measurement model has been embraced. This model is depicted in Figure 1.1.

The rectangle labeled “Q” in Figure 1.1 represents the rating on the 0-to-6 scale. This rating does not, by fiat, reveal “true” quality perceptions of the respondent, which is conceptualized as an unobservable latent construct and represented in Figure 1.1 by the circle with the word “quality” in it. The researcher assumes that the observed “Q” is influenced by true, latent quality perceptions, but that

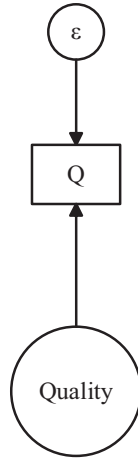


FIGURE 1.1 Measurement Model

the correspondence between latent and observed constructs is less than perfect. Ratings on Q are thus a function of both the consumers' true evaluations and measurement error (represented as " ε " in Figure 1.1). This can be expressed algebraically in the form of a linear model:

$$Q = \alpha + \lambda \text{Quality} + \varepsilon \quad [1]$$

where α is an intercept, λ is a regression coefficient (also frequently called a *loading*), and ε is measurement error. When the relationship is linear, as assumed in Equation 1, then Q is an interval-level measure of the latent construct of perceived quality. If the relationship is non-linear but monotonic, Q is an ordinal measure of the latent construct. Articulation of this formal model focuses attention on one of the primary challenges facing researchers who wish to create self-report questions—the need to reduce the influence of error on observed ratings. We next consider two sources of error, random and systematic, as well as their implications for characterizing the reliability and validity of self-report items.

Random Error and Reliability

Random error represents random influences, known or unknown, that arbitrarily bias numeric self-reports upward or downward. Often referred to as "noise," random error can be generated by such factors as momentary distractions, fluke misunderstandings, transient moods, and so on. This form of error is commonplace, but its relative magnitude can vary considerably from one question to the next. As such, it is meaningful to think about the degree to which a given question

or set of questions is susceptible to random error. This represents the concept of *reliability*.

The reliability of observed scores conveys the extent to which they are free of random error. Statistically, a reliability estimate communicates the percentage of variance in the observed scores that is due to unknown, random influences as opposed to systematic influences. Thus, if the reliability of a set of scores is 0.80, then 80% of their variation is systematic and 20% is random. The presence of random error in measures can bias statistical parameter estimates, potentially attenuating correlations and causing researchers to think they have sufficiently controlled for constructs in an analysis, when they have not.

Systematic Error and Validity

Another form of measurement error is called *systematic error*. This source of error often introduces variance into observed self-report items that is non-random; i.e., that is a function of one or more psychological constructs that are something different than the construct of interest. Consider the model in Figure 1.2. Here a researcher hopes to measure both drug use and grade-point average (GPA) via self-report. Each of these constructs are influenced by the true latent constructs that are of interest (as in Figure 1.2), but another latent construct is also exerting influence on the two measures, social desirability.

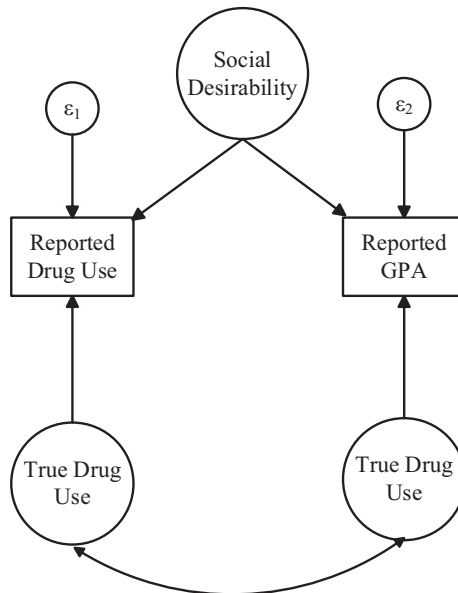


FIGURE 1.2 Example of Systematic Error

The dynamic in Figure 1.2 can arise if those most concerned with projecting a positive image are under reporting their true drug use and over reporting their true GPA. The systematic influence of this “third variable,” social desirability, might cause a researcher to overestimate (or underestimate) the strength of the relationship between drug use and academic performance.

Systematic error of the type in Figure 1.2 is a threat to the *validity* of a measure. Ratings on a self-report are valid to the extent that they accurately reflect the construct that is of interest, as opposed to constructs that are not of interest. In the above example, the two measures are partly influenced by the constructs that were of interest (drug use and GPA) but they also are partly influenced by a construct that was not (social desirability), and so the validity of these measures was undermined. In more extreme cases, a measure might be so strongly biased by systematic, confounding influences that it is best viewed as invalid; i.e., it should be viewed as a measure assessing something other than the construct of interest.

Statistical and Methodological Approaches to Measurement Error

One way of handling the presence of measurement error is to embrace modern analytic methods that can correct for biasing influences. Structural equation modeling (SEM) is a particularly useful analytic tool, well-suited to estimating statistical parameters while adjusting for both systematic and random sources of measurement error (Kline, 2016). Incorporated into these analytic approaches might also be attempts to formally measure known or anticipated sources of systematic error (“confounds”), so that their influences can be statistically controlled (or “covaried”). For instance, if a researcher has the concern that social desirability tendencies will influence ratings, a separate measure socially desirability can be administered (e.g., Fleming, 2012; Uziel, 2010), so that its influence on ratings can be formally estimated and statistically controlled during parameter estimation.

The Aggregation Approach to Measurement Error

A common approach to reducing the impact of random measurement error is aggregation. Because one can rarely expect to create a single perfect self-report item that captures all of the variance in a complex psychological construct, researchers often construct multi-item instruments to measure constructs. The logic of aggregation is that, even if a given item is influenced to a consequential degree by random error, different items will be influenced in different and largely idiosyncratic ways. The result is that when ratings on multiple items are summed or averaged, the error in specific items will “wash out” in the aggregate score, resulting in a more reliable and valid estimate of the latent construct of interest. This is a generally sound and accepted practice but there are common misperceptions and misapplications of aggregation. We explore these shortly, but we first

consider the different types of constructs one might wish to assess, and how this might affect aggregation.

Defining the Construct of Interest

Broad versus Narrow Constructs

Some constructs are fairly concrete and easily referenced in self-report items (e.g., age, height, income). With such “narrow” constructs, there will often be little gained from constructing multiple items to represent them and then aggregating the questions because they will yield identical information about a respondent (e.g., “How old are you in years and months?” “How old are you in months?”). In contrast, many concepts that are of basic and applied interest in psychology are by their nature abstract and hard to translate into a single question (e.g., intelligence, depression, social support). With such “broad” or abstract constructs, aggregation can have value, as multiple questions give respondents imperfect but non-redundant ways of expressing their standing.

Breadth versus Dimensionality

A construct might be broad in multiple senses. One way is that it can take a myriad of roughly equivalent, interrelated forms, such that a larger number of items might capture more of distinct ways it can manifest itself, leading to improved measurement. Consider extraversion. Highly extraverted people might evidence this quality by seeking to interact with new people, by seeking to have many friends, through their comfort speaking in groups, by their willingness to tell jokes, and so on. Our understanding of this construct is simply too big to be captured by any single item. That said, extraversion needn’t necessarily be expressed by any one of these specific behaviors. Some extraverts are known for their joy of speaking in public and others for their love of telling jokes. When aggregating across these and many other distinct expressions, a general tendency to be extraverted can emerge in an aggregate scale total, resulting in a meaningful unitary score that captures relative standing on this broad dimension.

A second way in which a construct might be broad is that it might be multidimensional, in that it is made up of interrelated but distinct facets. As examples, the construct of depression is often thought to be represented by four different (and also broad) facets: a cognitive dimension, an affective dimension, a somatic dimension, and an apathy dimension. Anxiety is thought to have three facets: social anxiety, generalized anxiety, and panic-related anxiety. Social support is thought to have three facets: tangible support, emotional support, and informational support.

To measure a broad construct, it is thus incumbent on researchers to clearly define it, specifying its dimensional structure based on theory or on past research. In the case of extraversion, where a researcher assumes a broad but unidimensional

attribute, the goal in generating items will be to approximate a selection of items, drawn from a theoretical and infinitely large pool of equally good items, each of which is influenced by a person's true extraversion (in a manner consistent with Equation 1). In contrast, in the case of depression, the goal will be to first define four facets of depression, and to repeat this same process of item generation four different times. In truth, whether pursuing items to capture a broad unidimensional construct or multiple broad dimensions of a multidimensional construct, some items will almost assuredly be better than others (as expressed by the relative size of λ and ϵ in Equation 1). However, through the creation of multiple imperfect items that vary in their quality, the resulting aggregate score can produce an observed estimate that is far better than can be generated by the pursuit of the single "best" self-report item.

An Iterative Process

How successful one will be at generating sets of questions that combine to estimate a broad construct should be viewed as an empirical question, one that often can be evaluated through reference to the results of analyses performed on the test items themselves. It is beyond the scope of this chapter to detail this process other than to note that scale construction is often an iterative process, one in which many potential items are generated, the "bad" items are revised or rejected, and the "good" items are retained. In the course of evaluating items, assumptions about the dimensionality of the construct should be scrutinized and perhaps revised, in light of empirical results. A construct that was first conceptualized as unidimensional might through trial and error reveal itself to be multidimensional, and vice versa. There are many useful texts to offer guidance on this iterative process and the standards one should apply to reevaluating measurement assumptions (see Furr & Bacharach, 2018; Nunally & Bernstein, 2004). Rather than review this well-covered material, we seek in the following sections to point to some of the more common misperceptions surrounding multi-item scales.

Internal Consistency versus Homogeneity

One common source of confusion is the distinction between the *internal consistency* of a multi-item scale and its degree of *homogeneity*. Internal consistency refers to the degree of interrelatedness of items, whereas homogeneity refers to their dimensionality or the extent to which the covariance structure among items can be accounted for by a single latent factor. These properties are not isomorphic. For example, if 10 items are all intercorrelated at $r = 0.20$, the correlational pattern among them can be accounted for by a single latent variable (i.e., they are unidimensional), but their internal consistency is relatively modest. As the intercorrelation between items increases, so too will the internal consistency, everything else being equal. Coefficient alpha is a common index thought to reflect the internal consistency of items, the homogeneity of items, or both. However,

despite the isomorphism this example highlights, reliability estimates do not make for good homogeneity estimates. Consider the correlation patterns for two six-item scales, each with an alpha of 0.86:

<i>Item</i>	1	2	3	4	5	6	1	2	3	4	5	6
1	—						—					
2	.8	—					.5	—				
3	.8	.8	—				.5	.5	—			
4	.3	.3	.3	—			.5	.5	.5	—		
5	.3	.3	.3	.8	—		.5	.5	.5	.5	—	
6	.3	.3	.3	.8	.8	—	.5	.5	.5	.5	.5	—

The scale on the left clearly is not unidimensional despite its large coefficient alpha. The scale on the right is unidimensional. To determine unidimensionality, one should assess it directly and not infer it from reliability (see Cortina, 1993).

Assessing Unidimensionality

So how is homogeneity to be assessed? One useful strategy is to conduct a confirmatory factor analysis on items. If a one-factor model fits the data well, then one can assume unidimensionality. A common practice after a factor analysis of items (be it confirmatory or exploratory) is to select only items that load on the same factor and the use these as the core items for aggregation in a final scale. Unfortunately, there are no clear standards for what constitutes a large enough factor loading for an item to be said to adequately represent the underlying factor. Loadings in the 0.30 to 0.40 range are often suggested, but closer inspection of what these values mean suggest that one might want higher standards. For example, in a traditional confirmatory factor analysis, the square of a standardized factor loading is the proportion of variation in an indicator that is due to the underlying factor, and one minus this value is the proportion of unique variance associated with the indicator. A factor loading of 0.50, for example, implies that just 25% of the variation in the indicator is due to the underlying factor whereas 75% of its variation is unique and has nothing to do with the factor. With this in mind, suppose a researcher created a four-item scale measuring perceived stigma of having a mental health problem, finding that all four items load on a single factor as follows:

<i>Item</i>	<i>Loading</i>
Sometimes I am talked down to because of my mental health problems	0.60
I believe I would be discriminated against by my employers because of my mental health problems	0.50
I would have had better chances in life if I had not had a mental illness	0.52
People's reactions to my mental health problems make me keep to myself	0.55

A global index of perceived stigma can be obtained by aggregating across the four items but as a result, attention is drawn away from the unique variance of each item—even though each item is dominated by unique variance. Perhaps this unique variance is most relevant to predicting an outcome rather than the common variance among the items. Suppose that a researcher wished to determine the extent to which perceived stigma predicts discrimination in an employment setting. The second item has the lowest loading, 0.50, and this means that it has about 75% unique variance relative to the underlying generalized stigma factor. However, this item is the only item focused on perceptions of stigma in employment settings. As a general rule, the accuracy of prediction generally will increase to the extent features of the judgment closely correspond to features of the criterion one wishes to predict (for review, see Fishbein & Ajzen, 2010). Perhaps as a result, this particular researcher—given the nature of the research question at hand—should focus attention on this one item, not the scale total. There are many other contexts where one might not want to be too quick to focus exclusively on common variance but instead work with *both* the common and unique sources of variance. We caution researchers to consider both the ways that aggregate estimates of broad concepts (pro-environmentalism) might predict broad behavioral outcomes (e.g., carbon footprint), and how they might be separated out into more narrow constructs (e.g., aluminum recycling attitudes) to predict specific behavioral tendencies (e.g., aluminum recycling; see Davidson & Jaccard, 1979; Kallgren & Wood, 1986).

Assessing the Reliability of a Composite Through Item Analysis

As noted, items are often aggregated to capitalize on the fact that random error in individual items will tend to cancel out, yielding a more reliable composite. We often want to estimate the reliability of a composite, with coefficient alpha being the most frequently used index for doing so. However, psychometricians argue against its use (Sijtsma, 2009), and recommend an alternative index that makes fewer assumptions, called *composite reliability*. Both composite reliability and coefficient alpha assume unidimensionality, but only coefficient alpha also assumes (a) that the factor loadings for items are all equal and (b) there is no correlation between any of the errors of individual items. Such assumptions are often violated and, as such, composite reliabilities are generally preferred to coefficient alphas as an index of the reliability of a scale composite (see Raykov, 2001). By the same token, item elimination from a scale based on the value of the “coefficient alpha if item is eliminated” has been shown to be flawed and is better approached in terms of how the composite reliability is affected if a given item is eliminated (Raykov & Marcaloudis, 2015).

Making Your Scaling Function Explicit

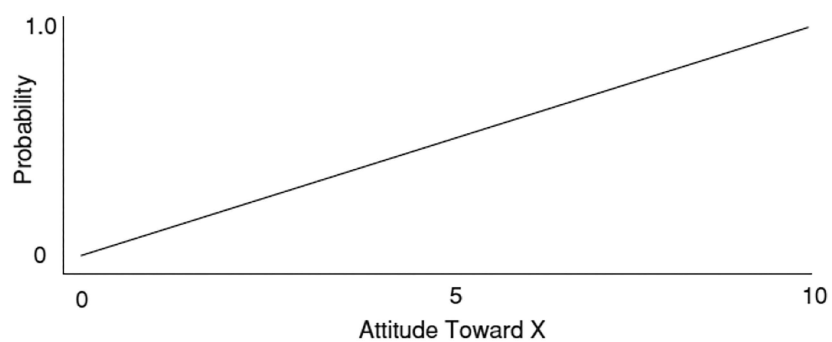
Equation 1 presented the operative measurement model used throughout this chapter. Although a linear function between a measure and a latent construct is

by far the most common function assumed by researchers, it is just one of many possible functions that can be operating. This is particularly important to keep in mind when conducting item analyses for multi-item scales. In traditional scaling models, the researcher assumes that a given response to an item is generated as a linear function of the latent construct (per Equation 1). However, psychometricians have elaborated other functions that have implications for how one writes items. To this end, it is useful to consider a construct central to psychometrics, *item-operating characteristic* (IOC). An IOC specifies the relationship between true score on the construct of interest and how the probability of endorsement of an item changes as the true score increases (see Green, 1954). Consider as an example a researcher interested in measuring someone's attitude towards a given attitude object, X. There exist a number of plausible IOCs for measures designed with this purpose. One type of IOC derives from the logic of Thurstone scaling and states that the probability of endorsing an item should be highest for an individual whose attitude toward X matches the "scale value" of the item with respect to X. For example, an individual with a neutral attitude toward X should be most likely to endorse an item that conveys neutrality with respect to X; an individual with a moderately positive attitude toward X should be most likely to endorse items that express moderately positive favorability towards X; and a person with an extremely unfavorable attitude toward X should be most likely to endorse items that express extreme unfavorability towards X. The more discrepant an individual's attitude is from the particular scale value of the item, in either a positive or a negative direction, the less likely the individual should be to endorse the item.

Figure 1.3 presents the IOCs based on this logic for three items that differ in their scale values. The scale values, in principle, vary from 0 to 10, with higher scores indicating higher degrees of favorability and 5 representing a neutral point. The first item in Figure 1.3 has an extreme positive scale value (of 10), and it can be seen that the IOC for this behavior is linear in form: The more positive the person's attitude towards X, the more likely the person will be to endorse the item. Consider the second item. This item has a scale value of 5, which represents neutral affect. In this case, individuals with neutral attitudes are most likely to endorse the item and the probability of endorsement decreases as one's attitude becomes more negative or more positive. This IOC is curvilinear in form and one would expect a low correlation between item endorsement and a person's attitude, because a correlation coefficient is primarily sensitive to linear relationships. Thus, using Thurstonian logic, one cannot identify "good" items purely by examining item-total correlations. Rather, one needs to use analytic strategies that allow for non-linearity in the probability of endorsement of an item and the total score, depending on the item's scale value.

An alternative conceptualization of the IOC derives from the basic logic of Guttman scaling (Edwards, 1957). Guttman assumed step-shaped IOCs: If an individual's attitude is less favorable than the degree of favorability implied by an item (i.e., its scale value), then the probability of endorsing the item is zero. However, if the individual's attitude is as favorable or more favorable than the scale value of

(a) Scale Value of 10



(b) Scale Value of 5



(c) Scale Value of 7

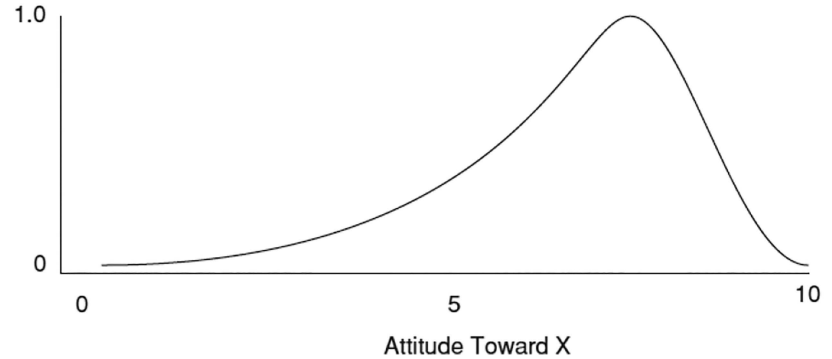


FIGURE 1.3 IOCs for Thurstone Scaling

the item, the probability of endorsement is 1.0. Figure 1.4 presents IOCs for the same three items using Guttman's logic (see Edwards, 1957, for elaboration of this rationale). Again, item-total correlations will not be helpful in identifying strong items under this form of measurement model. Rather, we require analytic strategies that are sensitive to step-shaped functions.

The general point is that the way we write items and the analyses we use to identify strong items for a multi-item scale are highly dependent on the measurement model we assume and the presumed item-operating characteristics for that scale. A measurement model that assumes simple linear IOCs (which is typical of Likert scaling) is but one model that can be adopted. It is important to be explicit about the measurement model one seeks to use.

Writing Self-Report Items

In this next section, we translate the measurement principles discussed above to provide concrete advice about writing self-report items and measures. The first step in generating questions is to step back and define the construct one wishes to measure. We start there and move through a wide range of tips to writing stronger questions for quantitative analyses.

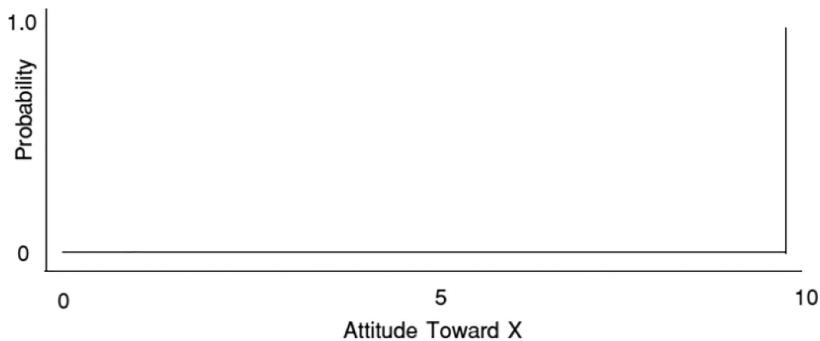
Defining the Scope of the Construct

If a construct is conceived as narrow in scope, then a single, straightforward question might be sufficient to produce a sufficiently valid and reliable estimate. When trying to estimate the intention to vote for Candidate X, for instance, a single rating scale measuring perceived likelihood of voting for Candidate X will cover a lot of ground. If multiple items are attempted (e.g., *intention* to vote, *willingness* to vote, and *expectation* of voting), the inter-correlations will likely be so high that little information is gained, although the cancelation of random errors may increase measure reliability.

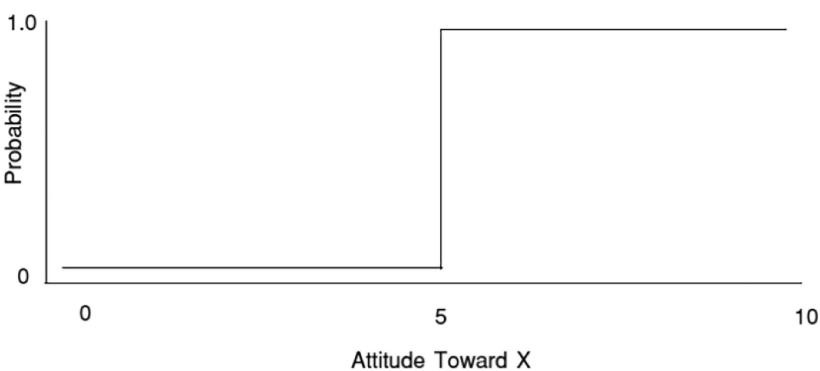
In contrast, if the construct is conceived as broad and manifested in many ways, greater thought must be given to the nature and types of items needed to fully sample the construct universe of interest. Is the construct broad but unidimensional? If so, unidimensionality should be a priority when generating questions that might load on a single factor. With a broad, unidimensional construct, one should be to produce items that sample liberally from a larger pool of potential expressions. Although random (and systematic) sources of error might affect each individual item to some extent, potentially resulting in lower inter-item correlations, higher reliability can be produced through aggregation.

In the process of generating items to assess broad constructs, however, one should give consideration to the unique variance introduced by specific items and whether any given item taps unique facets of the construct that have value in their own right, as stand-alone items. When generating items to estimate a person's overall

(a) Scale Value of 10



(b) Scale Value of 5



(c) Scale Value of 7

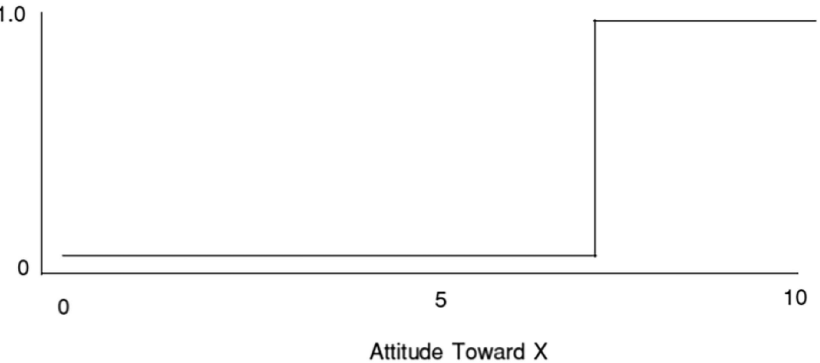


FIGURE 1.4 IOCs for Guttman Scaling

level of extraversion, for instance, a researcher might see different applications for items that tap sociability (e.g., being outgoing) social assertiveness (e.g., likes taking charge), each of which is an expression of extraversion. As attention turns from common to unique item variance, the researcher might consider if there is cause to generate sets of multiple, interchangeable items taping different distinct types expressions of the original construct (see, for instance, Soto & John, 2017).

Articulating the Item-Operating Characteristic

We noted earlier that the traditional and most common approach to measurement is to assume linear IOCs for items comprising a multi-item scale. In such cases, one should write questions with an eye for generating items that will have high inter-correlations and high item-total correlations. Consider for instance a researcher interested in measuring attitudes towards smoking marijuana. Respondents are asked to rate their endorsement of the statement “I can be friends with a person who has smoked marijuana in the past” but this might be problematic. Most anyone with even the most modest of pro-marijuana attitudes will endorse this statement highly, restricting the range of responses and yielding a data pattern that is at odds with a linear IOC. When working with non-linear IOCs, as much as possible, items should be constructed such that across the full set of items, they capture incremental, linear movement along the full range of potential scale values that might occur along the theoretical metric of interest. One should avoid ending up with a scale where the items, as a collective, truncate the range of scale values.

Tip 1: Consider a “Linear Wording” Approach to Asking Questions

When working with linear IOCs (as is typically the case), one should generate items whose probability of endorsement will clearly vary linearly as the underlying construct changes. Be particularly aware of base rate issues surrounding ceiling and floor effects. Suppose, for instance, a researcher wishes to assess attitudes towards getting pregnant among high school seniors. An item like “Getting pregnant now would be bad” would probably be of limited use, because almost all high school girls will agree strongly with the statement; it will not discriminate those with highly negative attitudes from those with moderately negative attitudes. However, this item can be modified to read, “Getting pregnant now would be one of the worst things that could happen to me.” This likely would avoid the ceiling effect that the prior version of the item exhibits. Even a subtle shift in phrasing from “I was sad last week” versus “I was very sad last week” can affect response distributions for items in ways that improve the range of responses one obtains as a function of the underlying construct.

In some cases, one also can adopt a strategy of simply asking people directly how they stand on the construct in question. To sort individuals in terms of their

attitudes towards smoking marijuana, for instance, one might simply ask respondents “how do you feel about smoking marijuana?” where answers are made on a numeric scale that ranges on a dimension from negative to positive, with a neutral midpoint. Or, one might ask respondents to complete the phrase “I feel _____ about smoking marijuana,” by choosing from a set of options that range from “extremely negative” to “extremely positive” (and see below where we list rating scale options to use in such a scenario). Such questions can be framed in a way that they reflect a linear relationship between responses to the item and the underlying latent construct. It is not always possible to articulate a construct in this manner, especially for constructs that are broad and require assessments of a wide range of interrelated manifestations. But, in many instances, a good way to obtain an index of a latent construct is to clearly articulate to the respondent the type of judgment (or IOC) that is desired and then to ask them to provide a rating accordingly.

Reducing Random Error

Random error is an unfortunate fact of life and researchers should expect it to influence responses to some degree. However, there are many ways to minimize it through the design of questions and in this section, we provide tips that might help.

Tip 2: Keep Items Short, Simple, and Understandable

The more cognitively demanding a question, the greater likelihood that transient differences in motivation, attention, and interest will affect responding. As much as possible, avoid long sentences, large or obscure words, complex phrasing, and unnecessary words. In most common instances, try to keep the reading level to about the fourth or fifth grade.

Tip 3: Make Sure the Item Measures Only One Concept

Items that are open to multiple interpretations will be more prone to error. An item like “My therapist was expert and sincere” is double-barreled and thus inherently ambiguous. Respondents who view their therapist as expert but not sincere have no valid response. Similarly, an item like “I intend to go to my appointment because it will help me get better” might be difficult for respondents who intend to attend their appointment, despite holding doubt it will help.

Tip 4: Avoid Negations, Particularly Double Negations

Inclusion of a negation in a question can be confusing, particularly if the item offers an opportunity to reject the original negation. For instance, respondents

who like to dance might fail to notice or skim over the word “not” when asked to “agree or disagree” with the statement, “I do not like to dance.” Problems only increase if a question is a double negation. An item like “Students should not fail to go to school” can be cognitively demanding (especially for respondents given the option to “disagree” with the statement). As a general rule, it is best to avoid the word “not” altogether, as people often misinterpret or fail to notice the negation and misreport.

Tip 5: Look for and Remove Potentially Ambiguous Terms and Phrases

One problem where some psychologists have difficulty is with the use of jargon. We become so fluent as “psychologizers,” that we forget how confusing we can be to many of the non-psychologists we wish to study. It would probably be a bad idea, for instance, to ask respondents how “reactant” they felt while reading a health message. Lack of understanding can be subtle, however, and can also arise from familiar, non-jargon words. Even a simple item, like, “I smoked marijuana last month” can introduce ambiguities, because some respondents will interpret the “last month” as some time in the last 30 days, whereas others will interpret it based on a calendar month. It is an unfortunate fact that some ambiguities only become obvious to researchers after the data have been collected but someone good at writing questions will put considerable energy into identifying any potential source of confusion a priori.

Tip 6: Personalize the Item and Provide Contextual Information and Time Frames

If not made explicit in a question, respondents will often impute their own time frames and other contextual information into questions, leading to item unreliability. The item “Joining a gang would be good” can elicit a very different response than the item “For me, joining a gang in my neighborhood at this time would be good.” The first statement not only fails to indicate a time period, it fails to clarify whether the respondent is being asked about his or her own gang-related decisions or the decisions of people in general. As much as is possible, clarify the “who,” the “what,” the “where,” and the “when” as well as the effect that is of interest.

Tip 7: Avoid Slang and Abbreviations

Our earlier warning against jargon points to the importance of writing in familiar and accessible language, but pursuit of the colloquial can misfire when the researcher drifts into the use of slang. Although many respondents might refer

to marijuana as “weed,” it would be unwise to assume that this term is universally understood. Abbreviations carry related problems. An item like “I know the whereabouts of my child 24/7” might fit the way many parents talk, but it might also be confusing to parents unfamiliar with the phrase. As “square” as it might seem to clearly define your terms and stick to dry, clinical language, this approach to writing questions will often reduce the influence of random error.

Reducing Systematic Error

As with random error, the potential sources of systematic error might extend far beyond a researcher’s ability to anticipate. There are some common culprits, however, and researchers should be on the lookout for them. Chief among these are sources of systematic error that can come about as a function of respondent demographics: gender, age, race, education, and socioeconomic status, to name a few. Self-report items written by a researcher from the viewpoint of his or her own social groups and life experiences might have far different meanings to respondents reading them from the vantage of different groups and experiences. For instance, a female researcher asking questions concerning “sexual harassment” might fail to realize that her male (but not female) respondents bring far different interpretations to this term than she had in mind while she was constructing her questions. Similarly, questions assessing “attitudes towards education” might be interpreted differently by children whose parents are college graduates, compared to those whose parents are high school dropouts. Much as with random error, one way of reducing systematic error is to define one’s terms and write questions clearly, such that a single, unambiguous meaning dominates.

Importantly, however, demographic differences are not the only factors that can exert systematic influences on ratings. Error can also be introduced as a result of any number of psychological attributes that exert influence on ratings. Earlier (Figure 1.2), we pointed to one potential source of systematic bias, social desirability. Concern for one’s public and private image can undermine self-reports on a wide range of sensitive topics. Practices that have been shown to reduce the effects of social desirability on self-reports include:

- Use of self-administered as opposed to face-to-face reports, such that respondents do not have to report sensitive behaviors directly to another person.
- Use of anonymous or confidential conditions, offering respondents reassurance that identifying information will not be associated with their data.
- Delivery of motivational instructions at the outset, encouraging honest reporting.
- Instructing respondents not to answer question at all, if they are not going to be truthful in their response (and using state of the art analytic methods to handle the missing data that results).

- Obtaining a measure of social desirability tendencies and using it as a statistical covariate when modeling the data.

Any or all of these methods might be applied to lessen the impact of desirability concerns on reporting, but it also is important to consider ways to eliminate bias through the design of better self-report items. This leads us to two new tips:

Tip 8: Avoid Leading Questions

Sometimes while writing questions, we reveal our own assumptions and values. The linguistic cues that lead a research participant to respond in certain ways can appear subtle but still exert influences on the ratings given. An item phrased as “To what extent does your mother disapprove of marijuana?” might elicit different answers than an item phrased as “To what extent does your mother approve or disapprove of marijuana?” The former item might lead or encourage respondents to communicate disapproval, as it fails to acknowledge that some mothers do hold favorable views towards marijuana.

Tip 9: Convey Your Acceptance of Potentially Undesirable Answers

Questions can be worded such that they reduce the sting of providing socially undesirable (but truthful) responses. For example, research suggests that older adults are less comfortable reporting their age than they are reporting the year they were born. They also are at times more comfortable checking off age categories than listing out their own specific age. One can also write questions in a manner that conveys acceptance. For instance, it is often the case that far more people indicate to pollsters that they voted in previous elections than voting rolls would indicate. People who did not vote might feel embarrassed to admit this, but some degree of embarrassment might be removed with careful questioning (e.g., “There are many reasons why people don’t get a chance to vote. Sometimes they have an emergency, or are ill, or simply can’t get to the polls. Did you vote in the last election?”). This strategy might make what was undesirable feel acceptable, but one has to be careful when using it not to be leading.

There are many other common forms of systematic error that one might also consider. For instance, psychometricians have identified a range of specific response styles, including (a) acquiescence response sets (i.e., the tendency make ratings indicating agreement), (b) disacquiescence response sets (i.e., the tendency to make ratings indicating disagreement), and (c) a middle-category response set (i.e., the tendency to move the midpoint of rating scales). The empirical evidence for prevalence of the contaminating influence of these artifacts is somewhat inconsistent, but it is clear they operate for some populations, in some contexts

(see Conway & Lance, 2010; Podsakovv, MacKenzie, & Podsakovv, 2012; Rorer, 1965; Wiggins, 1973). These possibilities do point to another tip:

Tip 10: Write Both Positively and Negatively Keyed Items, as Appropriate

One approach to dealing with acquiescence and disacquiescence response sets is to pursue a balance of positively and negatively keyed items. If one is seeking to measure extraversion, for instance, it might be a good idea to include positively keyed items (assessing such things as comfort talking to people), as well as negatively keyed items (assessing such things as interest in being alone). This advice comes with two large caveats, however.

First, it is important to treat as an empirical question the factor structure of a multi-item scale containing both positively and negatively keyed items. It may be that as a result of one general factor, extraversion, the greater comfort someone has talking to people, the less interest that person has in being alone. Or, it may be that the construct measured by positively keyed items (extraversion) is empirically distinct from the constructs measured by negatively keyed items (introversion). In research on attitude structure, for instance, researchers often find evidence that positive and negative evaluations of the same object are empirically distinct. Positive and negative evaluations can have distinct cognitive and emotional antecedents, as well as distinct consequences for judgment, decisions, and behavior (Cacioppo, Gardner, & Bernston, 1997), and so unidimensionality should not be assumed.

Second, when writing questions, it is important to generate positively and negatively keyed items that are non-redundant and equally sensible (see Weijters & Baumgartner, 2012). People can run afoul on both counts if their strategy for generating negatively keyed items is to try to “reverse” other, positively keyed items in an inventory. By simply reversing a sensible question, a nonsensible sentence might result. This is particularly likely if the new item is created through negation, which we noted earlier can introduce error. Whereas respondents might find it easy to answer “to what extent does your mother approve or disapprove of marijuana,” they might react with confusion when asked “to what extent does your mother NOT approve or disapprove of marijuana?”

Another problematic approach to producing reverse-keyed items is to include reverse-oriented, counterintuitive scales. Whereas this response metric is intuitive:

How much do you like going to parties?								
<i>Not at All</i>	0	1	2	3	4	5	6	<i>Extremely</i>

This metric might seem odd (and highly confusing) to many respondents:

How much do you like going to parties?								
<i>Extremely</i>	0	1	2	3	4	5	6	<i>Not at all</i>

It is reasonable to expect that some respondents answering the second question will wonder why enjoyment implies a lower number. Are they misunderstanding the question? Does the researcher have some trick up a sleeve? By introducing provocative and counterintuitive metrics into the mix, respondents might slow and perhaps become confused, producing misreporting.

Designing Item Metrics

The Pursuit of Rating Precision

Items are often rated on metrics using judgments such as agree-disagree, true-false, approve-disapprove, or favorable-unfavorable. Such metrics can be dichotomous (“yes” versus “no”) or many-valued (such as “strongly agree,” “moderately agree,” “neither,” “moderately disagree,” and “strongly agree”). The *precision* of a metric or scale refers to the number of discriminations it allows the respondent to make. Earlier we showed how precision might be reduced if questions are worded in a way that yields ceiling or floor effects, but the metrics one employs can have similar effects, if they force respondents who have meaningfully different evaluations to use the same category to describe their states of mind. Consider an item and response scale like this:

How much do you approve or disapprove of the Affordable Care Act?

_____ *Disapprove* _____ *Approve*

This question creates a reality in which respondents who “slightly disapprove” of the Affordable Care Act will receive the same score as those who “strongly oppose” it. Treating such people as if they are the same when analyzing data can introduce bias into parameter estimates and adversely affect statistical power. A simulation study by Bollen and Barb (1981) is informative. These authors created data, such that the true population correlations between two continuous variables were either 0.20, 0.60, 0.80, or 0.90. They then created “coarse” measures from the continuous measures for each population, by breaking the continuous measures into anywhere from 2 to 10 categories. For example, a continuous variable that ranges from -3 to $+3$ can be turned into a two-point scale by assigning anyone with a score of 0 or less a “0” and anyone with a score greater than 0 a “1.”

They found that true correlations were relatively well reproduced by coarse measures, as long as the coarse measures had 5 or more categories. For example, the reproduced correlations for five-category measures were within about 0.06 correlation units of the continuous-based correlations, when the true correlations were at or below 0.60. They concluded that five categories were probably sufficient for many research applications, and this recommendation has been borne out in many other simulation studies (although some research suggests seven or more categories may be best in some contexts; see Green, Akey, Fleming,

Hershberger, & Marquis, 1997; Lozano, García-Cueto, & Muñiz, 2008; Lubke & Muthén, 2004; Taylor, West, & Aiken, 2006). Thus, coarse measurement is not necessarily problematic, unless it is very coarse, namely less than five categories, leading us to the next tip:

Tip 11: In Most Instances, Orient Questions Around Five or More Response Categories

There are some caveats to this tip as well. First, this only applies to psychological attributes that are continuous in form. For ratings that orient around nominal categories (e.g., country of origin) the number of categories are dictated by the substantive content of the construct. For populations where researchers believe the cognitive demands of using a rating scale with five or more points is problematic, precision often can be had by delivering responses orally and in multiple steps. For example, one might ask respondents if they “agree,” “disagree,” or have “no opinion” about a given statement. Those who agree can then be asked in a follow-up if they “strongly” or “moderately” agree, just as those who disagree can be asked if they “strongly” or “moderately” disagree. Across the two questions, the researcher can then classify the respondent as having chosen one of the five categories (“strongly disagree,” “moderately disagree,” “neither,” “moderately agree,” or “strongly agree”).

Inclusion of a “Don’t Know” Response?

A common criticism of ratings scales is that they structure answers to such a degree that respondents are able to report evaluations that mean nothing to them (Sniderman, Tetlock, & Elms, 2001). One strategy that is sometimes used to combat this is to offer respondents a “don’t know” or “no opinion” response option. With this option, a respondent does not have to answer a question. According to some theorists, people indeed often have “no opinion” on a topic and if forced to respond to an item without them allowing to indicate “don’t know,” they will either respond randomly or in a non-meaningful way based on situational features in the testing context or their mood. If we include “don’t know” options, however, we may end up with a large number of answers that must be coded as non-responses. Despite plausible predictions to the contrary, extant research on this matter does not support the universal assertion that inclusion of a “don’t know” response category increases the reliability or validity of a measure, although there are some exceptions. In our view, a better strategy for generating meaningful data is to conduct qualitative research before questions are created to gain a better understanding of what questions are or are not meaningful to those in the population of interest, and to write questions accordingly (see Fisher, Fisher, & Aberizk, this volume).

Choosing Adverb Qualifiers

Data analyses promote stronger conclusions when a researcher works with measures that have interval or ratio-level properties, rather than nominal or ordinal properties. Interval properties often can be better approximated with rating scales using adverb qualifiers, as when respondents are asked if they agree with a statement “a little” or “a lot.” Interval-level properties can prove elusive if one utilizes a discrete set of adverbs that create unequal intervals or “spacing” between them, as with this question and response:

How much do you love puppies?			
<i>Not at All</i>	<i>A Little</i>	<i>Somewhat</i>	<i>Completely</i>

The difference between puppy-loving “a little” and “somewhat” seems slight, especially compared to the difference between a puppy-loving level of “somewhat” versus “completely.” This set of response categories illustrates the importance of pursuing anchors that create equal-appearing intervals, covering the full range of possible evaluations from low to high. There are large literatures in psychometrics that researchers can consult to identify adverb sets that help produce equal-appearing intervals (Budescu & Wallsten, 1994; Cox, 1980; Czaja & Blair, 2005; Dawes & Smith, 1985; Krosnick & Fabrigar, 1998; Tourangeau & Rasinski, 1988). As one example, an early study in psychophysics attempted to determine the modifying value of different adverbs. Cliff (1959) found that describing something as “slightly good” was perceived to be about 0.33 times as “good” as the simple, unmodified “good” (also see 1966a, 1966b). By consulting research on modifying values of adverbs, one can choose adverb qualifiers to more closely approximate equal-appearing intervals, thus producing ratings that more closely approximate interval-level properties. To be sure, some care must be taken in doing so, because qualifying values have been found to vary somewhat as a function of the population being studied and the type of judgment being made. However, more often than not, use of carefully selected adverb qualifiers will produce data that reasonably approximate interval-level properties. As a practical aid to readers, the Appendix provides sets of adverb qualifiers that produce roughly interval-level data for a wide range of judgments (and see Vagias, 2006).

Combining Numeric Ratings With Adverb Anchors

Data analyses generally are more straightforward and efficient if one works with measures that have at least interval-level properties. In this pursuit, self-report scales often orient around simple numeric rating scales. However, the mere fact that respondents can answer your question within provided numeric sequences

does not mean that you have achieved interval measurement. As an example, consider the following:

How do you feel?	<i>Sad</i>	1	2	3	4	<i>Happy</i>
------------------	------------	---	---	---	---	--------------

This researcher is interested in measuring mood but there is a mismatch between the numbering system and the anchors. The anchors suggest interest in a bipolar construct, anchored at one pole with “Sad” and the other by “Happy” but the rating system is unipolar, moving from 1 to 4. How would a respondent indicate a neutral mood? It also is unclear why “Sad” is associated with a small quantity (the value of 1), compared to “Happy”—can’t sad be felt with intensity? This example leads us to introduce a number of additional tips.

Tip 12: Communicate a “Zero Judgment” on Your Scale and Give It the Value of Zero

Some researchers include midpoints in rating scales (e.g., a “neutral” or “neither agree nor disagree” category), whereas other researchers omit them in order to force a respondent to “take a stand.” Use of a midpoint is theoretically warranted if it represents a valid psychological response for the judgment in question. Indeed, respondents may become irritated if they are not allowed to express their true feelings or opinions. There has been considerable research on the use or non-use of midpoints and, although somewhat mixed, overall the research tends to favor the use of midpoints as long as they are theoretically meaningful.

Tip 13: Communicate Bipolar Dimensions With Bipolar Rating Systems, Centered on Zero

Consider a researcher interested in measuring mood on a scale that ranges from extreme sadness to extreme happiness, with a midpoint of neutrality. This can be expressed as:

How do you feel?	–3	–2	–1	0	1	2	3
	<i>Extremely</i>			<i>Neutral</i>			<i>Extremely</i>
	<i>Sad</i>						<i>Happy</i>

A scale such as this communicates clearly the researcher’s conceptualization of mood as a bipolar evaluative dimension. It also utilizes a middle anchor to clarify the meaning of the zero-point; i.e., as the absence of either sadness or happiness.

Tip 14: Place Anchors That Approximate Interval-Level Distinctions at Equal-Appearing Numeric Intervals on the Response Scale

Error can be introduced in rating scales if they lack anchors at key points on the scale. Consider the following:

How happy are you? *Not at All* 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 *Extremely*

Researchers might use scales such as these in the hopes of increasing precision, but error can be introduced by such a numbering system, because it requests discriminations in terms of magnitude that likely go beyond the respondents' abilities to discern and/or communicate. What is the difference between happiness of 5 and 7 or between 7 and 12? Verbal anchors can help eliminate confusion about such rating systems. Earlier we discussed adverb modifiers that can be used to approximate interval-level rating systems (see also the Appendix). One fruitful approach is to combine these with the rating systems just discussed, reducing "number val-ues" requesting fine discriminations. Here are two such examples:

How happy are you?

0	1	2	3	4	5	6	7	8	9	10
<i>Not at All</i>			<i>Slightly</i>			<i>Quite</i>			<i>Extremely</i>	
<i>Happy</i>			<i>Happy</i>			<i>Happy</i>			<i>Happy</i>	

How do you feel?

-3	-2	-1	0	1	2	3
<i>Extremely</i>	<i>Quite</i>	<i>Slightly</i>	<i>Neutral</i>	<i>Slightly</i>	<i>Quite</i>	<i>Extremely</i>
<i>Sad</i>	<i>Sad</i>	<i>Sad</i>		<i>Happy</i>	<i>Happy</i>	<i>Happy</i>

With each of these ratings scales, the evaluative dimension is communicated through the use of a sensible numbering system and well-chosen anchors.

Tip 15: Add Extreme Anchors, When There Is Meaningful Variability at the Extremes

Sometimes opinions of interest will be endorsed extremely and with a high degree of consensus in populations of interest. As one example, Sweeney, Blanton, and Thompson (2009) sought to measure soldier's trust in their "most trusted leader," in the days before they participated in the launch of the second Gulf War. Needless to say, soldiers in these instances tended to have exceptionally high trust in this individual—so much so that one could reasonably anticipate that a ceiling effect would make this a meaningless rating. However, these researchers were able

to avoid this problem adding an extreme anchor and expanding precision around the extreme. The resulting question thus read:

To what extent do you trust your most trusted leader?												
0	1	2	3	4	5	6	7	8	9	10	11	12
<i>Not at All</i>			<i>Slightly</i>			<i>Quite a Bit</i>			<i>Extremely</i>			<i>Completely</i>

Although this scale might appear to request a wide range of evaluations from respondents, the researchers effectively administered a 4-point scale to this group of soldiers, as all but a handful of made ratings that ranged from 9 to 12. Similar strategies can be useful for predicting such things as adolescent health-risk tendencies, as even those likely to engage in risky behaviors tend to express negative evaluations, but to varying (and predictive) degrees (Gibbons, Gerrard, Blanton, & Russell, 1998). Burrows and Blanton (2015) reported results of a pilot study where they successfully predicted the likelihood of driving under the influence of alcohol (DUI), using a response scale that asked respondents to discriminate whether they were “completely” unwilling to drive under the influence or just “extremely” unwilling to DUI. In our view, one of the arguments for utilizing implicit measures rather than self-reports—i.e., that people often will not report socially undesirable attitudes—might in some instances be more easily addressed by giving respondents the option of making extreme ratings. Consider the measurement of racial bias, for instance, where it is often argued that respondents will not report socially undesirable attitudes they possess. Rather than pursuing implicit measurement strategies as a response, however, one might seek to measure how “completely” or “absolutely” individuals reject prejudicial beliefs and attitudes, using where the more moderate position is simply to reject prejudicial attitudes “extremely” (see Blanton & Jaccard, 2015).

Conclusion

Self-report is and will likely remain the most ubiquitous method of psychological assessment, in part because self-report items are easy to construct. Often missed, however, is the ease with which self-report items might be constructed, badly. We hope this chapter illustrates that the likelihood of writing strong questions can be increased through rigorous application of measurement principles. Researchers can improve their questions by clearly defining their constructs in terms of breadth and dimensionality, articulating scaling functions desired of questions, and paying close attention to sources of random and systematic error, such that they write stronger questions, and provide more informative ratings scales, and combine multiple items when conditions suggest this will improve measurement of the construct of interest.

APPENDIX

Across a wide range of psychometric studies, the following two sets of adverb qualifiers tend to produce roughly interval-level data:

For agreement judgments, two sets of reasonable qualifiers are:	
<i>Strongly agree</i>	<i>Strongly agree</i>
<i>Moderately agree</i>	<i>Agree</i>
<i>Neither</i>	<i>Neither</i>
<i>Moderately disagree</i>	<i>Disagree</i>
<i>Strongly disagree</i>	<i>Strongly disagree</i>
For frequency judgments, two sets of reasonable qualifiers are	
<i>Very frequently</i>	<i>Always or almost always</i>
<i>Frequently</i>	<i>Usually</i>
<i>Occasionally</i>	<i>About half the time</i>
<i>Rarely</i>	<i>Sometimes</i>
<i>Never</i>	<i>Never or almost never</i>
For importance judgments, two useful sets of adverb qualifiers are	
<i>Extremely important</i>	<i>Very important</i>
<i>Quite important</i>	<i>Moderately important</i>
<i>Slightly important</i>	<i>Slightly important</i>
<i>Not at all important</i>	<i>Unimportant</i>
For bipolar affective judgments, two sets of reasonable adverb qualifiers are	
<i>Extremely favorable</i>	<i>Very good</i>
<i>Quite favorable</i>	<i>Quite good</i>
<i>Slightly favorable</i>	<i>Slightly good</i>
<i>Neither</i>	<i>Neither</i>
<i>Slightly unfavorable</i>	<i>Slightly bad</i>
<i>Quite unfavorable</i>	<i>Quite bad</i>
<i>Extremely unfavorable</i>	<i>Very bad</i>

For extreme ratings, where ceiling or floor effects appear likely, consider adding extreme options (e.g., “absolutely” or “completely”) to expend beyond traditional endpoint (e.g., “extremely”).

References

- Beckstead, J. (2014). On measurements and their quality. Paper 4: Verbal anchors and the number of response options in rating scales. *International Journal of Nursing Studies*, 51(5), 807–814. doi:10.1016/j.ijnurstu.2013.09.004
- Blanton, H., & Jaccard, J. (2015). Not so fast: Ten challenges to importing implicit attitude measures to media psychology. *Media Psychology*, 18(3), 338–369.
- Bollen, K. A., & Barb, K. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, 46, 232–239.
- Budescu, D. V., & Wallsten, T. S. (1994). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from the perspective of cognitive psychology*. New York, NY: Academic Press.
- Burrows, C. N., & Blanton, H. (2015). Real-world persuasion from virtual world campaigns: How transportation into virtual worlds moderates in-game influence. *Communication Research*, 43(4), 542–570.
- Cacioppo, J. T., Gardner, W. L., & Bernston, C. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, 1(1), 3–25.
- Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, 66, 27–44.
- Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business Psychology*, 25, 325–334.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17, 402–422.
- Czaja, R., & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures*. Thousand Oaks, CA: Pine Forge Press.
- Davidson, A. R., & Jaccard, J. J. (1979). Variables that moderate the attitude—behavior relation: Results of a longitudinal survey. *Journal of Personality and Social Psychology*, 37(8), 1364–1376. doi:10.1037/0022-3514.37.8.1364
- Dawes, R. M., & Smith, T. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (pp. 509–566). New York, NY: Random House.
- Edwards, A. L. (1957). *Techniques of attitude scale construction*. East Norwalk, CT: Appleton-Century-Crofts.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press (Taylor & Francis).
- Fleming, P. (2012). Social desirability, not what it seems: A review of the implications for self-reports. *The International Journal of Educational and Psychological Assessment*, 11(1), 3–22.
- Furr, M., & Bacharach, V. (2018). *Psychometrics: An introduction* (2nd ed.). Newbury Park: Sage Publications.
- Gibbons, F. X., Gerrard, M., Blanton, H., & Russell, D. (1998). Reasoned action and social reaction: Intention and willingness as independent predictors of health risk. *Journal of Personality and Social Psychology*, 74(5), 1164–1180.
- Green, B. B. (1954). Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology* (pp. 335–469). Cambridge, MA: Addison-Wesley.
- Green, S. B., Akey, T., Fleming, K., Hershberger, & Marquis, J. (1997). Effect of the number of scale points on chi square fit indices in confirmatory factor analysis. *Structural Equation Modeling*, 4, 108–120.

- Howe, E. S. (1966a). Verb tense, negatives and other determinants of the intensity of evaluative meaning. *Journal of Verbal Learning and Verbal Behavior*, 5, 147–155.
- Howe, E. S. (1966b). Associative structure of quantifiers. *Journal of Verbal Learning and Verbal Behavior*, 5, 156–162.
- Kallgren, C. A., & Wood, W. (1986). Access to attitude-relevant information in memory as a determinant of attitude-behavior consistency. *Journal of Experimental Social Psychology*, 22(4), 328–338. doi:10.1016/0022-1031(86)90018-90011
- Kline, R. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.
- Krosnick, J. A., & Fabrigar, L. R. (1998). *Designing good questionnaires: Insights from psychology*. New York, NY: Oxford University Press.
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of ratings scales. *Methodology*, 4, 73–79.
- Lubke, G., & Muthén, B. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, 11, 514–534. Retrieved from <https://doi.org/10.1027/1614-2241.4.2.73>
- Nunnally, J., & Bernstein, I. (2004). *Psychometric theory*. New York, NY: McGraw-Hill.
- Podsakov, P. M., MacKenzie, S. B., & Podsakov, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569.
- Raykov, T. (2001). Bias of Cronbach's coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25, 69–76.
- Raykov, T., & Marcoulides, G. A., (2015). Scale reliability evaluation under multiple assumption violations. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2), 302–313. doi:10.1080/10705511.2014.938597
- Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63(3), 129–156. <http://dx.doi.org/10.1037/h0021888>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120.
- Sniderman, P. M., Tetlock, P. E., & Elms, L. (2001). Public opinion and democratic politics: The problem of nonattitudes and the social construction of political judgment. In J. H. Kuklinski & J. H. Kuklinski (Eds.), *Citizens and politics: Perspectives from political psychology* (pp. 254–288). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511896941.013
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113, 117–143.
- Sweeney, P. J., Thompson, V., & Blanton, H. (2009). Trust in combat: A test of an interdependence model and the links to leadership in Iraq. *Journal of Applied Social Psychology*, 39(1), 235–264.
- Taylor, A., West, S., & Aiken, L. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, 66, 228–239.
- Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299–314.
- Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, 5(3), 243–262. doi:10.1177/1745691610369465

- Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University. Retrieved November, 2017, from www.clemson.edu/content/dam/uc/sas/docs/Assessment/likert-type%20response%20anchors.pdf
- Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, 49(5), 737–747. doi:10.1509/jmr.11.0368
- Wiggins, J.S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley

2

IMPLICIT MEASURES

Procedures, Use, and Interpretation

Bertram Gawronski and Adam Hahn

There is no doubt that self-report measures have provided invaluable insights for a wide range of psychological questions (see Jaccard and Blanton, Chapter 1, this volume). After all, a straightforward way to find out what is on a person's mind is to directly ask the person about his or her thoughts and feelings. Yet, self-report measures have been criticized for their inability to capture mental contents that people are either unwilling or unable to report. First, self-report measures are known to be susceptible to self-presentation and socially desirable responding (Crowne & Marlowe, 1960). Second, self-report measures are not well-suited to capture thoughts and feelings that are outside of conscious awareness (Greenwald & Banaji, 1995). To overcome these limitations, psychologists have developed performance-based instruments that (a) limit participants' ability to strategically control their responses, and (b) do not rely on introspection for the measurement of thoughts and feelings. Based on their indirect approach in the assessment of mental contents, these performance-based instruments are often referred to as *implicit measures*, whereas traditional self-report measures are described as *explicit measures*.

Despite the popularity of implicit measures as a tool to overcome the two well-known problems of explicit measures, an accumulating body of research suggests that the relation between implicit and explicit measures involves a much more complex set of factors that cannot be reduced to motivational distortions and lack of introspective access. In a nutshell, the available evidence indicates that (a) strategic control is just one among several factors that can lead to dissociations between implicit and explicit measures and (b) the thoughts and feelings captured by implicit measures are consciously accessible (see Gawronski, LeBel, & Peters, 2007). Together, these findings pose a challenge to the common practice of interpreting dissociations between implicit and explicit measures as indicators

of socially desirable responding or lack of introspective awareness. Thus, to ensure accurate conclusions for theory development and real-world applications, it is imperative to use and interpret implicit measures in a manner that is consistent with the available evidence (for an overview, see Gawronski & Payne, 2010).

The current chapter provides an introduction to implicit measures that takes these issues into account. The overarching goal is to offer empirically based guidance for the appropriate use and interpretation of implicit measures. Toward this end, we first explain what it means for a measure to be implicit and then provide a brief overview of the most popular measurement instruments. Expanding on this overview, we discuss various factors that lead to converging versus diverging outcomes on implicit and explicit measures, and how implicit measures can complement explicit measures in individual difference and experimental designs. In the final section, we discuss some caveats against widespread, yet empirically unfounded, assumptions in research using implicit measures.¹

What Is “Implicit” About Implicit Measures?

A frequent question in research using implicit measures concerns the meaning of the terms *implicit* and *explicit*. This issue is a common source of confusion, because some researchers use the terms to describe features of measurement instruments, whereas others use them to describe the psychological constructs assessed by particular measurement instruments. For example, it is sometimes argued that participants are aware of what is being assessed by an explicit measure but they are unaware of what is being assessed by an implicit measure (e.g., Petty, Fazio, & Briñol, 2009). Yet, other researchers assume that the two kinds of measures tap into distinct memory representations, such that explicit measures capture conscious representations whereas implicit measures capture unconscious representations (e.g., Greenwald & Banaji, 1995).

Although these conceptualizations are relatively common in the literature on implicit measures, we deem it more appropriate to classify different measures in terms of whether the to-be-measured mental content influences participants' responses on the task in an automatic fashion (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). Specifically, measurement outcomes may be described as *implicit* if the impact of the to-be-measured mental content on participants' responses is unintentional, resource-independent, unconscious, or uncontrollable. Conversely, measurement outcomes may be described as *explicit* if the impact of the to-be-measured mental content on participants' responses is intentional, resource-dependent, conscious, or controllable (cf. Bargh, 1994). For example, a measure of racial attitudes may be described as implicit if it reflects participants' racial attitudes even when they do not have the goal to express these attitudes (i.e., unintentional) or despite the goal to conceal these attitudes (i.e., uncontrollable).

An important aspect of this conceptualization is that the terms *implicit* and *explicit* describe the process by which mental contents influence measurement

outcomes rather than the measurement instrument (cf. Petty et al., 2009) or the to-be-measured psychological construct (cf. Greenwald & Banaji, 1995). Moreover, whereas the classification of measurement outcomes as implicit or explicit depends on the processes that underlie a given measurement instrument, the instruments themselves may be classified as direct or indirect on the basis of their objective structural properties (De Houwer & Moors, 2010). Specifically, a measurement instrument can be described as direct when the measurement outcome is based on participants' self-assessment of the to-be-measured mental content (e.g., when participants' racial attitudes are inferred from their self-reported liking of Black people). Conversely, a measurement instrument can be described as indirect when the measurement outcome is not based on any self-assessment (e.g., when participants' racial attitudes are inferred from their reaction time performance in a speeded categorization task) or when it is based on a self-assessment that does not involve the to-be-measured mental content (e.g., when participants' racial attitudes are inferred from their self-reported liking of a neutral object that is quickly presented after a Black face). In line with this conceptualization, we use the terms *direct* and *indirect* to describe measurement instruments and the terms *explicit* and *implicit* to describe measurement outcomes. However, because claims about the automatic versus controlled nature of measurement outcomes have to be verified through empirical data, any descriptions of measures as *implicit* should be interpreted as tentative (for a review of relevant evidence, see De Houwer et al., 2009).

A popular way to conceptualize the mental contents captured by implicit measures refers to the idea of *mental association* (Greenwald et al., 2002). For example, the construct of *attitude* has been defined as a mental association between an object and its evaluation (Fazio, 2007). Expanding on this definition, *prejudice* can be defined as evaluative association involving a social group, and *self-esteem* as evaluative association involving the self. Similarly, *stereotypes* can be defined as semantic associations between a social group and stereotypical attributes, whereas the *self-concept* refers to semantic associations between the self and its attributes. In general, the concept of mental association is applicable to any kind of target objects (e.g., consumer products, political candidates) and their evaluative and semantic attributes. Implicit measures are based on the idea that activation of a mental concept can spread to other associated concepts in memory (Collins & Loftus, 1975). To the extent that the associative link between two concepts is sufficiently strong, spread of activation is assumed to occur automatically (i.e., unintentionally, unconsciously, efficiently, uncontrollably; see Bargh, 1994). Implicit measures make use of such automatic processes by assessing the effect of stimuli or stimulus features on participants' performance (e.g., response times, error rates) in responding to other stimuli or stimulus features. Although some theorists have proposed alternative frameworks that reject the notion of mental associations (e.g., De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011), associative theorizing has been a driving force in the development of implicit measures,

and it still serves as a prominent framework for their application in basic and applied research.

Measurement Instruments

Although there are more than a dozen performance-based instruments whose measurement outcomes may be described as implicit, some of them tend to be more popular than others. In current section, we briefly explain the procedural details of the most frequently used instruments and provide a list of less frequently used instruments for the sake of comprehensiveness.

Sequential Priming Tasks

Historically, the first type of performance-based instruments that has been used to measure social-psychological constructs is based on the notion of sequential priming (for a review, see Wentura & Degner, 2010). In a typical sequential priming task, participants are briefly presented with a prime stimulus, which is followed by a target stimulus. Depending on the nature of the task, participants are asked to (a) classify the target as positive or negative (i.e., evaluative decision task; see Fazio, Jackson, Dunton, & Williams, 1995), (b) classify the target in terms of a semantic property (i.e., semantic decision task; see Banaji & Hardin, 1996), or (c) decide whether the target is a meaningful word or a meaningless letter string (i.e., lexical decision task; see Wittenbrink, Judd, & Park, 1997). The basic idea underlying sequential priming tasks is that quick and accurate responses to the target should be facilitated when the target is congruent with the mental contents that were activated by the prime stimulus. In contrast, quick and accurate responses to the target should be impaired when the target is incongruent with the mental contents that were activated by the prime stimulus.

For example, if a person has a positive attitude toward Donald Trump, this person should be faster and more accurate in identifying the valence of positive words when the person has been primed with an image of Donald Trump compared to priming trials with a neutral baseline stimulus (e.g., Lodge & Taber, 2005). Conversely, evaluative classifications of negative words should be slower and less accurate when the person has been primed with an image of Donald Trump compared to priming trials with a neutral baseline stimulus. Similarly, a person who holds strong gender stereotypes should show better performance in identifying the gender of female pronouns after being presented with stereotypically female professions (e.g., nurse) than stereotypically male professions (e.g., doctor), and vice versa (e.g., Banaji & Hardin, 1996). Finally, using a lexical decision task to assess racial stereotypes, a person may show facilitated classifications of target words related to positive and negative stereotypes of African Americans (e.g., athletic, criminal) after being primed with Black faces compared to priming trials with a neutral baseline stimulus (e.g., Wittenbrink et al., 1997). Although

sequential priming tasks are among the most widely used instruments in research using implicit measures, they have been criticized for their low reliability, which rarely exceed Cronbach's Alpha values of .50 (Gawronski & De Houwer, 2014).

Implicit Association Test (and Variants)

The most prominent instrument in research using implicit measures is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). In the critical blocks of the IAT, participants are asked to complete two binary categorization tasks that are combined in a manner that is either congruent or incongruent with the to-be-measured mental content. For example, in the commonly used race IAT, participants may be asked to categorize pictures of Black and White faces in terms of their race and positive and negative words in terms of their valence. In one critical block of the task, participants are asked to press one response key for Black faces and negative words and another response key for White faces and positive words (i.e., prejudice-congruent block). In the other critical block, participants are asked to complete the same categorization tasks with a reversed key assignment for the faces, such that they have to press one response key for White faces and negative words and the other response key for Black faces and positive words (i.e., prejudice-incongruent block). The basic idea underlying the IAT is that responses in the task should be facilitated when two mentally associated concepts are mapped onto the same response key. For example, a person who has more favorable associations with Whites than Blacks should show faster and more accurate responses when White faces share the same response key with positive words and Black faces and share the same response key with negative words, compared with the reversed mapping.

IAT scores are inherently relative in the sense that they conflate four conceptually independent constructs. For example, in the race IAT, a participant's performance is jointly determined by the strength of White-positive, Black-positive, White-negative, and Black-negative associations (see Blanton, Jaccard, Gonzales, & Christie, 2006). This limitation makes the IAT inferior to sequential priming tasks, which permit the calculation of separate priming scores for each of the four associations if the tasks include appropriate baseline primes (see Wentura & Degner, 2010). Yet, the IAT is superior in terms of its internal consistency, which is typically in the range of .70 to .90 (Gawronski & De Houwer, 2014). At the same time, the IAT has been criticized for its blocked presentation of "congruent" and "incongruent" trials, which has been linked to several sources of systematic measurement error (see Teige-Mocigemba, Klauer, & Sherman, 2010). To address these and various other limitations, researchers have developed several variants of the standard IAT that avoid blocked presentations of congruent and incongruent trials, permit non-relative measurements for individual targets and attributes, and reduce the overall length of the task. These IAT variants include the Recoding-Free IAT (IAT-RF; Rothermund, Teige-Mocigemba, Gast, &

Wentura, 2009), the Single-Block IAT (SB-IAT; Teige-Mocigemba, Klauer, & Rothermund, 2008), the Single-Category IAT (SC-IAT; Karpinski & Steinman, 2006), the Single-Attribute IAT (SA-IAT; Penke, Eichstaedt, & Asendorpf, 2006), and the Brief IAT (BIAT; Sriram & Greenwald, 2009).

Go/No-Go Association Task

Another task that has been developed with the goal of overcoming the relative nature of measurement scores in the standard IAT is the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). On the GNAT, participants are asked to press a button (*go*) in response to some stimuli, and to withhold a response (*no go*) to other stimuli. Different types of stimuli are then paired with the “go” response on different blocks of the task. For example, in one block of a GNAT to measure racial attitudes, participants may be asked to press the “go” button when they see a picture of a Black face or a positive word, and not respond to any other stimuli (which may include pictures of White faces, negative words, and distractor stimuli). In another block, participants may be asked to press the “go” button for pictures of Black faces and negative words, and not respond to any other stimuli. The same task may be repeated in two additional blocks for White instead of Black faces. Because GNAT scores are calculated on the basis of participants’ error rates (rather than response times) using signal detection theory (Green & Swets, 1966), the GNAT typically includes a response deadline (e.g., 600 ms) to increase the number of systematic errors. The GNAT has shown lower reliability estimates compared with the standard IAT (Gawronski & De Houwer, 2014). Yet, a clear advantage is the possibility to calculate GNAT scores for individual target objects (e.g., attitudes toward Blacks) instead of relative scores involving two target objects (e.g., relative preference for Whites of Blacks).

Extrinsic Affective Priming Task

Another measure that has been designed to address structural limitations of the IAT is the Extrinsic Affective Simon Task (EAST; De Houwer, 2003). On the EAST, participants are presented with target words (e.g., Pepsi) that are shown in two different colors (e.g., yellow vs. blue) and positive and negative words in white color. Participants are asked to respond to the colored words in terms of their color and to the white words in terms of their valence. In the critical block of the task, participants are asked to respond to positive white words and words of one color (e.g., yellow) with the same key and to negative white words and words of the other color (e.g., blue) with another key (or vice versa). Because the target words are presented in different colors over the course of the task, each target is sometimes paired with the response key for positive words and sometimes with the response key for negative words. The critical question is whether participants respond faster and more accurately to a given target depending on whether its

color requires a response with the “positive” or the “negative” key. A major advantage of the EAST is that it does not include blocked presentations of congruent and incongruent trials, which resolves the problems associated with the blocked structure of the IAT (see Teige-Mocigemba et al., 2010). Yet, the EAST has been shown to be inferior to the IAT in terms of its reliability and construct validity, which has been attributed to the feature that participants do not have to process the semantic meaning of the target words (De Houwer & De Bruycker, 2007a). To address this limitation, De Houwer and De Bruycker (2007b) have developed a modified variant of the EAST that ensures semantic processing of the target words, which they called the Identification-EAST (ID-EAST). Although the EAST has originally been designed to measure evaluative associations, some studies have demonstrated its applicability to the measurement of semantic associations (e.g., Teige, Schnabel, Banse, & Asendorpf, 2004).

Affect Misattribution Procedure

The Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) was designed to combine the structural advantages of sequential priming tasks with the superior psychometric properties of the IAT (for a review, see Payne & Lundberg, 2014). Two central differences to traditional priming tasks are that (a) the target stimuli in the AMP are ambiguous and (b) participants are asked to report their subjective evaluations of the targets. The basic idea is that participants may misattribute the affective feelings elicited by primes to the neutral targets, and therefore judge the targets more favorably when they were primed with a positive stimulus than when they were primed with a negative stimulus. For example, in an AMP to measure racial attitudes, participants may be asked to indicate whether they find Chinese ideographs visually more pleasant or visually less pleasant than average after being primed with pictures of Black versus White faces. A preference for Whites over Blacks would be indicated by a tendency to evaluate the Chinese ideographs more favorably when the ideographs followed the presentation of a White face than when they followed the presentation of a Black face. Interestingly, priming effects in the AMP emerge even when participants are explicitly informed about the nature of the task and instructed not to let the prime stimuli influence their evaluations of the targets (Payne et al., 2005).

Although the AMP has shown satisfactory reliability estimates that are comparable to those of the IAT (Gawronski & De Houwer, 2014; Payne & Lundberg, 2014), the task has been criticized for being susceptible to intentional use of the primes in evaluations of the targets (Bar-Anan & Nosek, 2012). However, this criticism has been refuted by research showing that correlations between AMP effects and self-reported intentional use of the primes reflect retrospective confabulations of intentionality (i.e., participants infer that they must have had such an intention when asked afterwards) rather than actual effects of intentional processes (e.g., Gawronski & Ye, 2015; Payne et al., 2013). Although the AMP was

originally designed to measure evaluative associations, modified procedures have been used to measure semantic associations (e.g., Krieglmeier & Sherman, 2012; Sava et al., 2012).

Approach-Avoidance Tasks

Approach-avoidance tasks are based on the idea that positive stimuli should elicit spontaneous approach reactions, whereas negative stimuli should elicit spontaneous avoidance reactions. In line with this idea, Solarz (1960) found that participants were faster at pushing a lever towards them (approach) in response to positive as opposed to negative stimuli, and pushing it away from them (avoidance) for negative as opposed to positive stimuli. Chen and Bargh (1999) expanded on this finding by instructing participants to make either an approach or an avoidance movement as soon as a stimulus appeared on screen. They then calculated participants' response time to a given stimulus depending on whether they had to show an approach or an avoidance movement in response to that stimulus. Their results showed that participants were faster in making an approach movement in response to positive compared to negative stimuli. Conversely, participants were faster in making an avoidance movement in response to negative compared to positive stimuli.

Initial accounts of approach-avoidance tasks interpreted the obtained response patterns as reflecting direct links between particular motor actions and motivational orientations (e.g., contraction of arm extensor = avoidance; contraction of arm flexor muscle = approach). However, in contrast to these accounts, more recent findings suggest that congruency effects in approach-avoidance tasks depend on the evaluative meaning that is ascribed to a particular motor action in the task. For example, Eder and Rothermund (2008) found that participants were faster in moving a lever backward in response to positive words than negative words when this movement was described as "pull" (positive) and the opposite movement as "push" (negative). In contrast, participants were faster in moving a lever backward in response to negative words than positive words when this movement was described as "downward" (negative) and the opposite movement as "upward" (positive). Corresponding patterns emerged for forward movements. These results suggest that the labels used to describe particular motor actions in approach-avoidance tasks are essential for accurate interpretations of their measurement outcomes. Although some versions of approach-avoidance tasks have shown satisfactory estimates of internal consistency, their reliability varies considerably depending on the variant that is used (Krieglmeier & Deutsch, 2010).

Other Instruments

Although the reviewed instruments are the most popular examples in the current list of available measures, there are several other instruments with unique

features that make them better suited for particular research questions. Although we do not have the space to explain the procedural details of these instruments here, we briefly list them for the sake of comprehensiveness. For example, the Action Interference Paradigm (AIP; Banse, Gawronski, Rebetez, Gutt, & Morton, 2010) has been developed for research with very young children who may not be able to follow the complex instructions of other tasks. The Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010) and the Relational Responding Task (RRP; De Houwer, Heider, Spruyt, Roets, & Hughes, 2015) have been designed to measure automatically activated propositions (rather than automatically activated associations). Other instruments have targeted various methodological limitations of existing measures (e.g., blocked structure, relative measurement, low reliability), including the Evaluative Movement Assessment (EMA; Brendl, Markman, & Messner, 2005), the Implicit Association Procedure (IAP; Schnabel, Banse, & Asendorpf, 2006), and the Sorting Paired Features Task (SPFT; Bar-Anan, Nosek, & Vianello, 2009).

Convergence vs. Divergence Between Implicit and Explicit Measures

The broader idea underlying the use of implicit measures is that they provide information that cannot be gained from explicit measures. This idea is prominently reflected in (a) research on the relation between implicit and explicit measures, (b) research using implicit and explicit measures to predict behavior, and (c) experimental research using implicit and explicit measures as dependent variables.

Relations Between Implicit and Explicit Measures

Correlations between implicit and explicit measures tend to be relatively low overall. Several meta-analyses have found average correlations in the range of .20 to .25 (e.g., Cameron, Brown-Iannuzzi, & Payne, 2012; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). These correlations have been interpreted as evidence that implicit and explicit measures capture related, yet conceptually distinct, constructs (e.g., Nosek & Smyth, 2007). However, such interpretations provide little insight into what these constructs are and why they are weakly related. More seriously, there is evidence that the average correlations obtained in meta-analyses underestimate their actual relation, in that the average correlations are suppressed by various methodological factors. One of the most essential factors in this regard is the low internal consistency of many implicit measures (see Gawronski & De Houwer, 2014). To the extent that the internal consistency of an implicit measure is relatively low, its correlation with explicit measures will be attenuated by measurement error (e.g., Cunningham, Preacher, & Banaji, 2001). Yet, such attenuated correlations may not necessarily reflect distinct psychological constructs.

Other factors that contribute to low correlations can be broadly interpreted in terms of the correspondence principle in research on attitude-behavior relations (Ajzen & Fishbein, 1977). In general, correlations between implicit and explicit measures tend to be higher if the two kinds of measures correspond in terms of their dimensionality and content. For example, Hofmann et al. (2005) found that implicit measures reflecting relative preferences for one group over another tend to show higher correlations to explicit measures of the same relative preference compared to explicit measures of absolute evaluations. Similarly, implicit measures of race bias using Black and White faces as stimuli tend to show higher correlations to evaluative ratings of the same faces compared to evaluative ratings of anti-discrimination policies and perceptions of racial discrimination (e.g., Payne, Burkley, & Stokes, 2008).² Thus, without correspondence at the measurement level, it seems premature to interpret low correlations as evidence for distinct constructs at the conceptual level.

In addition to these methodological factors, there are a number of psychological factors that influence correlations between implicit and explicit measures. Overall, correlations tend to be larger for self-reported feelings, affective reactions, and “gut” responses compared to judgments that are more cognitive in nature (e.g., Gawronski & LeBel, 2008; Smith & Nosek, 2011). For example, in a study by Banse, Seise, and Zerbes (2001), scores of a gay-straight IAT showed higher correlations to self-reported affective reactions towards gay people (e.g., self-reported affect when seeing two men kissing each other) compared to self-reported cognitive reactions (e.g., agreement with the statement that gay men should not be allowed to work with children). Implicit and explicit measures also show higher correlations when participants are given less time to think about their judgments than when they are encouraged to deliberate about their response (e.g., Ranganath, Smith, & Nosek, 2008).

Theoretically, varying relations between implicit and explicit measures have been explained in terms of the *activation* of mental contents versus the *application* of activated contents for overt judgments (for a review, see Hofmann, Gschwendner, Nosek, & Schmitt, 2005). For example, the MODE model (Motivation and Opportunity as DEterminants) assumes that implicit measures capture the activation of automatic associations in response to an object (Fazio, 2007). Depending on a person’s motivation and opportunity, the person may engage in deliberate processing to scrutinize specific attributes of the object. In this case, people are assumed to base their judgments on the nature of relevant attributes instead of automatically activated associations. Hence, to the extent that both the motivation and the opportunity for deliberate processing are high, correlations between implicit and explicit measures should be low. Yet, when either the motivation or the opportunity for deliberate processing are low, people are assumed to rely on their automatic reactions, leading to higher correlations between implicit and explicit measures.

A similar explanation is offered by the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011). According to the APE

model, implicit measures reflect the activation of mental associations on the basis of feature similarity and spatiotemporal contiguity. In contrast, explicit measures are assumed to reflect the outcome of propositional processes that assess the validity of activated mental contents for overt judgments. A central assumption of the APE model is that the propositional validation of activated mental contents involves an assessment of consistency, in that inconsistency requires a reassessment and potential revision of one's beliefs. Thus, correspondence between implicit and explicit measures is assumed to depend on whether the association captured by an implicit measure is consistent with other information that is considered for a self-reported judgment. To the extent that it is consistent with other salient information, it is usually regarded as valid and therefore used as a basis for self-reported judgments. However, if it is inconsistent with other salient information, people may reject this association in order to restore cognitive consistency (e.g., Gawronski, Peters, Brochu, & Strack, 2008; Gawronski & Strack, 2004). Thus, a central difference to the MODE model is that deliberate processing may not necessarily reduce the relation between implicit and explicit measures. Instead, the APE model predicts that such reductions should occur only when the additionally considered information is inconsistent with the association captured by the implicit measure. To the extent that deliberate processing involves a selective search for information that supports the validity of this association, deliberate processing may in fact increase rather than decrease the relation between implicit and explicit measures (e.g., Galdi, Gawronski, Arcuri, & Fries, 2012; Peters & Gawronski, 2011).

Prediction of Behavior With Implicit and Explicit Measures

A common question about implicit measures is whether they predict behavior, with several independent meta-analyses suggesting different conclusions (e.g., Cameron et al., 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Although this question is perfectly justified, it does not reflect the more nuanced theoretical views that have guided research on the prediction of behavior with implicit and explicit measures (for reviews, see Fries, Hofmann, & Schmitt, 2008; Perugini, Richetin, & Zogmaister, 2010). Instead of testing zero-order relations between implicit measures and behavioral criteria, a substantial body of research aimed at gaining a deeper understanding of predictive relations by focusing on the following three questions: (a) What kinds of behaviors do implicit and explicit measures predict? (b) Under which conditions do implicit and explicit measures predict behavior? (c) For whom do implicit and explicit measures predict behavior?

Inspired by the assumptions of dual-process theories (e.g., Fazio, 1990), one of the earliest findings was that implicit measures tend to outperform explicit measures in the prediction of spontaneous behavior (e.g., eye gaze in interracial interactions predicted by implicit measures of racial prejudice), whereas explicit

measures tend to outperform implicit measures in the prediction of deliberate behavior (e.g., content of verbal responses in interracial interactions predicted by explicit measures of racial prejudice). This double dissociation has been replicated in a variety of domains with several different measures (e.g., Asendorpf, Banse, & Mücke, 2002; Dovidio, Kawakami, & Gaertner, 2002; Fazio et al., 1995).

Expanding on the idea that the predictive validity of implicit and explicit measures is determined by automatic versus controlled features of the to-be-predicted behavior (Fazio, 2007; Strack & Deutsch, 2004), several studies have investigated contextual conditions under which implicit versus explicit measures are superior in predicting a given behavior. The main finding of this research is that explicit measures outperform implicit measures in the prediction of a given behavior under conditions of unconstrained processing resources, whereas implicit measures outperform explicit measures under conditions of constrained processing resources. For example, Hofmann, Rauch, and Gawronski (2007) found that candy consumption under conditions of cognitive depletion showed a stronger relation to an implicit measure of candy attitudes, whereas candy consumption under control conditions showed stronger relations to an explicit measure (see also Frieze, Hofmann, & Wänke, 2008). Similar findings have been obtained for the prediction of interpersonal behavior in interracial interactions (Hofmann, Gschwendner, Castelli, & Schmitt, 2008).

Adopting an individual difference approach, Hofmann, Gschwendner, Frieze, Wiers, and Schmitt (2008) found a similar moderation pattern for individual differences in working memory capacity (WMC). In a series of studies, Hofmann and colleagues found that implicit measures outperform explicit measures in the prediction of a given behavior for people with low WMC, whereas explicit measures outperform implicit measures for people with high WMC. The broader idea underlying this research is that individual differences in WMC and situationally available resources are functionally equivalent, such that the implementation of behavioral decisions via reflective processes requires cognitive resources (Strack & Deutsch, 2004). To the extent that cognitive resources are scarce, behavior will be determined by impulsive tendencies that result from automatically activated associations. This idea also resonates with another individual difference factor that has been found to moderate the prediction of behavior: a person's preferred thinking style. Several studies have shown that explicit measures are better predictors of behavior for people with a preference for a deliberative thinking style, whereas implicit measures are better predictors of behavior for people with a preference for an intuitive thinking style (e.g., Richetin, Perugini, Adjali, & Hurling, 2007).

Deviating from approaches in which implicit and explicit measures are seen as competitors in the prediction of behavior, several studies have investigated interactive relations between the two kinds of measures. The general assumption underlying these studies is that discrepancies between implicit and explicit measures are indicative of an unpleasant psychological state that people aim to reduce

(Rydell, McConnell, & Mackie, 2008). For example, people showing large discrepancies on implicit and explicit measures of a particular psychological attribute (e.g., attitude, self-concept) have been shown to elaborate attribute-related information more extensively than people with small discrepancies (e.g., Briñol, Petty, & Wheeler, 2006). In a similar vein, combinations of high self-esteem on explicit measures and low self-esteem on implicit measures have been shown to predict narcissistic and defensive behaviors (e.g., Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003).

Implicit and Explicit Measures as Dependent Variables in Experimental Designs

The available evidence for dissociations in the prediction of behavior raised the question of what determines the outcomes on implicit and explicit measures. This question has been particularly dominant in research on attitude formation and change, which has shown various dissociations in the antecedents of attitudes captured by implicit and explicit measures (for a review, see Gawronski & Bodenhausen, 2006). Whereas some studies found effects on explicit, but not implicit, measures (e.g., Gawronski & Strack, 2004; Gregg, Seibt, & Banaji, 2006), others showed effects on implicit, but not explicit, measures (e.g., Gibson, 2008; Olson & Fazio, 2006). Yet, other studies found corresponding effects on explicit and implicit measures (e.g., Olson & Fazio, 2001; Whitfield & Jordan, 2009). These inconsistent patterns posed a challenge to traditional theories of attitude formation and change, which inspired the development of new theories that have been designed to explain potential dissociations between implicit and explicit measures of attitudes (e.g., Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Petty, Briñol, & DeMarree, 2007).

One example is the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011), which distinguishes between the activation of associations in memory (*associative process*) and the validation of momentarily activated information (*propositional process*). According to the APE model, processes of association activation are driven by principles of similarity and contiguity; processes of propositional validation are assumed to be guided by principles of cognitive consistency. The distinction between associative and propositional processes is further linked to implicit and explicit measures, such that implicit measures are assumed to reflect the outcomes of associative processes, whereas explicit measures are assumed to reflect the outcomes of propositional processes. Drawing on several assumptions about mutual interactions between associative and propositional processes, the APE model implies precise predictions regarding the conditions under which a given factor should lead to (a) changes on explicit but not implicit measures; (b) changes on implicit but not explicit measures; (c) corresponding changes on explicit and implicit measures, with changes on implicit measures being mediated by changes on explicit measures; and (d) corresponding

changes on explicit and implicit measures, with changes on explicit measures being mediated by changes on implicit measures.

For example, consistent with the predictions of the APE model, cognitive dissonance has been shown to change explicit, but not implicit, evaluations (e.g., Gawronski & Strack, 2004). Conversely, repeated pairings of a neutral conditioned stimulus (CS) with a valenced unconditioned stimulus (US) have been shown to change implicit evaluations of the CS. Yet, explicit evaluations were influenced only when participants were instructed to focus on their feelings toward the CS, which presumably led to a validation of the affective reaction resulting from the newly formed associations (e.g., Gawronski & LeBel, 2008). The central implication of this research is that implicit measures can be more or less resistant to external influences than explicit measures, with their relative resistance depending on whether (a) a given factor targets the content of mental associations (leading to changes on implicit measures) or the perceived validity of activated contents (leading to changes on explicit measures), and (b) proximal changes in one of the two processes lead distal changes in the other process (i.e., when a newly formed association is perceived as valid or when propositional inferences influence the structure of mental associations).

Some Caveats

In the final section of this chapter, we discuss some caveats against widespread assumptions in research using implicit measures. Although the accuracy of these assumptions is often taken for granted, they are either conceptually problematic or inconsistent with the available evidence. Thus, it seems prudent to take these issues into account to ensure appropriate interpretations of the data obtained with implicit measures.

The Metric of Implicit Measures Is Arbitrary

Many of the scoring procedures for implicit measures involve the calculation of difference scores, in which latencies or error rates on “compatible” trials are compared with the latencies or error rates on “incompatible” trials (or neutral baseline trials). The resulting numerical values are often used to infer a psychological attribute on one side of a bipolar continuum if the resulting score is higher than zero (e.g., a preference for Whites over Blacks) and a psychological attribute on the other side of the continuum if the score is lower than zero (e.g., a preference for Blacks over Whites), with a value of zero being interpreted as a neutral reference point. Although such metric interpretations are very common (for a discussion, see Blanton & Jaccard, 2006), they are conceptually problematic because incidental features of the stimulus materials have been shown to influence both the size and the direction of implicit measurement scores (e.g., Bluemke & Fries, 2006; Bluemke & Fiedler, 2009; Scherer & Lambert, 2009; Steffens & Plewe, 2001).

Because it is virtually impossible to quantify the contribution of such material effects, absolute interpretations of implicit measurement scores are therefore not feasible regardless of whether they involve characteristics of individual participants (e.g., participant X shows a preference for Whites over Blacks) or samples (e.g., 80% of the sample showed a preference for Whites over Blacks).

Yet, it is important to note that most research questions in social and personality psychology do not require absolute interpretations, but instead are based on relative interpretations of measurement scores. The latter applies to experimental designs in which measurement scores are compared across different groups (e.g., participants in the experimental group show higher scores compared to participants in the control group) as well as individual difference designs in which measurement scores are compared across different participants (e.g., participant A has a higher score compared to participant B). Hence, the abovementioned problems do not necessarily undermine the usefulness of implicit measures in social and personality psychology, although they do prohibit the widespread practice of absolute interpretations of measurement scores of individual participants or samples.

Implicit Measures Do Not Provide a Window to the Unconscious

A common assumption in research using implicit measures is that they provide a window to the unconscious, including unconscious attitudes, unconscious prejudice, unconscious stereotypes, unconscious self-esteem, etc. (e.g., Bosson, Swann, & Pennebaker, 2000; Cunningham, Nezlek, & Banaji, 2004; Rudman, Greenwald, Mellott, & Schwartz, 1999). Such claims are based on the notion that implicit measures rely on performance-related indicators, and therefore do not require introspective access for the assessment of mental contents. However, this methodological fact does not permit the reverse inference that the mental contents captured by implicit measures are introspectively inaccessible (see Gawronski & Bodenhausen, 2015). Any such claims are empirical hypotheses that require supportive evidence. Importantly, the available evidence clearly contradicts the assumption that the mental contents captured by implicit measures are unconscious (for a review, see Gawronski, Hofmann, & Wilbur, 2006). Using multiple IATs capturing attitudes toward different social groups, Hahn, Judd, Hirsh, and Blair (2014) found that participants were quite accurate in predicting the patterns of their IAT scores. The median within-subjects correlation between predicted and actual scores across four studies (total $N = 430$) was $r = .68$ (average $r = .54$). Interestingly, the same analysis applied to the relation between explicit measures and IAT scores showed much lower correlations (average $r = .20$), similar to the ones typically observed in this area (see Hofmann et al., 2005). These findings pose a challenge to the claim that implicit measures provide a window to the unconscious. Yet, they are consistent with theories that explain dissociations

between implicit and explicit measures in terms of other processes that involve a deliberate rejection of consciously accessible associations (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006).

Dissociations Do Not Necessarily Reflect Motivated Distortions

Another common assumption in research using implicit measures is that they resolve the well-known problems of social desirability. This assumption is based on the notion that it is much more difficult to strategically influence one's scores on an implicit measure compared to one's scores on an explicit measure. Although it is correct that motivated distortions on explicit measures can lead to dissociations between implicit and explicit measures, the validity of this proposition does not permit the reverse conclusion that any dissociation reflects motivated distortions on explicit measures (see Gawronski & Bodenhausen, 2015). After all, dissociations can also result from cognitive processes, such as the deliberate analysis of specific attributes (see Fazio, 2007) or the consideration of additional information that is inconsistent with automatically activated associations (see Gawronski & Bodenhausen, 2006). Although either of these processes may elicit motivational concerns, they can lead to dissociations between implicit and explicit measures for purely cognitive reasons (e.g., Gawronski & LeBel, 2008).

Implicit Measures Do Not Provide Superior Access to Old Representations

Some theories suggest that implicit measures reflect highly stable, old representations whereas explicit measures reflect recently acquired, new representations (e.g., Petty, Tormala, Briñol, & Jarvis, 2006; Rudman, 2004; Wilson, Lindsey, & Schooler, 2000). The central idea underlying these theories is that previously formed representations may not be erased from memory when people acquire new information that is inconsistent with these representations. To the extent that earlier acquired knowledge is often highly overlearned, older representations are assumed to be activated automatically upon encounter of a relevant stimulus. In contrast, more recently acquired knowledge is usually less well learned, which implies that the retrieval of newer representations requires controlled processing. Based on these assumptions, implicit measures have been claimed to reflect highly stable, old representations whereas explicit measures reflect more recently acquired, new representations.

Conceptually, these assumptions imply two related, yet empirically distinct, predictions: (a) implicit measures are more resistant to change than explicit measures; (b) implicit measures are more stable over time than explicit measures. Both predictions are at odds with the available evidence. The first prediction stands in contrast to the large body of studies showing experimentally induced changes on

implicit, but not explicit, measures (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Olson & Fazio, 2006). The second prediction stands in contrast to the finding that implicit measures tend to show lower stability over time than explicit measures (Gawronski, Morrison, Phillips, & Galdi, 2017). Together, these findings pose a challenge to the widespread assumption that implicit measures reflect highly stable, old representations.

Implicit Measures Are Not Immune to Context Effects

Another common assumption about implicit measures is that they can help researchers to resolve the problem of context effects on self-reports. Research on response processes in self-report measures has identified a wide range of contextual factors that can undermine accurate assessments (for a review, see Schwarz, 1999). With the development of performance-based instruments that do not rely on self-assessments, many researchers expected to gain direct access to people's "true" characteristics without contamination by contextual factors. However, the available evidence suggests that implicit measures are at least as susceptible to contextual influences as explicit measures (for reviews, see Blair, 2002; Gawronski & Sritharan, 2010). Theoretically, most of these context effects can be explained with the distinction between activation and application discussed earlier in this chapter. The basic idea is that contextual factors may influence either (a) the activation of mental contents, which should lead to context effects on implicit measures or (b) the application of activated contents for overt judgments, which should lead to context effects on explicit measures. The bottom-line is that neither implicit nor explicit measures are immune to context effects, which poses a challenge to the idea that implicit measures provide context-independent reflections of people's "true" characteristics.

Implicit Measures Do Not Speak to the Automaticity of an Experimental Effect

A defining characteristic of implicit measures is that the to-be-measured mental content influences measurement outcomes in an automatic fashion (see De Houwer et al., 2009). Based on this assumption, implicit measures are sometimes included as dependent measures in experimental studies to test whether the employed manipulation influences the observed outcomes in an automatic fashion. However, such interpretations conflate the impact of mental contents on measurement outcomes with the impact of experimental manipulations on mental contents (Gawronski & De Houwer, 2014). To illustrate this difference, consider a study by Peters and Gawronski (2011) in which participants were asked to recall past behaviors reflecting either extraversion or introversion, and then to complete an IAT designed to measure associations between the self and extraversion (versus introversion). Results showed that IAT scores of self-extraversion associations

were higher when participants were asked to recall extraverted behaviors than when they were asked to recall introverted behaviors. Based on the (flawed) assumption that implicit measures can be used to identify automatic effects of experimental manipulations, one might be tempted to conclude that recalling past behaviors influenced self-associations in an automatic fashion. However, the task of recalling past behaviors was fully conscious, intentional, and controllable. Thus, a more appropriate conclusion is that (a) the experimental manipulation influenced the activation of self-associations in a non-automatic fashion, and (b) the activated self-associations influenced participants' responses on the IAT in an automatic fashion. Whereas the former refers to the effect of the experimental manipulation on mental contents, the latter refers to the effect of mental contents on measurement outcomes. The distinction between implicit and explicit measures speaks only to the latter effect, but it does not provide any insight about the nature of the former effect.

Implicit Measures Are Not Process-Pure

As we noted earlier in this chapter, implicit measures are often assumed to provide direct proxies for mental associations. However, in a strict sense, implicit measures reflect behavioral responses, and these responses should not be equated with their presumed underlying mental constructs (De Houwer, Gawronski, & Barnes-Holmes, 2013). Although the impact of mental associations on implicit measures is rarely disputed (for a notable exception, see De Houwer, 2014), a considerable body of research suggests that implicit measures do not provide process-pure reflections of mental associations (Sherman, Klauer, & Allen, 2010). To disentangle the contributions of multiple qualitatively distinct processes to implicit measures, theorists have developed formal models that provide quantitative estimates of these processes, including applications of process dissociation (Payne & Bishara, 2009), multinomial modeling (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Meissner & Rothermund, 2013; Stahl & Degner, 2007), and diffusion modeling (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007).

An illustrative example is Conrey et al.'s (2005) quad-model, which distinguishes between four qualitatively distinct processes underlying responses on implicit measures: (a) activation of an association (AC); (b) detection of the correct response required by the task (D); (c) success at overcoming associative bias (OB); and (d) guessing (G). Research using the quad-model has provided more fine-grained insights into the mechanisms underlying previous findings obtained with implicit measures. Whereas some effects have been shown to be genuinely related to underlying associations (e.g., changes on implicit measures of racial bias that result from extended training to associate racial groups with positive or negative attributes; see Calanchini, Gonsalkorale, Sherman, & Klauer, 2013), others stem from non-associative processes, such as successful versus unsuccessful

inhibition of activated associations (e.g., increases in implicit measures of racial bias after alcohol consumption; see Sherman et al., 2008).

The Reliability of Implicit Measures Varies Widely Across Instruments

A final issue concerns the reliability of implicit measures. Unfortunately, measurement error is an issue of concern for several of the reviewed measures, showing estimates of internal consistency that seem unsatisfactory from a psychometric point of view (for a summary, see Gawronski & De Houwer, 2014). The only two measures that have consistently shown acceptable estimates of internal consistency (e.g., Cronbach's Alpha values in the range of .70 to .90) are the IAT (Greenwald et al., 1998) and the AMP (Payne et al., 2005). Most other measures (e.g., GNAT, EAST) have shown estimates of internal consistency that are slightly lower than what might be deemed acceptable (e.g., Cronbach's Alpha values in the range of .50 to .70). The lowest estimates of internal consistency have been observed for sequential priming tasks (e.g., Cronbach's Alpha values below .50). Although concerns about reliability tend to be more common in personality psychology than in social psychology, low internal consistency can be a problem in both correlational and experimental designs. On the one hand, low internal consistency can distort the rank order of participants in terms of a particular construct, which reduces correlations to other measures (e.g., in studies on the prediction of behavior). On the other hand, low internal consistency can reduce the probability of identifying effects of experimental manipulations (e.g., in studies on attitude change), which includes both initial demonstrations of an experimental effect and replications of previously obtained effects (LeBel & Paunonen, 2011). Thus, regardless of whether implicit measures are used in correlational or experimental designs, it seems prudent to take their varying levels of internal consistency into account.

Conclusions

Historically, the use of implicit measures is rooted in the idea that they overcome the well-known limitations of explicit measures in capturing thoughts and feelings that people are either unwilling or unable to report. As should be clear from the evidence reviewed in this chapter, the relation between implicit and explicit measures is much more complex, in that it cannot be reduced to social desirability or lack of awareness. Moreover, many widespread assumptions about implicit measures are either conceptually problematic or inconsistent with the available evidence. However, these conclusions do not imply that implicit measures are useless. If implicit measures are used and interpreted in a manner that is consistent with the available evidence, they can provide valuable insights into the processes underlying social judgment. In addition, they can serve as a useful complement

in the prediction of behavior and in research on the formation and change of mental representations. Conceptually, dissociations between implicit and explicit measures in any of these applications can be interpreted as reflecting differences between (a) the activation of mental contents and (b) the application of activated contents for overt judgments. Given that the distinction between activation and application is relevant for almost any question regarding the mental processes underlying judgments and behavior, implicit measures still represent one of the most significant additions to the tool-box of psychological instruments.

Notes

- 1 Because of its shared concern with implicit measures, their use, and their conceptual meaning, the current chapter has overlap with previous publications by the authors addressing the same issues (e.g., Gawronski, 2009; Gawronski & De Houwer, 2014; Gawronski, Deutsch, & Banse, 2011; Gawronski et al., 2007; Hahn & Gawronski, 2015, 2018).
- 2 Judgments of anti-discrimination policies and perceptions of racial discrimination are central themes in the Modern Racism Scale (McConahay, 1986), which is often used as an explicit measure in research using implicit measures of racial bias.

References

- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84, 888–918.
- Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, 83, 380.
- Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, 7, 136–141.
- Banse, R., Gawronski, B., Rebetez, C., Gutt, H., & Bruce Morton, J. (2010). The development of spontaneous gender stereotyping in childhood: Relations to stereotype knowledge and stereotype flexibility. *Developmental Science*, 13, 298–306.
- Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie*, 48, 145–160.
- Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the Affective Misattribution Procedure. *Personality and Social Psychology Bulletin*, 38, 1194–1208.
- Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, 56, 329–343.
- Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1–40). Hillsdale, NJ: Erlbaum.
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, 60, 527–542.
- Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6, 242–261.

- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61, 27–41.
- Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, 42, 192–212.
- Bluemke, M., & Fiedler, K. (2009). Base rate effects on the IAT. *Consciousness and Cognition*, 18, 1029–1038.
- Bluemke, M., & Friesse, M. (2006). Do irrelevant features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, 42, 163–176.
- Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, 79, 631–643.
- Brendl, C. M., Markman, A. B., & Messner, C. (2005). Indirectly measuring evaluations of several attitude objects in relation to a neutral reference point. *Journal of Experimental Social Psychology*, 41, 346–368.
- Briñol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. *Journal of Personality and Social Psychology*, 91, 154–170.
- Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, 43, 321–325.
- Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, 16, 330–350.
- Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, 25, 215–224.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, 89, 469–487.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349–354.
- Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, 30, 1332–1346.
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measurement: Consistency, stability, and convergent validity. *Psychological Science*, 12, 163–170.
- De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, 50, 77–85.
- De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, 8, 342–353.
- De Houwer, J., & De Bruycker, E. (2007a). The implicit association test outperforms the extrinsic affective Simon task as an implicit measure of interindividual differences in attitudes. *British Journal of Social Psychology*, 46, 401–421.
- De Houwer, J., & De Bruycker, E. (2007b). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry*, 38, 133–143.

- De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, 24, 252–287.
- De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology*, 6, 319.
- De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135, 347–368.
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82, 62–68.
- Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, 137, 262.
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology*, 23, 75–109.
- Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, 25, 603–637.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, 19, 285–338.
- Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behavior. *British Journal of Social Psychology*, 47, 397–419.
- Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin*, 38, 559–569.
- Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology*, 50, 141–150.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692–731.
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, 44, 59–127.
- Gawronski, B., & Bodenhausen, G. V. (2015). Theory evaluation. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 3–23). New York, NY: Guilford Press.
- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). New York, NY: Cambridge University Press.
- Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, A. Voss, & C. Stahl (Eds.), *Cognitive methods in social psychology* (pp. 78–123). New York, NY: Guilford Press.

- Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit” attitudes unconscious? *Consciousness and Cognition*, 15, 485–499.
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, 44, 1355–1361.
- Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, 2, 181–193.
- Gawronski, B., Morrison, M., Phillips, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, 43, 300–312.
- Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.
- Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin*, 34, 648–665.
- Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). New York, NY: Guilford Press.
- Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, 40, 535–542.
- Gawronski, B., & Ye, Y. (2015). Prevention of intention invention in the affect misattribution procedure. *Social Psychological and Personality Science*, 6, 101–108.
- Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, 35, 178–188.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley-Blackwell.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97, 17–41.
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90, 1–20.
- Hahn, A., & Gawronski, B. (2015). Implicit social cognition. In J. D. Wright (Ed.), *The international encyclopedia of the social and behavioral sciences* (2nd ed., pp. 714–720). Oxford: Elsevier.
- Hahn, A., & Gawronski, B. (2018). Implicit social cognition. In J. T. Wixted (Ed.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., pp. 395–427). Malden, MA: Wiley-Blackwell.

- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143, 1369.
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31, 1369–1385.
- Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes and Intergroup Relations*, 11, 69–87.
- Hofmann, W., Gschwendner, T., Friese, M., Wiers, R., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Towards and individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology*, 95, 962–977.
- Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates implicit-explicit consistency? *European Review of Social Psychology*, 16, 335–390.
- Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology*, 43, 497–504.
- Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, 61, 465–496.
- Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology*, 85, 969–978.
- Karpinski, A., & Steinman, R. B. (2006). The single category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, 91, 16–32.
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process-components of the Implicit Association Test: A diffusion model analysis. *Journal of Personality and Social Psychology*, 93, 353–368.
- Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach-avoidance behavior: The manikin task vs. two versions of the joystick task. *Cognition and Emotion*, 24, 810–828.
- Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, 103, 205–224.
- LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, 37, 570–583.
- Lodge, M., & Taber, C. S. (2005). Automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, 26, 455–482.
- McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–126). New York, NY: Academic Press.
- Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, 104, 45–69.
- Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, 19, 625–666.
- Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology*, 54, 14–29.

- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, 12, 413–447.
- Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, 32, 421–433.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105, 171–192.
- Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology*, 20, 272–314.
- Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, 39, 375–386.
- Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94, 16–31.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277–293.
- Payne, B. K., & Lundberg, K. B. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, 8, 672–686.
- Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single Attribute Implicit Association Tests (SA-IAT) for the assessment of unipolar constructs: The case of sociosexuality. *Experimental Psychology*, 53, 283–291.
- Perugini, M., Richetin, J., & Zogmeister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). New York, NY: Guilford Press.
- Peters, K. R., & Gawronski, B. (2011). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology*, 47, 436–442.
- Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, 25, 657–686.
- Petty, R. E., Fazio, R. H., & Briñol, P. (2009). The new implicit measures: An overview. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 3–18). New York, NY: Psychology Press.
- Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST Model. *Journal of Personality and Social Psychology*, 90, 21–41.
- Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, 44, 386–396.
- Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit and explicit measures. *European Journal of Personality*, 21, 529–546.
- Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *The Quarterly Journal of Experimental Psychology*, 62, 84–98.

- Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science*, 13, 79–82.
- Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, 17, 437–465.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91, 995–1008.
- Rydell, R. J., McConnell, A. R., & Mackie, D. M. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance and increased information processing. *Journal of Experimental Social Psychology*, 44, 1526–1532.
- Sava, F. A., Maricutoiu, L. P., Rusu, S., Macsinga, I., Virga, D., Cheng, C. M., & Payne, B. K. (2012). An inkblot for the implicit assessment of personality: The semantic misattribution procedure. *European Journal of Personality*, 26, 613–628.
- Scherer, L. D., & Lambert, A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology*, 97, 383–403.
- Schnabel, K., Banse, R., & Asendorpf, J. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology*, 53, 69–76.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.
- Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. A., & Groom, C. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, 115, 314–335.
- Sherman, J. W., Klauer, K. C., & Allen, T. J. (2010). Mathematical modeling of implicit social cognition: The machine in the ghost. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 156–175). New York, NY: Guilford Press.
- Smith, C. T., & Nosek, B. A. (2011). Affective focus increases the concordance between implicit and explicit attitudes. *Social Psychology*, 42, 300–313.
- Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, 59, 239–245.
- Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, 56, 283–294.
- Stahl, C., & Degner, J. (2007). Assessing automatic activation of valence: A multinomial model of EAST performance. *Experimental Psychology*, 54, 99–112.
- Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie*, 48, 123–134.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.
- Teige, S., Schnabel, K., Banse, R., & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task. *European Journal of Personality*, 18, 495–520.
- Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment*, 24, 237–245.
- Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to Implicit Association Tests and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook*

- of implicit social cognition: Measurement, theory, and applications* (pp. 117–139). New York, NY: Guilford Press.
- Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). New York, NY: Guilford Press.
- Whitfield, M., & Jordan, C. H. (2009). Mutual influences of explicit and implicit attitudes. *Journal of Experimental Social Psychology, 45*, 748–759.
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101–126.
- Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationships with questionnaire measures. *Journal of Personality and Social Psychology, 72*, 262–274.

3

ELICITATION RESEARCH

*William A. Fisher, Jeffrey D. Fisher,
and Katrina Aberizk*

This chapter reviews the conceptual rationale that underlies elicitation research, describes methodological approaches for conducting elicitation research, and provides illustrations of the conduct and value of such research. At its core, elicitation research uses participant-informed, “bottom-up” research methods to identify salient factors that influence the formation of attitudes and beliefs and the performance of behavior (Fishbein, 1967; Fishbein & Ajzen, 1975, 2010). As such, elicitation research assiduously consults “the experts”—the thinking and acting human beings who form attitudes and beliefs and enact behaviors—and solicits their spontaneously occurring, “top-of-the-head” reports of factors that are relevant to these outcomes. At the same time, elicitation research is sensitive to the role of investigator-imposed assessment approaches that could provide occasions for artifactual responding. From the perspective of elicitation research, studies that rely solely on intuitively derived, investigator-imposed, close-ended assessment approaches may fail to identify salient factors that impact individual attitudes, beliefs, and behavior.

Researchers turn to elicitation methods when they are concerned that use of traditional, top-down assessment methods (e.g., self-report ratings scales) can inadvertently direct research participants to respond in ways that otherwise would not have occurred to them, had they not been asked, or that are otherwise not relevant to them. Responses to structured assessment techniques can therefore provide convincing but misleading evidence concerning the correlates and determinants of attitudes, beliefs, and behavior (Fishbein & Ajzen, 1975, 2010). This conceptualization of the vital importance of conducting “bottom-up,” open-ended elicitation research is far from novel and is not the authors’ own. Classical social psychological assessment (e.g., Fishbein, 1967), qualitative research

procedures (e.g., Denzin & Lincoln, 1994), and core traditions in anthropological research (Russell, 2006) have long emphasized the foundational importance of participant-informed research approaches. Here we draw on these approaches to provide a synthesis for readers interested in understanding their uses in basic and applied research endeavors.

Four Research Aims

It would be a mistake to think of elicitation as merely a “technique” or “tool” that a researcher might use to assess constructs of interest. Elicitation methods are only as informative as their relevance to achieving specific research aims. The first step for any researcher is to articulate the purpose and the role that elicitation methods might serve. Here we identify four such aims.

Descriptive Research

First, elicitation research may be conducted as a purely descriptive, qualitative effort to improve the understanding of some population of interest. In research of this nature, the goal typically is to illuminate previously unknown aspects of the population’s attitudes, beliefs, and behavior, and the factors that may influence them (e.g., Fishbein & Ajzen, 1975). Even very straightforward, open-ended qualitative approaches and associated systematic thematic analysis may be used to achieve this aim. For example, Shumlich (2017) simply asked several hundred undergraduates to write fine-grain descriptions of both (a) a first-time sexual encounter with a new partner, and (b) a sexual encounter with a long-term partner. Although many interesting descriptions emerged, what was striking was that not a single account of a sexual encounter mentioned an overt attempt to clarify a sexual partner’s consent to the sexual interaction. Through thematic analysis of these accounts and participant responses to related open-ended questions, researchers identified a widely held belief in the sample that “lack of partner resistance” indicates consent. It was further revealed that actually asking a partner if they consent to a sexual interaction was viewed by some in the sample as completely unacceptable, because it would “break the mood.”

The study by Shumlich (2017) illustrates what can be gained from even minimal elicitation prompts. This research relied on the most straightforward of questions (e.g., “Describe a first time sexual encounter”), yet it revealed an enormous gap between “what is” and “what should be” with respect to verification of consent in sexual interactions. Additional questions pursued by Shumlich provided hints as to why this gap exists and identified areas for future intervention. Readers interested in applying similar techniques might refer to Braun and Clarke (2006), who provide step-by-step instructions for the systematic conduct of thematic analysis of qualitative data in response to open-ended questions (and see Moustakas, 1994).

Early Stage Research

Second, elicitation research may be employed as a foundational qualitative step in the development of closed-ended quantitative assessments of attitudes, beliefs, or behavior. Here, the aim is to create a valid assessment of attitudes, beliefs, or behaviors that is grounded in participants' experience of the attitude, belief, or behavior at focus. Open-ended qualitative approaches, including individual written responses, semi-structured interviews, and focus-group discussions can be employed as a first step in the development of participant-informed quantitative measures. Assume that you are interested in what has come to be called "vaccine hesitancy" or resistance to or rejection of vaccination for one's self or one's children. Opel, Mangione-Smith, Taylor, and colleagues (2011) conducted elicitation research using focus groups to augment more traditional methods to scale development in this area. Their desire was to create a quantitative instrument assessing parental hesitancy to vaccinating their children. They began with a pool of items taken from existing measures that appeared to capture parents' hesitancy to vaccinate their children. Focus groups of parents were then asked independently for their views about vaccinating their children. Qualitative analysis of parents' focus group responses identified sources of parental hesitancy that were not addressed by the intuitively identified scale items. For instance, existing scales failed to tap sufficiently into parental skepticism about vaccination research described to them by their own doctors, just as they failed to tap into perceptions that vaccinations only protect against "rare" outcomes, or their fears that there is something "unnatural" about exposing children to vaccinations. These new concerns that surfaced in the focus groups were used to create new scale items that were added to the item pool to create a content valid scale, and the resulting scale can be used as an ecologically valid measure for predicting and understanding vaccine acceptance or refusal and as a baseline and outcome measure in intervention research to reduce vaccine hesitancy (see Sadaf, Richards, Glanz, Salmond, & Omer, 2013).

Theory Development

Third, elicitation research findings may assist in the design of ecologically valid tests of theory. Here the aim is to create participant-informed assessments of theoretically specified constructs so that the relationships among them may be tested with appropriate measures. The Information-Motivation-Behavioral Skills (IMB) model of health behavior (J. Fisher & Fisher, 1992; W. Fisher, Fisher, & Shuper, 2014) assumes that when one has information about a health behavior, is motivated to practice it, and has the requisite behavioral skills, one will practice the behavior in question. It assumes that the particular content of each of the IMB model's information, motivation, and behavioral skills constructs will vary as a function of the health behavior and the population at focus. Suppose you are a pharmaceutical manufacturer who has spent years and millions of dollars

developing a microbicide women in Africa can use to prevent contracting HIV. To test the assumptions of IMB model for predicting women's behavior in this context you would first conduct elicitation research with African women to illuminate the particular information, motivation, and behavioral skills elements they would need to use the microbicide. They might tell you they would require information about the proper dosage, administration, side effects, and safety of the drugs. Moreover, to be motivated to use the microbicide, they might say they would need to have positive attitudes toward the drug, and support from important others to use it. Finally, elicitation research might reveal that women would need the skills to apply the microbicide properly, to discuss its use with their sexual partner or to conceal its use, and to deal with any side effects. Based on elicitation research, the pharmaceutical manufacturer would then know how to design measures of microbicide information, motivation, and behavioral skills to test the basic assumption of the IMB model that women who are high on each of these indicators will practice higher levels of microbicide use (Ferrer, Morrow, Fisher, & Fisher, 2010). If the model test is successful, it would suggest that intervening to increase women's microbicide-relevant information, motivation, and behavioral skills would increase their levels of microbicide use, which could save lives (Mansoor et al., 2014) and assure profits for the pharmaceutical company.

Intervention Research

Finally, elicitation research findings may assist in the design of theory-based, empirically targeted, behavior-change interventions. Here the aim is to collect population-specific data, concerning theoretically specified constructs, so that theory maybe applied in a participant-informed fashion to guide behavior-change interventions. As an initial step in designing a theoretically guided behavior change intervention, and as depicted above, open-ended and closed-ended measures are employed to inform interveners as to where an intervention population is positioned with respect to the constructs that are assumed to influence their behavior and that are to be targeted for change. This type of elicitation research may be considered to be a form of needs assessment to serve as a basis for intervention design (Altschuld & Watkins, 2014).

We cannot understate how essential elicitation research is to the design of effective social psychological interventions to promote behavior change in diverse domains, and the wasted time and energy that might result when researchers seek "short cuts" that involve skipping this step. In health behavior intervention research, for example, elicitation studies may be used to map out pre-intervention levels of a problematic health behavior and identify existing social, personal, and environmental factors that may contribute to the problematic behavior and that need to be addressed to change it. They may also aid in understanding the dynamic interactions among these determinants, and their potential interplay in changing the behavior (J. Fisher & Fisher, 1992; W. Fisher et al., 2014).

Consider the example of intervention researchers who are seeking to understand how to stem an urban HIV outbreak heavily driven by needle and equipment sharing among homeless individuals who inject drugs. Researchers could recruit outreach workers who have daily contact with members of this population to conduct individual open-ended interviews with a sample of these persons at a street-level drop in center in the urban area. Interview foci could be completely bottom-up and participant driven: “Could you tell me about when you use needles and works without sharing them? Could you tell me about when you share needles and works? Why might you do one, or the other?” Individual interviews could also involve the elicitation of data based on the constructs of a theory (e.g., the severity, vulnerability, and barriers to preventive behavior constructs specified in the Health Belief Model, Rosenstock, 1990; beliefs, attitudes, and norms concerning needle sharing, as discussed by Fishbein & Ajzen, 1975). A combination of open-ended questions followed by theoretically based, participant-informed responses to structured questions could be used in a comprehensive elicitation research approach that remains sensitive to participant response burden. Whether involving participants’ descriptions of their behavior and motivation (“I share with my friends when I can’t get a needle”), or their expectations (e.g., “What are the advantages and disadvantages of sharing?”), the voice of “experts” is sought to provide foundational information for understanding and changing risky behavior. Quantitative elicitation research concerning base-rates of the health risk behavior, prevalence of HIV in the intervention target population, maps of the social networks involved in needle sharing, and geographical location of needle sharing settings would also be essential participant-informed data for understanding and guiding intervention efforts. Without this information, interveners would be, to a greater or lesser extent, “shooting in the dark.”

Elaboration of the Four Research Aims

Descriptive Research

Out of a desire to improve our understanding of the causes, consequences, and social significance of attitudes, beliefs, and behaviors, researchers have employed a range of descriptive methods, each of which relies to varying degrees and in varying ways on elicitation methods. These methods include but are not limited to the use of open-ended questions (Creswell, 2003), focus groups (Morgan, 1998), semi-structured individual interviews (Erickson & Stull, 1998), and highly structured interviews (Ajzen & Fishbein, 1980). Here we review each of these methods, drawing attention in the end to how they might be combined.

Open-Ended Questions

A simple way to gain participant-informed understanding of an attitudinal or behavior domain is to simply ask a participant to provide you with such an

understanding. A recent and illustrative example of the use of open-ended elicitation research questions was reported by Kohut, Fisher, and Campbell (2017) in a study of the impact of pornography on the couple relationship. In their review of research on this topic, these researchers observed that research has generally found that pornography harms the sexual and romantic aspects of couple relationships (and see Montgomery-Graham, Kohut, Fisher, & Campbell, 2015). However, these researchers also observed that such findings have been obtained by investigator-imposed, top-down research assessments, which tended to selectively assess harmful effects of pornography on the couple relationship.

Breaking from the top-down methods that dominated this research tradition, Kohut and colleagues (2016) asked simple open-ended questions on the topic of pornography (e.g., “What effect, if any, has pornography (defined as pictures, text, or audio descriptions of nudity and sexual interaction) use had on your couple relationship?” “What negative effects, if any, has pornography use had on your couple relationship?” and “What positive effects, if any, has pornography use had on your relationship?”). They then applied systematic methods of thematic analysis of these responses (see Braun & Clarke, 2006) to elucidate participant perspectives on the impact pornography might have had on their relationship.

The most common response to these open-ended questions, by a very wide margin, was that pornography use had “no effect” on respondents’ couple or sexual relationship. More intriguing, however, was that a large number of positive effects were also spontaneously reported, including reports of increased knowledge of sexual techniques, increased couple intimacy, increased sexual autonomy, and the like. In fact, responses emphasizing positive effects were considerably more common than were reports emphasizing negative effects. Results also indicated that, in contrast to common belief, pornography use was not always a solitary male vice. It was relatively common among women and was also an occasionally shared couple activity. The contrast between these results and those using top-down methods was striking and it arguably provided new and ecologically valid descriptive information.

Focus Groups

Focus groups involve a discussion leader and a small sample of members of a population of interest (typically 5–8 participants; see Morgan, 1998). The discussion group leader asks a series of loosely structured questions concerning a given topic, and participants are told that there are no “correct” answers and that a diversity of views is common and in fact sought. This is another method designed to elicit salient perspectives and concepts, and it can provide richer descriptions than might be obtained by open-ended questions, alone. As an example, consider Salisbury and Fisher’s (2014) focus group elicitation research concerning the meaning young adult men and women attach to the occurrence, or nonoccurrence, of female orgasm in heterosexual intercourse. Young women relatively infrequently experience orgasm during heterosexual intercourse, and it was deemed important

to explore how women and men perceive and react to this situation. Perceptions and reactions to the occurrence or nonoccurrence of female orgasm in heterosexual intercourse were elicited in a number of small, female-only and male-only focus groups. Discussion leaders asked a number of general questions related to the issue of interest. Focus group questions included “How important is it for a woman to have an orgasm during intercourse? Why?” “Would it bother a woman when orgasm does not occur? Why?” and “How do males tend to react when their female partner does not orgasm during intercourse?” The complete focus group discussion guide may be found in Salisbury and Fisher (2014).

Themes emerging from the focus group discussions were systematically coded and analyzed by multiple raters independently and then, together, employing Braun and Clarke’s (2006) qualitative analysis methodology. Findings converged on several themes—shared by both young women and young men—concerning the presence or absence of women’s orgasm in heterosexual intercourse. First, young adult men and women spontaneously and consistently reported that women’s orgasm in heterosexual intercourse is *vastly more important to the male partner than to the female partner*. Protection of the male ego emerged as primary and orgasm for the female as secondary, in the spontaneously elicited and consistent view of both sexes. Second, young adult men and women spontaneously reported the belief that men are responsible for “giving” the woman an orgasm and that the woman was only responsible for “receiving” the male’s ministrations. Third, there was consensus that female orgasm occurrence was more important in the context of an ongoing relationship than in the context of a casual sex “hook-up” but again, largely for the male ego’s benefit. The use of focus group discussions guided by semi-structured questions, which were then systematically and thematically analyzed, permitted identification of young women’s and men’s commonly held—and quite possibly problematic—gendered understandings of female orgasm that had not surfaced in other research approaches. An advantage of focus group methods is that they can be adapted to identify factors influencing decisions of a wide range of populations, including under-studied, harder-to-reach populations, where sample size is a limiting constraint and a researcher might not otherwise know where to start in mapping the research process (see Tennille, Solomon, Fishbein, & Blank, 2009).

Semi-Structured Individual Interviews

A potential benefit of focus groups is that group discussion might provoke participants to reflect upon and articulate a wide range of views. A potential downside is that group consensus, or a “follow the leader” phenomenon might emerge and manifest in ways that obscure individual differences and outlier opinions. Much can thus be gained by pursuing interviews of individual participants. To get the most of such interviews and to see the points of convergence and divergence among participants, interviews can be “semi-structured” in that they are guided

by established theory but still “loose” in the sense that interviewers are allowed to follow-up on answers with new questions (some but not all of which might be anticipated and thus also given some structure). Reliance on prior theory to add structure also increases the likelihood that a broad range of the salient, spontaneously expressed underpinnings of attitudes, beliefs, and behavior will be unearthed. In research using such methods, the theory at focus dictates the constructs that should be assessed, but the content of these constructs is supplied by the research participants, through their open-ended discussions.

A contemporary illustration of semi-structured interviewing comes from a study of vaccine hesitancy and vaccine acceptance by Fisher and colleagues (W. Fisher et al., 2016). These investigators explored parental intentions to inoculate their infant with a new vaccine against meningitis B, a rare but potentially catastrophic infection. The study was undertaken within the general context of expert views (often guided by top-down assessments and assumptions) that (a) hesitancy to inoculate one’s infant with a new vaccine is primarily a function of fears of novel side effects that are unique to new vaccines, and that (b) family income would heavily influence the adoption of a new vaccine which was relatively expensive (see Fisher, 2012; Larson, Jarrett, Eckersberger, & Smith, 2014; MacDonald & SAGE Working Group on Vaccine Hesitancy, 2015).

To test the accuracy of expert views, W. Fisher et al. (2016) adopted the structured, participant-informed technique of asking parents of infants several theoretically specified open-ended questions. The theory-based questions assessed parents’ perceptions of the advantages and disadvantages of vaccinating their infant with both the usual infant vaccines and the new vaccine, as well as parental perception of whether significant others would support or oppose these actions. Findings showed quite robustly that perceptions of the advantages and disadvantages of vaccinating one’s infant with the novel meningitis B vaccine were actually very similar to perceptions of the advantages and disadvantages of vaccinating one’s infant with routine infant vaccines, in use for many years. Results also indicated that sources of social support and opposition to routine infant vaccinations were very similar to those to novel infant vaccinations. This finding has potentially important consequences for public health, as it suggests that intervention efforts that target the intuitively assumed parental fears about unknown side effects of novel vaccines are unwarranted and misguided. Interviews further indicated that the cost of the novel vaccine was not particularly important in relation to intentions to vaccinate one’s infant with the new meningococcal B vaccine. This unexpected finding was used to inform physicians who—in separate open-ended question based elicitation research—indicated that they would not offer the vaccine to patients who were thought to be unable to afford it. This is but one time in which the decision to incorporate elicitation methods resulted in surprising findings, that in turn could promote counterintuitive intervention methods (as discussed in greater detail in a later section).

Structured Interviews

A possible misperception of elicitation research is that structure is the enemy. True, it can in some instances impose “top-down” views of the researcher that inadvertently put words in the mouth of participants. This need not be the case, however. In many instances, a researcher might possess sufficient understanding of the phenomenon of interest and the potential relevance of a guiding theory that they structure interviews in such a way that the results provide maximally informative new descriptive information. An example of a novel intervention that was informed through highly structured and standardized, but participant-informed assessment techniques, can be found in research by W. Fisher, Sand, Lewis, and Boroditsky (2000). These investigators applied a novel approach to explore the use of postmenopausal hormone replacement therapy (HRT) that employed the use of open-ended questions to create structured interviews that were then conducted in focus groups, rather than in individual sessions.

This research approach began with the administration of standard, theory-based, open-ended measures of perceived HRT advantages and disadvantages, along with perceptions of social support, to explore women’s intentions to use HRT. These measures were administered at the start of focus group sessions involving 205 pre-, peri-, and postmenopausal women in 33 “Women’s Health Discussion Groups” conducted in small, mid-size, and large cities across different regions in Canada. Rather than simply collecting the data from this questionnaire and stopping there, these researchers instead used women’s own responses as a way of facilitating informative (90 minute), structured group interviews that were moderated by a registered nurse.

These discussions illuminated an unsuspected pattern in the correlates of intentions to initiate HRT that would not have been gleaned from simply analyzing the self-report data. Whereas it was nearly uniformly assumed by healthcare “experts” that avoidance or adoption of HRT was driven by women’s perceptions of the health risks of using this therapy, it turned out that perceived risks of HRT had little to do with women’s intentions. Rather, findings indicated that women’s intentions to utilize HRT were more or less purely a function of their perceptions of the *advantages* of this therapy, and secondarily, a function of their perceptions of social support for doing so. Had the investigators conducted an intuitively driven study guided by the expert view that perceived health risks drive avoidance of hormone replacement therapy, this important result and its important implications for interventions to increase the use of HRT may never have emerged.

Summary

These examples illustrate the range of elicitation techniques one might employ, as well as the creative ways they might be combined to better understand the attitude, beliefs, and behaviors of research participants. These techniques provide

the descriptive foundation underlying the basic and applied research applications considered in the sections that follow.

Early Stage Research

Elicitation methods have historically been used in the development of quantitative inventories that assess both personality and social psychological variables (Robinson, Shaver, & Wrightsman, 1991). Researchers following in these traditions begin by eliciting participant-informed perceptions of the construct under study. This leads to the creation of a pool of items that assess aspects of the construct that were identified through elicitation. Item-selection procedures and evaluation of the measurement properties follow, with the focus at this stage on quantitative metrics (i.e., reliability, validity, and where appropriate, intervention responsiveness).

These stages of questionnaire development will be familiar to many readers, as they are often referenced in pedagogical articles teaching questionnaire development, but we think this sense of familiarity can obscure the potential contribution of elicitation research to novel measures development. In particular, elicitation methods take on considerable urgency and must be pursued with great vigor, when the goal is to develop quantitative inventories that respect the full diversity of a population of interest or when the researcher seeks to quantify attributes of a group or groups that are unfamiliar to them. We thus review one example of focus group methodology that we think highlights this value. This study used elicitation methods to develop a questionnaire that would assess a culture-specific source of stress in a minority immigrant population.

Willgerodt (2003) and colleagues (Willgerodt & Huang, 2004) were interested in assessing how different rates of acculturation in Chinese immigrant families can lead to intergenerational conflict. Such conflict had been identified in earlier research as a contributing factor in high rates of anxiety, depression, and suicide among older, foreign-born family members. However, prior research that might speak to this issue was potentially of limited relevance at best and possibly misleading, at worst. This is because past research relied on assessment tools developed for use among Euro-Americans or by the investigators themselves (Fuligni, 1998), or because it utilized measures designed for a single generations, which might be of limited use in multi-generational assessments (e.g., Lee, Choe, Kim, & Ngo, 2000).

Willgerodt and Huang (2004) thus began their investigation by pursuing elicitation research, with the ultimate goal of developing new measures that were both culturally and generationally sensitive. They recruited 11 Chinese adolescents, 15 first-generation parents, and 13 immigrant grandparents for generation- and gender-specific focus groups. Group discussions were guided by semi-structured questions, including: “What it is like to be a Chinese American family in the United States?”, “What are the things that upset or bother your family?”, and “What are the most positive things about your family?”, with alternative questions

and probes available to interviewers, allowing them to explore unanticipated questions and concerns of participants. A key feature of these interviews was the care with which researchers handled topics that might be culturally or generationally sensitive to members. The researchers were aware that talking about family issues in a group setting can be especially difficult among members of this cultural group, particularly older generations. The focus group discussion leader thus prepared for participants to cope with their discomfort by talking about personal issues “as if” they were the experience of someone else. Thus, the focus group discussion leader consistently employed general, not personal, probes; for example, “Why might this be a problem for *Chinese families*?” (as opposed to “*your family*”).

Participant responses in this culturally sensitive focus group approach surfaced two broad sources of intergenerational conflict, unlikely to have been uncovered using top-down, close-ended, investigator imposed assessments. One source of conflict was related to intergenerational and intercultural conflicts that emerge in daily life. For example, focus group participants said the following: “Non-Chinese kids get more free time than me” (adolescents); “My relationship with my husband is different now that we are in the United States” (parents); and “I am a burden to my children” (grandparents). Another source of conflict was the divergence of Chinese and Western value systems as a function of generational differences in rates of assimilation, which led to mistrust and confusion. Illustrating this theme, for example, focus group participants said the following: “I’m not allowed to argue with my parents” (adolescents); “My children are disrespectful to their grandparents” (parents); and “My grandchildren talk back to me” (grandparents).

The goal was then to use these rich qualitative descriptions to develop a new, quantitative, inventory. Here, too, elicitation methods assisted researchers. Focus group responses describing a conflict between two or more family members were delineated and identified by the researchers as a content domain, and phrases and terms were marked so that the language of the participants could be preserved in development of a quantitative assessment of family conflict in this cultural group. As many items as possible were generated from the focus group data for this initial instrument development phase, and then a second round of generation- and gender-specific focus groups with different members of the population was conducted to elicit feedback on the questionnaire items. After responding to the items in the questionnaire, second round participants were asked the following questions to ensure comprehension and construct validity “What was it like to fill out the questionnaire?”, “Were there any items you had difficulty understanding or answering?”, “Were there any items that offended you or to which you did not want to respond?”, and “Does the instrument cover the range of problems that Chinese families experience? Is there anything missing?” Revisions were made to the instrument if more than two participants raised similar concerns.

As this example illustrates, the commitment to cultural and generational diversity in elicitation research—dictated by the assessment objectives of a particular area of study—is essential when developing culturally valid quantitative

assessments of attitudes, beliefs or behaviors. This research program illustrates the way that elicitation research can ground development of a new measure in the conceptual space, perceptions, and beliefs of “the experts”—the thinking and acting individuals among whom the measure is to be used to understand and predict attitudes and behavior—even in cases in which the researcher enters into the research question with an impoverished understanding of the groups they wish to study (and see Willgerodt, 2003).

Theory Tests

Elicitation research also is often the launch pad for developing new theoretical understandings of psychological phenomena. Research in the tradition of the Theory of Reasoned Action provides a clear and well-structured example of this approach (Fishbein, 1967; Fishbein & Ajzen, 1975, Ajzen & Fishbein, 1980). Readers working from other theoretical frameworks might question the relevance of this theory to their own work in some cases, but we see great value in investigators developing a full understanding of this approach, in order to learn from this example how to build and test theory via elicitation methods.

According to the Theory of Reasoned Action, behavior (B) occurs as a function of an individual’s behavioral intention (BI) to perform an act. BI, in turn, is theorized to be a function of attitudes toward performing the act (Aact) and/or subjective norms (SN) concerning perceived social support for performing the act in question. The basic psychological underpinnings of Aact consist of salient perceptions of the outcomes of the act (B_i) multiplied by evaluations (e_i) of these outcomes. The basic psychological underpinnings of SN involve normative beliefs (NB_j) or perceptions of specific referent others’ wishes concerning performance of the behavior, and motivation to comply (Mc_j) with their wishes. Thus, $B \sim BI = Aact_{w_1} + SN_{w_2}$ where w_1 and w_2 are empirically determined regression weights and $Aact = \sum Bie_i$ and $SN = \sum NB_jMc_j$.

Elicitation research is essential to testing the constructs and relationships of the Theory of Reasoned Action (Ajzen & Fishbein, 1980) as well as to build specific theories about psychological dynamics in different behavioral domains. Specifically, in work using the Theory of Reasoned Action, open-ended elicitation research is conducted to identify salient population-specific beliefs and referent others that are the foundations of the theory’s Aact and SN constructs. Using structured elicitation research procedures, a representative subsample of the population of interest is asked “What are the advantages, if any, of (the behavior in question)?” and “What are the disadvantages, if any, of (the behavior in question)?” A simple frequency count of responses is used to identify salient beliefs (Bi) about the outcomes of performing the behavior in the population at focus. The subsample of the population is also asked, “What persons or groups would approve of you engaging in (the behavior in question)?” to identify specific referent others (NB_j) who would support the behavior. Generally, a cluster of five

to seven perceived outcomes of performing a behavior and a similar number of referent others are elicited.

Salient elicited beliefs and referent others are subsequently used as a basis for quantitative assessment of the underpinnings of Aact ($\sum B_{iei}$) and SN ($\sum NB_{ij}Mcj$) and testing the theory's predicted relationships. Generally, behavior proves to be a function of intentions, intentions prove to be a function of attitudes and/or norms concerning an act, and Aact is moderately to strongly correlated with $\sum B_{iei}$ and SN is moderately to strongly correlated with $\sum NB_{ij}Mcj$. Elicitation research thus is employed to create quantitative measures of the basic psychological underpinnings of attitudes and norms. This information is highly informative of the basis of attitudes and norms concerning a particular behavior and affords the ability to test the theory in relation to measures that are valid in a given population, with respect to a particular behavior. What is more, intervention efforts to promote behavior change can target elicited basic underpinnings of attitudes and norms in order to influence intentions and ultimately behavior. Specific step-by-step procedures for conducting elicitation research within the Theory of Reasoned Action are provided by Ajzen and Fishbein (1980), which can be used to test theorized relationships of its theoretical constructs.

This structured approach to theory development differs dramatically from various "top-down," inductive approaches that will be more familiar to many researchers, but it can lead to new understandings that become the basis for generating novel theories that are grounded in the understandings generated by participants, rather than ideas generated by researchers.

Interventions

We have provided examples of how elicitation research can point investigators in the direction of novel, sometimes counterintuitive understandings of the determinants and dynamics of behavior and ultimately, to social psychological interventions to change behavior. We close by fleshing out the process by which researchers might utilize these methods to design and empirically validate ecologically valid, theory-based interventions. On this topic, there is a large foundation from which one might draw. One of the great contributions of social psychology has been the development of behavior change models that can guide psychological intervention, most of which place heavy emphasis on the focused use of elicitation research to guide researchers at different stages of inquiry (e.g., Rosenstock, 1990; Ajzen & Fishbein, 1980; J. Fisher & Fisher, 2000). Since two of the authors (WAF and JDF) have spent much of their careers developing and testing the IMB model of health behavior change and building and evaluating behavior change interventions based on it (J. Fisher & Fisher, 1992; W. Fisher et al., 2014), we draw most heavily on this research tradition to illustrate the role of theory-guided elicitation research in the design of effective behavior change interventions. However, the specifics of this approach can be adapted for a wide

variety of interventions, in a range of different content domains, based on any number of different theories.

As noted earlier, the IMB model holds that an individual's store of health behavior information, their personal and social motivation to act on this information, and their objective behavioral skills and sense of self-efficacy for acting effectively, are fundamental determinants of the performance of health behaviors. In addition, the IMB model asserts that, generally speaking, health behavior information and health behavior motivation will activate health behavior behavioral skills that are applied in the performance of health behaviors, *per se*. When health behaviors do not involve the performance of complex or novel behavioral skills there may be direct relationships of information and motivation with health behavior (see Figure 3.1). The IMB model specifies a three-stage approach to behavior change intervention, which might be adapted by researchers utilizing the IMB model or pursuing other theories, in other content domains. A researcher orienting around a different theoretical framework in a different applied domain could adapt these same three phases, but the first task would be to articulate the theorized drivers of the behavior at focus, presumably based from a foundation of prior theory and research.

In the first phase, *assessment and planning*, elicitation research is employed to identify deficits in health-related information, motivational obstacles, and behavioral skills limitations, each of which may contribute to measured levels of a health risk behavior. Elicitation research can also tell us what information, motivation, and behavioral skills strengths exist in the intervention target population that can be mobilized in facilitating health behavior change. In the second

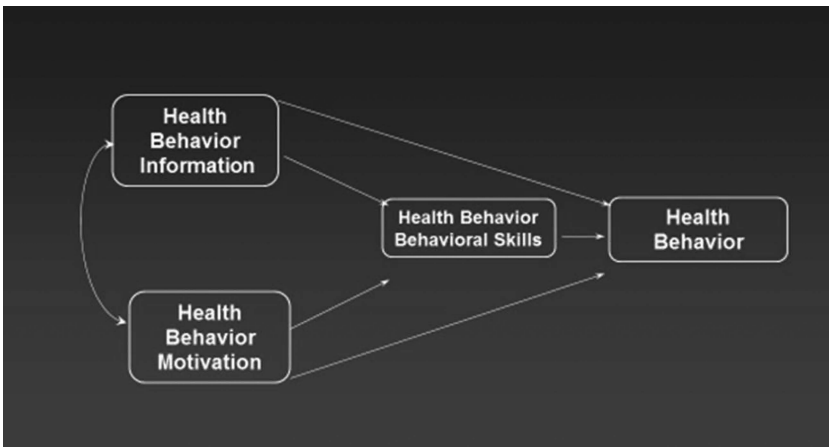


FIGURE 3.1 The Information-Motivation-Behavioral Skills Model of Health Behavior

Source: Fisher, J. D., & Fisher, W. A. (1992). Changing AIDS risk behavior. *Psychological Bulletin*, 111, 455–474.

phase, *intervention*, IMB model-based interventions are specifically tailored to address elicitation research findings concerning the information deficits, motivational obstacles, and behavioral skills limitations that are linked with a health risk behavior in the intervention target population. In the last phase, *evaluation*, rigorous research is conducted to determine if the intervention has had the desired effects on health related information, motivation, behavioral skills, and health behavior *per se*.

Critical to the success of any intervention utilizing this or related methods is that elicitation data collected must be keyed to the constructs, populations, and behaviors at focus. In addition to qualitative and quantitative assessment approaches discussed earlier (e.g., open-ended questions, focus group discussions, semi-structured or structured assessments), any additional assessment approaches should be guided by one's theoretical approach and intervention objectives. Close-ended quantitative assessments, some of which may be developed on the basis of elicitation research findings and some of which can be found from the literature, can be efficiently be employed in elicitation research. Direct observation, guided by one's theory and intervention objectives (e.g., observation of role played behavioral skills in condom negotiation) and measures that are uniquely relevant to the population and the health-risk behavior at focus (e.g., information about safer and risky sex behavior among men who have sex among men of color) can also contribute to elicitation and intervention efforts. Unobtrusive and indirect measures (e.g., reports of hospitalization for distinctively needle-sharing related diagnoses) and direct physical measures (e.g., DNA studies of discarded needles to determine singularity or multiplicity of users) can all add to baseline and intervention outcome data concerning risky and safer behavior.

In work using the IMB model, quantitative measures can then be created or identified to investigate the assumptions of the model in cross-sectional structural modeling tests of the relationships of health behavior-related information, motivation, behavioral skills, and behavior in a given setting. Such research can provide critical guidance concerning which elements—information, and/or motivation, and/or behavioral skills—may contribute most strongly to a specific health risk behavior and which should therefore be emphasized in an intervention to change it. Moreover, within the IMB model's motivation construct, analysis of quantitative elicitation data can determine whether motivation to perform a behavior is driven predominantly by a person's attitudes, social norms, or both (Fishbein & Ajzen, 1975, 2010). These data provide guidance with respect to whether attitude change, social normative support, or both should be emphasized to change the behavior at focus. Close-ended measures which are created or identified to assess information, motivation, and behavioral skills and used for the purpose of model-testing can also be used in intervention outcome research to assess whether the intervention has, in fact, produced pre- to post-intervention changes in the theoretical constructs which are assumed to be responsible for behavior change *per se*.

In addition to providing insight on determinants of health risk behavior, elicitation data can provide additional content that is essential for intervention design and implementation. For example, we have used elicitation research procedures to help us to identify what type of intervener the participants may respond to most favorably, what tone the intervention should take, whether it should consist, for example, of one 6-hour session or multiple shorter sessions, what type of intervention will be most congenial to both the interveners and the intervention recipients, how a pilot intervention must be modified before going to scale, what types of incentives should be offered for participation, and whether providing child care and refreshments would be necessary to ensure participant attendance in low resource settings, all of which have proved critical to ensure favorable intervention outcomes. In specific instances, we have conducted elicitation research with both the target population which is directly involved (e.g., individuals living with HIV in South Africa), as well as those who provide their medical care (nurses and physicians at community health centers) in order to comprehensively understand critical aspects of intervention content and context for a specific population (J. Fisher et al., 2014). In terms of choosing specific intervention content, incentives, and structure, we have long espoused involving behavioral scientists, healthcare providers, the actual interveners, and the population being intervened with as true collaborators in evaluating the elicitation findings and in deciding on their implications for intervention content and structure (J. Fisher, Cornman, Norton, & Fisher, 2006).

It is beyond the scope of the current chapter to outline all of the examples of research we have developed, guided by elicitation methods, but we have taken such interventions to scale to have dramatic influences on not just health-related information, motivation, and skills, but also on health behaviors that can sometimes make the difference between life and death (e.g., J. Fisher et al., 1996; W. Fisher et al., 2014). In each instance, elicitation research has been essential for capturing the target population-specific content of theoretical constructs driving behavior and creating elicitation-research based interventions to promote needed change in these theoretical constructs and the behaviors they influence. Elicitation research methodologies are diverse, must be sensitive to the population, health behavior, setting, and construct(s) at focus, and must provide participant-informed data concerning each of these issues.

Conclusion

A fundamental rationale for elicitation research involves the value of identifying salient, spontaneously accessible factors that are perceived by thinking and acting individuals as important influences on their attitudes, beliefs, and behavior. Together with this, our rationale for elicitation research alerts researchers to the potential danger of providing only top-down, investigator-imposed, close-ended opportunities for participants to respond with respect to presumed determinants

of their attitudes, beliefs, and behavior. We have explicated a conceptual rationale for the importance of appropriate elicitation research in connection with achieving participant-informed understandings of attitudes, beliefs, and behavior. We have also identified elicitation research as essential for the development of valid quantitative measures and for identification of salient content of theoretically specified constructs for theory testing purposes. Finally, we have outlined the importance of elicitation research for informing effective intervention research. We encourage researchers to implement such research and realize its benefits.

References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.
- Altschuld, J. W., & Watkins, R. (2014). A primer on needs assessment: More than 40 years of research and practice. *New Directions for Evaluation*, 144, 5–18.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Creswell, J. W. (2003). A framework for design. In C. D. Laughton (Ed.), *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage Publications.
- Erickson, K. C., & Stull, D. D. (1998). *Doing team ethnography: Warnings and advice*. Thousand Oaks, CA: Sage Publications. doi:10.4135/9781412983976.
- Ferrer, R. A., Morrow, K. M., Fisher, W. A., & Fisher, J. D. (2010). Toward an information-motivation-behavioral skills model of microbicide adherence in clinical trials. *AIDS Care*, 22(8), 997–1005.
- Fishbein, M. (Ed.). (1967). *Readings in attitude theory and measurement*. New York, NY: Wiley-Blackwell.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: Introduction to theory and research*. Reading, MA: Addison Wesley.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Taylor & Francis.
- Fisher, J. D., Cornman, D. H., Norton, W. E., & Fisher, W. A. (2006). Involving behavioral scientists, health care providers, and HIV-infected patients as collaborators in theory-based HIV prevention and antiretroviral adherence interventions. *JAIDS*, 43(Supplement 1), S10–S17. doi:10.1097/01.qai.0000248335.90190.f9
- Fisher, J. D., Cornman, D. H., Shuper, P. A., Christie, S., Pillay, S., MacDonald, S., . . . for the SA Options Team. (2014). HIV prevention counseling intervention delivered during routine clinical care reduces HIV transmission risk behavior in HIV-infected South Africans receiving antiretroviral therapy: The Izindela Zokuphila/Options for Health Randomized Trial. *JAIDS*, 67(5), 499–507. doi:10.1097/QAI.0000000000000348
- Fisher, J. D., & Fisher, W. A. (1992). Changing AIDS risk behavior. *Psychological Bulletin*, 111, 455–474.
- Fisher, J. D., Fisher, W. A. (2000). Theoretical approaches to individual-level change in HIV risk behavior. In J. Peterson & R. DiClemente (Eds.), *Handbook of HIV prevention* (pp. 3–55). New York, NY: Kluwer Academic and Plenum Press.

- Fisher, J. D., Fisher, W. A., Misovich, S. J., Kimble, D. L., & Malloy, T. E. (1996). Changing AIDS risk behavior: Effects of an intervention emphasizing AIDS risk reduction information, motivation, and behavioral skills in a college student population. *Health Psychology, 15*(2), 114–123. doi:10.1037/0278-6133.15.2.114
- Fisher, W. A., (2012). Understanding human papillomavirus vaccine uptake. *Vaccine, 30S*, F149–F156. doi:10.1016/j.vaccine.2012.04.107
- Fisher, W. A., Bettinger, J., Gilca, V., Sampalis, J., Brown, V., Yaremko, J., & Mansi, J. (2014, May). *Understanding the impact of approved but unfunded vaccine status on parental acceptance of a novel meningococcal serogroup B vaccine for infants*. European Society of Pediatric Infectious Diseases, Dublin, Ireland.
- Fisher, W. A., Fisher, J. D., & Shuper, P. A. (2014). Social psychology and the fight against AIDS: An Information–Motivation–Behavioral Skills model for the prediction and promotion of health behavior change. *Advances in Experimental Social Psychology, 50*, 105–193.
- Fisher, W. A., Jannini, E. A., Gruenwald, I., Lev-Sagie, A., Pyke, R., Revicki, D., & Reisman, Y. (2016). *Standards for clinical trials in male and female sexual dysfunction*. International Consultation on Sexual Medicine, Madrid, Spain.
- Fisher, W. A., Sand, M., Lewis, W., & Boroditsky, R. (2000). Canadian menopause study—I: Understanding women's intentions to utilize hormone replacement therapy. *Maturitas, 37*, 1–14.
- Fulgini, A. J. (1998). Authority, autonomy, and parent–adolescent conflict and cohesion: A study of adolescents from Mexican, Chinese, Filipino, and European backgrounds. *Developmental Psychology, 34*, 782–792.
- Kohut, T., Fisher, W. A., & Campbell, L. (2017). Perceived effects of pornography on the couple relationship: Initial findings of open-ended, participant-informed, “bottom-up” research. *Archives of Sexual Behavior, 46*, 585.
- Larson, H. J., Jarrett, C., Eckersberger, E., & Smith, D. (2014). Understanding vaccine hesitance around vaccines and vaccination from a global perspective: A systematic review of published literature, 2007–2012. *Vaccine, 33*(19), 2150–2159. doi:10.1016/j.vaccine.2014.01.081
- Lee, R. M., Choe, J., Kim, G., & Ngo, V. (2000). Construction of the Asian American family conflicts scale. *Journal of Counseling Psychology, 47*, 211–222.
- MacDonald, N., & SAGE Working Group on Vaccine Hesitancy. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine, 33*(34), 4161–4164. doi:10.1016/j.vaccine.2015.04.036
- Mansoor, L. E., Abdool Karim, Q., Werner, L., Madlala, B., Mgcobo, N., Cornman, D. H., . . . Abdool Karim, S. S. (2014). Impact of an adherence intervention on the effectiveness of Tenofovir Gel in the CAPRISA 004 Trial. *AIDS and Behavior, 18*(5), 841–848. doi:10.1007/s10461-014-0752-9
- Montgomery-Graham, S., Kohut, T., Fisher, W., & Campbell, L. (2015). How the popular media rushes to judgment about pornography and relationship while research lags behind. *The Canadian Journal of Human Sexuality, 24*(3), 243–256.
- Morgan, D. L. (1998). *The focus group guidebook*. Thousand Oaks, CA: Sage Publications.
- Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: Sage Publications.
- Opel, D. J., Mangione-Smith, R., Taylor, J. A., Korfiatis, C., Wiese, C., & Martin, D. P. (2011). Development of a survey to identify vaccine-hesitant parents: The parent attitudes about childhood vaccines survey. *Human Vaccines, 7*(4), 419–425.
- Robinson, J. P. Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes*. San Diego: Academic Press.

- Rosenstock, I. M. (1990). The health belief model: Explaining health behavior through expectancies. In K. Glanz & B. K. Reimer (Eds.), *Health behavior and health education: Theory, research, and practice* (pp. 39–62). San Francisco, CA: Jossey-Bass.
- Russell, H. R. (2006). *Research methods in anthropology: Qualitative and quantitative approaches* (4rth ed.). Lanham, MD: Altimira Press.
- Sadaf, A., Richards, J. L., Glanz, J., Salmond, D. A., & Omer, S. B. (2013). A systematic review of interventions for reducing parental vaccine refusal and vaccine hesitancy. *Vaccine*, 31(40), 4293–4304.
- Salisbury, C. M. A., & Fisher, W. A. (2014). “Did you come?” A qualitative exploration of gender differences in beliefs, experiences, and concerns regarding female orgasm occurrence during heterosexual sexual interactions. *Journal of Sex Research*, 51(6), 616–631.
- Shumlich, E. J. (2017). *Attitudes and behaviours surrounding consent in a university population*. Manuscript in preparation. Department of Psychology, Western University, London, Canada.
- Tennille, J., Solomon, P., Fishbein, M., & Blank, M. (2009). Elicitation of cognitions related to HIV risk behaviors in persons with mental illnesses: Implications for prevention. *Psychiatric Rehabilitation Journal*, 33(1), 32–37.
- Willgerodt, M. A. (2003). Using focus groups to develop culturally relevant instruments. *Western Journal of Nursing Research*, 25(7), 798–814.
- Willgerodt, M. A., & Huang, H. Y. (2004). The Asian American family conflict assessment tool: Pilot findings [Abstract]. *Communicating Nursing Research*, 37, 251.

4

PSYCHOBIOLOGICAL MEASUREMENT

Peggy M. Zoccola

Introduction

There has been a longstanding interest in mind-body phenomenon in psychology. Indeed, psychologist William James' (1884) influential theory on emotion unequivocally linked physiological changes to emotions. Although his theory has been modified in the many ensuing years, James' work was among the first to suggest that measuring patterns of arousal would allow psychologists to index emotional states. Nearly a century later, multiple texts on psychophysiological methods are available for students and researchers. Together with many technological advances in understanding and quantifying human anatomy and physiology in more nuanced ways, we now see burgeoning interest and rigorous psychobiological research among social scientists.

Some social scientists look to psychobiological assessment as objective measures to replace or complement self-report measures. Just as implicit or behavioral observations may be used to circumvent the limitations of introspection or self-reported psychological processes, so is the hope for biological measures. The collection of biological data alongside self-report or behavioral measures may confirm participant reports or provide converging evidence of a particular phenomenon, such as arousal or threat perception. Researchers might also turn to biological measurement to identify underlying mechanisms that contribute to psychological processes. For instance, one might test whether specific hormone secretions precede particular social behaviors or psychological phenomenon. Still others might focus on biological measures as consequences of psychological or behavioral phenomenon. For example, does psychosocial stress lead to greater susceptibility to the common cold?

Although all such approaches in the context of psychobiological assessment are addressed in this chapter, it is important to note that psychobiological data are often collected in a way that makes it difficult to tease apart cause and effect. In some instances, it seems clear that psychological processes are driving physiological changes; other times physiological processes seem to be leading to psychological responses. For those who turn to psychobiological assessment for alternatives to self-report or behavioral data, the cause and effect may not matter so much, as long as the presence or absence of a correlation can be determined. For others, clearly identifying cause and consequence is critical for confirming and refining theory.

This chapter begins by providing basic anatomical and physiological information about each of three pertinent bodily systems: autonomic, endocrine, and immune. Along with each system overview are procedures for assessing specific system parameters or biomarkers and select examples of research to illustrate topics that can be addressed with such methods. The chapter then reviews general methodological, practical, and analytical issues one should consider and closes with concluding remarks, sharing thoughts on next steps for this research area, and by providing additional resources for the reader.

Autonomic Nervous System

Anatomy and Physiology

The autonomic nervous system (ANS) is a major division of the peripheral nervous system. It regulates critical peripheral bodily functions that are generally thought to be outside of our conscious control. ANS nerve fibers transmit signals between the brain and spinal cord and nearly every other tissue and vessel throughout the body's periphery, including smooth and cardiac muscle, as well as glands, and the gastrointestinal track. Blood pressure, heart rate, digestion, body temperature, respiration, and inflammation are just some of the many functions under ANS control (Andreassi, 2007).

The ANS is further divided into two functionally distinct branches: the parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS). Broadly speaking, SNS activity facilitates bodily processes associated with energy mobilization, such as those that help an individual deal with emergency, or "fight-or-flight" types of situations. Consider how your body might respond if a fire alarm goes off in your building. As described by physiologist Walter Cannon (1915), one might experience

the contraction of blood vessels with resulting pallor, the pouring out of "cold sweat," the stopping of saliva-flow so that "the tongue cleaves to the roof of the mouth," the dilation of pupils, the rising of the hairs, the rapid

beating of the heart, the hurried respiration, the trembling and twitching of the muscles.

(p. 3)

Such cardinal features of sympathetic arousal divert blood flow and energy to major muscles throughout the body, and are understood to enhance survival of physical threats to the self.

The other branch of the ANS, the PNS, exerts actions in opposition to the SNS and leads to vegetative and restorative functions. As such, PNS activity is sometimes characterized as “rest-and-digest” or “feed-and-breed,” to exemplify some of the functions this system serves. Consider how your body might respond after you finish your extra helping on Thanksgiving and you are about to settle in for the evening. The PNS stimulates the digestive system (salivary flow, peristalsis in the intestines), and it promotes rest, relaxation, and sleep. PNS neurons also innervate the smooth muscles of the iris and lead to constriction of the pupil. A particularly important cranial nerve of the ANS, the vagus nerve, transmits signals between the brain and periphery, including control of the heart. When activated, it slows the beating of the heart. Of particular relevance to those interested in psychobiological measures, the vagus has both “afferent” and “efferent” pathways. In other words, because messages flow both from the central nervous system to peripheral organs (afferent) and away from peripheral organs to the central nervous system (efferent), the body and brain can influence one another as well as coordinate a response to external changes in the physical and social environment.

Combined, the SNS and PNS are integrated subsystems that together facilitate homeostasis, or the underlying processes of maintaining a relatively stable internal environment essential for survival. As integrative systems, the two can at times promote the same responses, through opposing mechanisms. For instance, signs of “fight-or-flight” activity, such as increased beating of the heart, may be influenced by both stimulation of the SNS and a withdrawal of PNS activity. As a result, there has been some interest in determining the balance of ANS activity, or the degree to which one ANS branch is more activated than the other. To properly assess the complex, interdependent aspects of physiological systems it is necessary to use strong measurement tools—the topic of this chapter.

Assessment

A variety of parameters are used to quantify SNS and PNS activity. Peripheral ANS parameters are derived most commonly from observed cardiovascular and electrodermal activity. See Table 4.1 for a summary of ANS indices in social psychological research. As noted earlier, the SNS and PNS systems work in conjunction with one another, and each parameter varies in the degree to which it reflects mostly SNS activity, PNS, or both.

TABLE 4.1 Autonomic Nervous System Indices in Social Psychological Research

<i>Parameter</i>	<i>Description</i>	<i>Sympathetic Response</i>	<i>Parasympathetic Response</i>	<i>Instrumentation</i>
Cardiovascular				
HR	Speed or rate of heart contractions in beats per minute.	Increase	Decrease	Electrocardiography, pulse meter
SBP	Systolic blood pressure occurs when the heart contracts (i.e., maximum pressure in artery) in millimeters of mercury.	Increase	Decrease	Blood pressure monitor
DBP	Diastolic blood pressure occurs in between contractions of the heart (i.e., minimum pressure in artery) in millimeters of mercury.	Increase	Decrease	Blood pressure monitor
MAP	Mean arterial pressure, typically derived with the following formula: $1/3(\text{SBP}-\text{DBP}) + \text{DBP} = \text{MAP}$.	Increase	Decrease	Blood pressure monitor
CO	Cardiac output reflects amount of blood pumped by the heart in liters per minute ($\text{HR} \times \text{stroke volume}$).	Increase	Decrease	Impedance cardiography and electrocardiography
TPR	Total peripheral resistance, or net constriction in the arterial system ($\text{MAP} \times 80 / \text{CO}$) in resistance units.	Increase	Decrease	Impedance cardiography and electrocardiography
PEP	Pre-ejection period is the amount of time between depolarization in the left ventricle and the opening of the aortic valve in milliseconds. PEP is also referred to as ventricular contractility.	Decrease		Impedance cardiography and electrocardiography
RMSSD	Root mean square of successive differences is a time-domain measure of variability in heart rate measured in milliseconds. RMSSD is calculated by taking the square root of the mean squared difference of successive normal-to-normal heart beats over a specified time period.		Increase	Electrocardiography

RSA or HF-HRV	Respiratory sinus arrhythmia, or high frequency HR variability, refers to the beat-to-beat alterations in heart rate occurring in the high frequency range (0.15-0.40 Hz) as measured by power spectral analysis. HF-HRV is often reported in squared millisecond units.	Increase	Electrocardiography, select pulse meters
Electrodermal			
SCL	Skin conductance level is the tonic level of the skin's electrical conductivity in microSiemens.	Increase	Electrodermal activity recording system
SCR	Skin conductance response is the phasic change in skin's electrical conductivity in response to a stimulus in micro-Siemens.	Increase	Electrodermal activity recording system
NS-SCR	Non-specific skin conductance response is the phasic change in skin's electrical conductivity that occurs in the absence of an identifiable stimulus in microSiemens.	Increase	Electrodermal activity recording system
Other			
RR	Respiratory rate is the number of breaths per minute.	Increase	Respiratory belt
Pupil size	Diameter of pupil, measured in millimeters.	Decrease Increase	Eye-tracking system

Cardiovascular indices

Heart rate, or the number of times the ventricles of the heart contract in a given period of time (usually a minute), has been widely used to capture ANS activity in psychological research. For example, many researchers have quantified phasic changes (i.e., pre-post increases) in HR to index psychological stress induced in the laboratory. One reason for HR's popularity is likely due at least in part to its relative ease of assessment. You can't get much more low-tech than simply placing two fingers on a wrist to count pulse rate! However, since both sympathetic and parasympathetic inputs control the rate and variability of HR (Saul, 1990), researchers have turned more recently to examine SNS and PNS activity separately.

Beyond HR, additional cardiac parameters allow for a more nuanced assessment of ANS activity (see Table 4.1). For example, high frequency HR reflects cardiac input from the vagus nerve, and thus can be used as an index of PNS activity. HR variability measures reflect the degree to which intervals between ventricular contractions vary. Variability in HR can be calculated in a multiple of ways (for review, see Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996).

Electrocardiography (ECG or EKG) and impedance cardiography are commonly used to derive ANS measures. ECG recordings capture multiple electrical inflections of the heart's electrical activity (e.g., QRS complexes), which can later be used to derive multiple cardiac parameters. Impedance cardiography recordings capture changes in blood flow to estimate volume of blood ejected from the heart and the timing of the opening and closing of the heart's valves. In a typical psychophysiology research laboratory, a three-lead ECG with self-adhering disposable spot electrodes are placed on a participant to acquire the cardiac signal. Electrodes are typically placed on the limbs or torso to create an imaginary triangle with the heart at the center (Andreassi, 2007). The exact configuration of the electrodes can vary somewhat to accommodate particular study constraints. For example, if movement of the limbs is anticipated during the laboratory procedures, the researcher may choose to place the leads on the torso. To acquire impedance data alongside ECG, additional spot or band/tape electrodes are necessary. With band electrodes, two may be placed to encircle the base of the neck and an additional two are placed near and below the base of the sternum to encircle the torso (for additional details, see Blascovitch, Vanman, Mendes, & Dickerson, 2011).

If you have been keeping count, we are now up to seven electrodes placed on the participant—with just as many cables fastened to them! Such tethering is important to take into account when designing study protocols and providing pre-visit information to participants. For instance, it would be a good idea to instruct participants to wear comfortable, two-piece clothing, leave their jewelry at home, and use the restroom before being hooked up to the physiological recording equipment.

Although wired recording systems remain normative, commercially available ambulatory devices can be used inside and outside the research laboratory. Indeed, modestly priced ambulatory HR monitoring devices have shown to have excellent test-retest reliability and agreement with ECG-derived HR variability in healthy participants (Weippert et al., 2010; Williams et al., 2016; for ambulatory cardiac measures, see Walsh, Topol, & Steinhubl, 2014). Fully equipped psychophysiological systems that amplify and record ANS-related activity can cost a few to tens of thousands of dollars, so purchasing decisions are not to be taken lightly. Consider the applications you and your collaborators may want now and in many years from the time of the purchase. Smaller systems or some ambulatory devices with more limited assessment options can be much lower in cost. Ambulatory devices worn in the laboratory might also be desirable given that traditional systems tether participants to equipment with multiple sensors and cords. It is always a good idea to consult with other researchers who have used the equipment under your consideration or to try it out yourself before making a large financial commitment.

Electrodermal indices

Electrodermal activity refers to all electrical phenomena in the skin. Of most relevance to the assessment of the ANS, is the electrical activity of eccrine sweat glands. Eccrine sweat glands are innervated by cholinergic fibers of the SNS, and thus electrodermal activity is commonly used to quantify sympathetic activity. Although the primary function of sweat glands are to regulate body temperature, or thermoregulation, they are also responsive to psychologically relevant stimuli. Some of the early electrodermal activity work focused “arousal,” which was quantified by using a galvanometer to measure changes in skin conductance in response to a variety of emotional stimuli (Neumann & Blanton, 1970).

There are two main types of methods for collecting electrodermal activity with commercially available instruments and software (Andreassi, 2007): skin conductance and skin potential. For skin conductance, a small electrical current is passed through the skin with two (bipolar) electrodes placed on the skin surface. The resistance to that current, or its reciprocal skin conductance, is measured. Skin conductance levels (SCL) refer to “baseline” values at a given time period. Skin conductance/resistance responses refer to momentary fluctuations, typically in response to a stimuli introduced by the research. The skin conductance response (SCR) typically appears one to three seconds after a stimulus is presented. It is important to note that SCRs also occur spontaneously, or in the absence of the researcher’s stimulus. Such a response is referred to as a “non-specific” SCR (NS-SCR). Researchers will commonly measure the number of NS-SCRs in a period of time (i.e., 1 minute) and examine whether the NS-SCR rate changes in response to particular stimuli of interest. In the second type of electrodermal activity assessment (skin potential), no electrical current is introduced. Instead,

one (unipolar) electrode is placed on the skin to measure skin potential at a given point in time.

Eccrine sweat glands are found widely throughout the surface of the body, and are most densely distributed across the palms of hands and soles of feet. Thus, typical electrode placement for skin conductance is on two adjacent fingers or palm of one hand with a reference electrode placed on an abraded forearm. For skin potential assessment, one electrode is placed on the palm and a reference electrode is placed on an abraded forearm. Often researchers will place electrodes on the non-dominant hand and arm so as to limit interference with participant movement. Though this may make good practical sense, the reader is cautioned that electrodermal activity may differ from one side of the body of the other. Indeed, some psychophysiologicals have specifically focused on electrodermal activity asymmetry in their research (e.g., Picard, Fedor, & Ayzenberg, 2016).

Just as with cardiac activity recording devices, there are stationary and portable devices for electrodermal activity (e.g., Poh, Swenson, & Picard, 2010). Although there are many advantages to being able to record electrodermal activity wirelessly and in daily life settings, it is important to note that electrodermal activity recorded from other skin locations (i.e., wrists) may differ from what has been found for palms, and may be smaller in magnitude (Payne, Schell, & Dawson, 2016).

Example Research Applications

Social psychological research applications with ANS measurement are extensive—ranging from simple measures of arousal in early cognitive dissonance studies to more recent and complex assessment of parasympathetic and sympathetic balance in response to stressful stimuli. One of the most common social psychological research applications is in the area of emotion. Although a thorough review of ANS-emotion findings are beyond the scope of this chapter (see, Kriebig, 2010; Levenson, 2014), it is fair to say that the use of emotion elicitation paradigms in modern psychophysiological laboratories has led to a variety of insights on the links between emotion, emotion regulation, and ANS activity. For example, HR variability has emerged as a potential index of regulated emotional responding (Appelhans & Luecken, 2006). Just as emotion regulation reflects one's ability to adjust one's arousal and emotional state in response to changing emotional stimuli, HR variability reflects the degree to which the beating of one's heart can be altered to meet the demands of the changing environment (with higher HR variability indicating greater ability to transition between arousal levels). Accordingly, higher resting HR variability has been linked to a number of variables that indicate regulated emotional responding (e.g., constructive coping, faster fear extinction). Other ANS indices have been shown to correlate with other social psychological states and processes, such as empathy, social support, and stress appraisals. See Table 4.2 for more examples of social psychological research utilizing ANS indices.

TABLE 4.2 Example Social Psychological Research Applications Utilizing Autonomic Nervous System (ANS) Indices

<i>Topic</i>	<i>ANS Parameter(s)</i>	<i>Citation</i>	<i>Description</i>
Emotion	*many*	Kreibig, 2010	This review demonstrated some patterns in ANS responses to emotion subtypes and discrete emotions.
Emotion regulation	HRV	Applehans & Luecken, 2006	This review indicated that resting HRV is associated with less negativity bias, increased approach to novelty, constructive coping, and faster fear extinction.
Empathy	SC	Levenson & Ruef, 1992	Participants in their study who more accurately rated negative affect experienced by another had a more similar SC response to that person.
Social support	HR, BP, SC	Thorsteinsson & James, 1999	This meta-analysis showed that experimental manipulations of social support provision have moderate to large effects on HR and BP reactivity.
Stress appraisal	CO, TPR, PEP, HR	Seery, 2011	This review demonstrated that in response to self-relevant motivated performance situations, challenge states are characterized by relatively greater CO and lower TPR and threat states are characterized by relatively lower CO and higher TPR.

Endocrine System

Anatomy and Physiology

Similar to the nervous system, the endocrine system functions as a major communication system that consists of glands and hormones, the chemical messengers that glands secrete. Although the glands and organs of the endocrine system are widely distributed across the body and not anatomically continuous, they are tied together as a system based on the functions they serve. Furthermore, the nervous system is well integrated with the endocrine system, which underlies the close connections between psychological processes and hormonal changes. The study of hormones and hormonal processes in relation to psychological processes

and human behavior is referred to as psychoneuroendocrinology or behavioral endocrinology.

There are three classes of hormones, which control and regulate the activity of cells and organs (Neave, 2008). The majority are protein-based (peptide hormones), but others are derived from amino acids (amine hormones) or lipids (steroid hormones). See Table 4.3 for a list of hormones by class, their source, and their primary effects. The chemical structure of a hormone, or class, is important to note because it influences the transportation, movement, and half-life of hormones, and thus their assessment. For instance, steroid hormones, such as cortisol, are derived from cholesterol and can pass through cell walls and thus enter most

TABLE 4.3 Major Hormones and Their Primary Effects

<i>Hormone</i>	<i>Source</i>	<i>Primary Effects</i>
Peptide Hormones		
Adrenocorticotropin hormone	Anterior pituitary	Release of glucocorticoids and androgens
Corticotropin-releasing hormone	Hypothalamus	Release of adrenocorticotropin hormone
Gastrin	Stomach	Stomach acid secretion
Glucagon	Pancreas	Increase in blood sugar
Growth hormone	Anterior pituitary	Body growth and metabolic processes
Insulin	Pancreas	Reduction in blood sugar, promotion of cellular uptake of glucose, formation of glycogen
Leptin	Adipose tissue	Appetite suppression
Oxytocin	Hypothalamus (via posterior pituitary)	Contraction of uterus and mammary glands
Prolactin	Anterior pituitary	Lactation
Vasopressin	Hypothalamus (via posterior pituitary)	Vasoconstriction, water retention
Amine Hormones		
Epinephrine (or adrenaline)	Adrenal medulla	Preparation for emergency: increase in blood sugar, vasoconstriction, HR and BP
Melatonin	Pineal gland	Regulation of circadian timing
Norepinephrine (or noradrenaline)	Adrenal medulla	Preparation for emergency: increase in blood sugar, vasoconstriction, HR and BP

<i>Hormone</i>	<i>Source</i>	<i>Primary Effects</i>
Thyroxine	Thyroid	Stimulation and regulation of metabolism
Steroid hormones		
Cortisol	Adrenal cortex	Help body cope for moderate/long-term stress: increase in blood sugar; immune function alteration
Dehydroepiandrosterone	Adrenal cortex	Serves as a precursor to gonadal hormones and to oppose effects of glucocorticoids
Estrogens (estradiol, estrone, estriol)	Ovaries	Regulation of menstrual cycle and development of female reproductive system
Progesterone	Ovaries	Regulation of menstrual cycle and development of female reproductive system
Testosterone	Testes	Sperm production, growth and development of male reproductive system

tissues throughout the body to bind to receptors. Steroid hormones also easily cross the blood-brain barrier to exert effects both on the brain and periphery. In contrast, protein-based hormones, such as the neuropeptide oxytocin, bind to surface receptors and activate secondary messenger systems to exert their effects.

Two endocrine glands of major significance include the hypothalamus and pituitary. Located near the center of the brain, the hypothalamus is made up of more than two dozen nuclei and serves as a “control” center, linking the nervous and endocrine systems. It plays a vital role in maintaining homeostatic processes, such as those described in the earlier section on the autonomic nervous system. The hypothalamus largely exerts its regulating effects by controlling the release of hormones from the pituitary gland, which is suspended off of the hypothalamus. Often referred to as the “master gland,” the pituitary gland regulates many bodily functions, including growth, pain, blood pressure, and reproduction. Technically, the pituitary is two glands: the posterior pituitary and the anterior pituitary. The posterior pituitary stores and releases a variety of hormones that are produced by the hypothalamic nuclei (e.g., oxytocin, vasopressin) upon neural input from the central nervous system. The anterior pituitary has no neural inputs. Instead, other hormones trigger the anterior pituitary to synthesize and release additional hormones.

Hormones are typically released in a pulsatile fashion, meaning that glands secrete hormones multiple times throughout the day in brief surges. As such, the concentrations of hormones can fluctuate minute to minute. The secretion of a

given hormone is controlled by the concentrations of another, which is referred to as a hormone or biological cascade. Although there are several biological cascades within the endocrine system, the present focus is on one of the most prominent: the hypothalamic-pituitary-adrenal (HPA) axis and its end-product cortisol.

The HPA axis is activated by the hypothalamus in the brain, which receives and integrates somatosensory and affective input from the prefrontal cortex and limbic structures. In response, corticotrophin releasing hormone (CRH) is secreted by the neurons of the hypothalamus' paraventricular nucleus. CRH in turn triggers the release of adrenocorticotropin hormone (ACTH) from the anterior pituitary. ACTH stimulates the adrenal cortex to release glucocorticoids, including the catabolic steroid hormone cortisol. Biologically active cortisol acts independently on receptors distributed throughout the body and in conjunction with other physiological systems to regulate many functions critical for survival. For instance, cortisol acts with catecholamines of the autonomic nervous system to mobilize energy stores in response to stressors and also suppresses some aspects of the immune system and can inhibit reproductive processes. The HPA axis is regulated by negative feedback loops; as such, activation of the HPA axis is suppressed in response to elevated cortisol concentrations in healthy individuals. There is a pronounced circadian influence on cortisol concentrations; healthy individuals typically exhibit a robust increase in cortisol within the first 30–45 minutes after awakening, followed by decline across the rest of the day.

Assessment

Cortisol

Cortisol levels can be quantified from multiple biological specimens, most frequently from saliva and to a lesser extent blood. By collecting multiple blood or saliva samples, a researcher can document fluctuations in cortisol concentration over relatively short time periods (minutes to hours). Under “baseline” or resting conditions, approximately 5%–10% of circulating cortisol in the blood is free unbound, active), while the remaining is bound to proteins (corticosteroid binding globulin, serum albumin). Cortisol derived from blood represents both bound and unbound cortisol, with the latter reflecting the portion that is bioavailable, or free to act on target cells. Salivary concentrations, in contrast, reflect all bioavailable cortisol.

The sample of choice for most social scientists is saliva as it represents the freely available portion of cortisol, is non-invasive to collect, and easy to handle and store (relative to urine or blood), which allows for both laboratory- and field-based research. Saliva samples can be collected by passively drooling into a plastic tube via a straw or straw-like device or by using a commercially available device,

such as the Salivette (Sarstedt, Inc., Newton, NC), which includes a dental roll that participants put in mouths and saturate with saliva. It is good practice to store saliva samples in a freezer as soon as possible after collection until they are ready to be processed in-batch. However, it bears noting that cortisol levels are stable at room temperature for several weeks (Clements & Parker, 1998). Competitive enzyme linked immunosorbant assay determines how much of a substance (cortisol) is present in a sample.

Short-term changes in salivary cortisol concentration are commonly assessed in one of two general ways: (a) daily trajectories or slopes, and (b) responses to stressful events or stimuli, such as in the laboratory or in everyday life. For researchers interested in quantifying longer-term exposure to cortisol (months), hair can be a useful specimen to collect (Russell, Koren, Rieder, & Van Uum, 2012).

Multiple guidelines and recommendations have already been published for the collection of cortisol (Granger et al., 2007; Kudielka, Gierens, Hellhammer, Wüst, & Schlotz, 2012). However, a few basic recommendations are worth repeating here. First, sampling time points should be standardized across participants and the actual times of collection should be recorded. This is important because of the dramatic diurnal fluctuation in cortisol concentrations. Second, medical histories (e.g., endocrine and psychiatric disorders), medication use (e.g., oral contraception), and lifestyle factors (smoking, physical activity) can shape resting and reactive cortisol levels. As such, researchers typically exclude from participation or analyses cases in which participants use steroid-based medication, are pregnant, or have a major psychiatric or endocrine disorder. Depending on a particular study, additional inclusion or exclusion criteria might be applied.

Other hormones

Many of the same recommendation are in place for the assessment of other steroid hormones, such as testosterone or estradiol. For example, as with cortisol, circadian variation is commonly observed for other hormones. For ovarian sex hormones (e.g., progesterone, estradiol), phase of menstrual cycle can also dramatically alter hormone concentrations. To address this, some researchers schedule premenopausal women participants based on the day or phase of their menstrual cycle or control for it in analyses. It is also important to account for sex differences in research on steroid hormones, such as testosterone or estrogens. Passive drool is the recommend method for collecting saliva for quantifying most steroid hormones as cotton swabs can interfere with immunoassays (Granger et al., 2007). Although circulating peptide and amine hormones tonic and phasic levels can be readily quantified from blood samples, they are not reliably quantified in saliva. Additionally, large peptide hormones such as oxytocin do not cross the blood-brain barrier in significant amounts; therefore, peripheral blood concentrations may not reflect central levels found in the brain (for review, see Leng & Ludwig, 2016).

Example Research Applications

Burgeoning research of human psychoneuroendocrinology and social endocrinology focus on the bidirectional influences of social psychological processes and endocrine activity. Psychoneuroendocrinology researchers ask: can internal hormone changes influence psychosocial processes and can psychological and behavioral processes influence hormonal states? One of the most well-developed research topics in this area is psychological stress. Although conceptual and operational definitions of stress vary substantially between researchers, there is robust evidence that psychological stressors and perceptions of stress can have wide ranging effects on circulating hormones. Thus, many researchers incorporate endocrine assessment (e.g., salivary cortisol) to index psychological stress.

Additional psychoneuroendocrine research focuses on the links between testosterone and psychosocial processes such as aggression and dominance. Recent theoretical and empirical work incorporates both the catabolic hormone of cortisol and the anabolic hormone of testosterone to understand status, dominance, and testosterone. According to the dual-hormone hypothesis (Mehta & Josephs, 2010), testosterone is expected to influence aggressive and dominant behaviors when cortisol is low, but not when cortisol levels are high because cortisol counteracts the effects of testosterone. Consistent with this account, several studies have linked a hormone profile of high testosterone and low cortisol with increased dominance (e.g., Sherman, Lerner, Josephs, Renshon, & Gross, 2016). For more examples of other psychoneuroendocrine topics, including memory and prosocial behavior, see Table 4.4.

Immune System

Anatomy and Physiology

Immunity refers to the body's ability to resist or eliminate potentially harmful foreign materials or abnormal cells, and the cells and molecules of the immune system are responsible for coordinating the body's response to such pathogens or abnormalities. At first glance, it may seem surprising to learn that psychologists would be so interested in understanding the immunity and immune processes. However, as the last several decades of research has revealed, there are extensive and complex relationships between psychological phenomenon and immunological processes. In 1964, Solomon and Moos first coined the term psychoimmunology to describe the field that examines interactions between psychological states and immune function. A short while later, the term was expanded to psychoneuroimmunology (PNI) to reflect the role of the neuroendocrine system in also linking psychological states and immunological activity. The immune system is directly innervated by the ANS (lymph nodes, spleen, thymus) and immune cells have receptors for catecholamines and glucocorticoids.

TABLE 4.4 Example Social Psychological Research Applications Utilizing Hormonal Indices

<i>Topic</i>	<i>Hormone Parameter(s)</i>	<i>Citation</i>	<i>Description</i>
Aggression and dominance	Testosterone (and cortisol)	Sherman et al., 2016	Salivary testosterone predicted greater number of subordinates for male executives, but only for executives with low cortisol.
Memory	Epinephrine, cortisol	Cahill et al., 1994	Compared to placebo, propranolol (which blocks epinephrine effects), impaired memory for emotional but not neutral material.
Prosocial behavior	Oxytocin, vasopressin	Poulin et al., 2012	Oxytocin and vasopressin receptor genes interacted with perceived threat to predict prosocial behavior in a representative U.S. sample.
Resilience	Dehydroepiandrosterone	Petros et al., 2013	Self-reported resilience was positively correlated with salivary dehydroepiandrosterone.
Social-evaluative threat	Cortisol	Dickerson & Kemeny, 2004	This large meta-analysis suggested that social-evaluative threat elicits greater salivary cortisol responses than non-evaluative situations.

In addition, the immune system communicates directly with the central nervous system. Thus, the field of PNI has been shaped by the interest in understanding the bidirectional relationships between psychological phenomenon and immune processes and disease.

Humans have a range of complicated physiological mechanisms to protect our bodies against the invasion and potential damage from foreign microorganisms and own self cells that have gone awry. The overarching function of the human immune system is to protect the body from disease-causing microorganisms and other potentially harmful substances as well resist and eliminate infected and abnormal cells, such as tumors. The immune system recognizes the surface molecules, or antigens, of foreign and host cells. Carrying out these important functions are two major integrated subdivisions of the immune system: innate immunity and adaptive immunity (Daruna, 2012). Evolutionarily older, innate immunity is often described as the nonspecific first line of defense against injury

and infection. Innate immunity is a broad system of action that can respond relatively rapidly (i.e., minutes to hours) to microbes and toxins. Innate immune responses are relatively short in duration; thus, they do not lead to long-lasting immune protection. Innate immunity is comprised of elements with which a person is born, always present and available to protect an individual. Innate actions include the body's physical barriers (skin), secretions (mucous, stomach acid), and mechanical processes to block, trap, inactivate, or expel pathogens. The innate system can recognize foreign substances and recruit other immune cells and molecules to initiate, maintain, or dampen inflammation.

The primary signaling molecules of innate inflammatory responses are pro-inflammatory cytokines. Pro-inflammatory cytokines are produced by immune cells and other cells of the nervous system. Beyond influencing local inflammatory processes, cytokines are responsible for regulating complex and far-reaching changes throughout the body and brain (Maier & Watkins, 1998). The name for such a widespread response to infection is often referred to as sickness behavior or the acute-phase response, and encompasses a variety of physiological, cognitive, behavioral, and affective changes. Cardinal physiological alterations include fever, alterations in sleep structure, and production of acute-phase proteins and white blood cells. Behavioral responses include social withdrawal or isolation, increased pain complaints, and reductions in physical activity, eating, and libido. Dysphoria and anhedonia are also observed. Cognitive alterations include deficits in attention and memory interference. Sickness behaviors are thought to be adaptive changes that develop during the course of an infection to influence an organism's motivational state, to reorganize priorities to best cope with pathogens.

The second branch, adaptive immunity (sometimes referred to as specific or acquired immunity), comes online when innate immune processes are not sufficient to address the infection and contact is made by the invading antigen. Due to the large number of possible antigens, the adaptive immune system generally has relatively few immune cells (i.e., lymphocytes) that can respond to any particular antigen. However, once the immune cell binds to its target, additional cells proliferate to mount a sufficient response to the infection. Depending upon which type of molecule is detected, distinct immune response occur; for instance, cell-mediated responses primarily occur to neutralize viruses that have invaded the host's cells or to eliminate abnormal self cells (e.g., cancerous cells); in contrast, humoral or allergic responses largely target foreign extracellular substances, such as bacteria and allergens. The process of a full proliferative response can take much longer (weeks in some cases), but the response can lead to long-term protection by leaving behind memory cells that may allow for a more efficient response upon subsequent exposure. Such immunological memory helps to explain how vaccinations work. For greater detail on the complexity of immune processes and the many cells, signaling molecules, and receptors that are of relevance, see Daruna (2012).

Assessment

Just as the process of immunity is complex, so is the assessment of the system's varied cells and processes. There is no "one" test to measure global, or overall, immune function. Rather there are a number of tests from which to choose to estimate various components and functions of the immune system (Kiecolt-Glaser & Glaser, 1995). At the time of this writing, the mostly widely used approaches of measuring different aspects of immune function in psychological and behavior research involve the collection and assaying of circulating blood samples. Circulating blood samples can allow for both enumerative assays and functional assays. See Table 4.5 for a summary of functional and enumerative psychoneuroimmune measures. Enumerative assays refer to the counts, percentages, or concentrations of specific immunological marker (e.g., total or specific lymphocytes, particular antibodies, cytokines or acute-phase proteins) typically quantified from blood samples and in some cases saliva. For example, greater levels of circulating acute-proteins (e.g., C-reactive protein) are interpreted as greater levels of inflammation. In contrast to enumerative assays, functional assays allow researchers to quantify immunological responses to a variety of challenges *in vitro* or *in vivo*. For example, a researcher may be interested in measuring natural killer (NK) cell activity. NK cells are important for the detection and destruction of tumor cells and virally infected cells. To test NK cell lysis, or NK cells' ability to destroy, or lyse, cells, an *in vitro* functional immune test is used. Blood samples are drawn from the participant and then separated from his or her blood in the laboratory. NK cells are then cultured in media with a radioisotope. Afterward, the tagged NK cells are incubated with target cells that they can kill (typically from a cancer cell line), thus releasing the isotope, which is subsequently measured by the researchers. Greater counts reflect increased NK cell efficacy, which has implications for cancer progression. Although blood samples can be used to test many different enumerative and functional immune parameters, it is important to recognize that such samples reflect peripheral processes, rather than localized processes; that is, much of the action in an immune response is localized to tissues and immune organs (e.g., lymph nodes). As such, blood draws may miss important information.

For all aforementioned immune tests (and those described in Table 4.5), specially trained staff and specific equipment is necessary to complete the assessments. Since many psychologists do not themselves have access to a nurse or phlebotomist and wet laboratory environment, it is important to partner with those who do have these skills and resources. As an alternative to blood-based immune tests and the constraints that come along with them, salivary assays of immune parameters have grown more common among behavioral scientists. For example, a recent review of the extant literature indicates that some cytokines (e.g., interleukin-1b, tumor necrosis factor alpha, and interleukin-6), increase

TABLE 4.5 Examples of Functional and Enumerative Psychoneuroimmune (PNI) Measures

<i>PNI parameter</i>	<i>Description</i>
Enumerative Measures	
Inflammatory cytokines	Test to measure amount of inflammatory signaling molecules (e.g., interleukin-6, tumor necrosis factor alpha)
Acute-phase molecules	Test to measure amount of acute-phase molecules produced in response to inflammation (e.g., C-reactive protein)
Specific antibodies	Test to measure specific antibodies (e.g., secretory immunoglobulin A)
Specific immune cell count	Test to measure number of specific immune cells (e.g., CD4+ cells)
Total white blood cell count	Test to measure number of circulating white blood cells, or leukocytes
Functional Measures	
Proliferative responses to challenge	Test to determine the degree to which lymphocytes proliferate or replicate in response to mitogens, or infectious agents, such as Lipopolysaccharide (and measure resulting increase in inflammatory cytokines, such as interleukin-6) or Concanavalin A or Phytohemagglutinin (and measure T- and B-lymphocyte responses, respectively)
Natural killer cell activity	Test to determine the cytotoxicity, or tumor-killing ability of NK cells, by incubating participants' NK cells with target cells. Degree of NK activity is quantified by amount of isotope released by lysed (killed) cancer cells.
Latent virus titers	Test to determine the immune system's ability to control a latent viral infection, such as herpesvirus or Epstein-Barr virus.
Response to vaccine	Vaccine (e.g., trivalent influenza vaccine) is administered to participants. Pre- and post-vaccine measures (via blood or self-report) are taken.
Response to inoculation	Virus (e.g., common cold) or bacteria is administered to participants. Pre- and post-inoculation measures are taken.
Wound healing	Wound (e.g., mouth puncture, skin abrasion) is administered to participants. Pre- and post-wound measures are taken.

fairly consistently in saliva in response to acute stress (Slavish, Graham-Engeland, Smyth, & Engeland, 2015).

Regardless of whether the researcher is collecting blood or saliva samples for quantifying aspects of the immune system, there are a number of methodological issues that are of relevant across all PNI research. Segerstrom and Smith (2012) describe three kinds of important error in PNI research: the good, the bad, and the ugly. The good error reflects variability in immune parameters that are of

interest to researcher. For instance, immune cells or inflammatory responses may vary as a function of psychosocial stress, pain, or other social psychological phenomena. Bad error is that which results from type I error, such as conducting studies with small sample sizes and many dependent variables. With the relatively high costs of conducting PNI research, sample sizes are often underpowered to perform complex statistical tests. Ugly error results from noise in the research process. For example, not controlling for time of day, using multiple assays, laboratories, or technicians to process samples, and improperly storing samples can all lead to error in immune parameters.

Example Research Applications

As with measures of the ANS and endocrine system, psychological stress is a major area of emphasis in understanding psycho-immune links. For example, in a series of studies by Sheldon Cohen and colleagues, psychological stress predicts greater susceptibility to developing the common cold and degree of symptom severity (e.g., Cohen, Tyrrell, & Smith, 1991). In addition, the experience of acute psychological stressors is reliably linked to increased plasma and salivary markers of inflammation (Segerstrom & Miller, 2004; Slavish et al., 2015). Other lines of research indicate that pro-inflammatory cytokines regulate psychosocial behaviors, such as social withdraw, and other sickness behaviors (anhedonia, dysphoria, fatigue, and cognitive and motor impairment). For example, research by Naomi Eisenberger and colleagues reveals how inflammation contributes to social, cognitive, and affective symptoms of depression (e.g., Eisenberger, Inagaki, Mashal, & Irwin, 2010). See Table 4.6 for additional examples of social psychological research with immune measures, including studies of emotional disclosure, perseverative cognition, and personality.

General Methodologic and Analytic Considerations

If it has not yet been made abundantly clear from the preceding sections, psychophysiological assessment is not for the faint of heart! When one begins to undertake the acquisition of biological data, many questions need to be answered, starting with: Which system should be measured? Which biomarkers within a given system? Once the system or biomarker has been identified, next questions are: What kind of equipment, staff, and training are needed for the acquisition of these data? What timing and number of assessments should be recorded? Alongside these questions are other practical considerations, including cost, equipment, space, degree of invasiveness and burden placed on the participant, availability of appropriately trained collaborators and research assistants, and procedures for storing, analyzing, and interpreting the data. The following sections discuss several broad methodologic and analytic issues that arise when collecting psychobiological data.

TABLE 4.6 Example Social Psychological Research Applications Utilizing Psychoneuro-immune (PNI) Measures

<i>Topic</i>	<i>PNI Parameter(s)</i>	<i>Citation</i>	<i>Description</i>
Emotional disclosure	Wound healing	Weinman et al., 2008	Compared to a control condition, an emotional disclosure writing intervention led to smaller punch biopsy wounds.
Perseverative cognition	Acute-phase molecules, inflammatory cytokines	Zoccola et al., 2014	Compared to distraction, post-stressor rumination led to greater C-reactive protein (but no difference in interleukin-6 or tumor necrosis factor alpha).
Personality	Response to cold virus	Cohen et al., 2003	Positive emotional style was associated with lower risk of developing a cold after rhinovirus exposure.
Social disconnection	Inflammatory cytokines	Eisenberger et al., 2010	Relative to placebo, endotoxin led to increases in interleukin-6 and tumor necrosis factor alpha as well as greater feelings of social disconnection.
Stress	*many*	Segerstrom & Miller, 2004	The large meta-analysis indicates that stress reliably alters immune parameters, and the nature of the alterations depend on the nature of stress (e.g., acute versus chronic).

Tonic versus Phasic Measures

Although some hypotheses are concerned with a tonic physiological parameter (e.g., resting HR variability), others address phasic responses (e.g., short term changes in cortisol concentration). Tonic levels refer to relatively stable values of a particular parameter. Tonic measures are sometimes referred to as resting or baseline levels (e.g., resting HR). Phasic measures refer to momentary fluctuations, typically occurring in response to some type of stimulus that has been introduced by the researcher. Phasic measures are sometimes also referred to as reactivity or responsivity (e.g., skin conductance response, HR reactivity). When testing the latter, it is vital to select appropriate comparison conditions (e.g., resting baseline) from which to derive a response. Sometimes a minimally demanding, or “vanilla,” baseline period in which the participant is alert but inactive is a more appropriate comparison than absolute rest (Jennings, Kamarck, Stewart, Eddy, & Johnson, 1992). A true “resting” basal period can be quite difficult to achieve, and participants who are asked to do “nothing,” may be anxiously awaiting the next

laboratory procedure, dozing off, or trying to find some other distracting activity to occupy them selves in the laboratory.

Timing and Number of Assessments

Biological changes in the autonomic, endocrine, and immune system vary in speed and duration. ANS changes can occur in fractions of a second, whereas it may take many minutes or hours to see full endocrine or immunological responses to particular stimuli. Additionally, psychological phenomena vary in their temporal characteristics and dynamic complexity (e.g., fleeting or fluctuating emotions, chronic ongoing stress). Researchers should carefully consider the temporal characteristics of the phenomena under study to ensure that the timing of assessment of psychological and biological measures are appropriately aligned. It bears noting, however, that with new measures and questions, there can often still be a bit of guesswork in selecting psychobiological sampling periods. Recent and future empirical and meta-analytic work will be useful in helping to establish such windows of time. For example, how soon after the introduction of a psychosocial stressor might a researcher expect to observe peak changes in salivary cortisol? Results from a meta-analysis of 204 laboratory stressor tells us that we should aim for 21–30 minutes post-stressor onset (Dickerson & Kemeny, 2004).

Given that some psychobiological data are collected continuously over extended periods of time, such as over the course of an hour-long laboratory visit with multiple tasks or over several days in an ambulatory study, researchers need to make decisions about how to quantify continuously collected data. What periods of time are of interest? The last minute of a resting baseline period? Each minute of a 5-minute cognitive task? The average of an entire task period? Answers to these questions should be dictated by the research questions of interest (and limits of psychophysiological recording equipment!). If the researcher plans to calculate difference scores to quantify change, or compare ANS activity during one procedure versus another, he or she should select sampling periods, or epochs, of equal duration so that variance estimates are comparable.

Data Storage, Processing, Analysis, and Interpretation

Data Synchronization and Storage

The synchronization of data across measurements for each individual is critical for analysis and interpretation. Careful timing and marking of continuous and intermittent data files is essential to compiling full datasets that link task or stimulus outcomes to physiological data. In addition to consulting with hardware and software companies that supply the products used in the laboratory, it may be necessary to consult with additional experts with backgrounds in biotechnology or software programming to facilitate this process. Furthermore, it is good practice

to pilot test data collection procedures and then practice data extraction, compilation, and analysis to ensure all necessary variables are properly collected and stored. Pilot testing is also necessary to ensure sufficient time for participants to habituate to the laboratory, complete tasks, and capture peak reactivity or recovery data.

Data Ranges and Norms

Many physiological parameters fluctuate over time as result of multiple factors, including variables of theoretical interest (e.g., emotional state) or as a result of methodological issues (e.g., movement, temperature, time of day). Once psychobiological data are collected, it should be inspected for possible artifacts and biologically implausible values (e.g., increased HR due to finger tapping). Although analysis software typically has algorithm-based functions to identify and address artifacts in continuous waveform data, it is useful to visually inspect data to some extent. Thus, it is important to train conscientious research assistants to assist in this biosignal data processing and editing role. In some cases, meaningful clinical cut-offs can be used when processing psychobiological data (e.g., systolic blood pressure over 140 may signify hypertension). Often, psychobiological data are non-normally distributed; positive skewness and outliers are common. For some, outliers may be of interest and relevant to the research question, so excluding extreme values may not always be the right approach. Nonetheless, data transformations are common (e.g., log transformations of salivary cortisol), but can make it more difficult to interpret findings. It is generally a good idea to become familiar with reporting guidelines and norms for your psychobiological measure of interest.

Statistical Analysis

Psychobiological research typically involves repeated measures or continuous assessments that are subsequently broken down into time-series data for each participant. Such data violate key assumptions of independence of many statistical methods. Despite this, statistical techniques based on analysis of variance (e.g., repeated measures ANOVA) with appropriate corrections for assumption violations are still the most commonly used in psychobiological research at present. Increasingly more common are empirical papers that contain regression-based techniques and multi-leveling modeling analytic procedures. For example, Houtveen, Hamaker, and Van Doornen (2010) detail a multilevel path analysis approach for analyzing 24-h ambulatory physiological recordings. Researchers interested in psychobiological research should consider taking advanced statistical courses at their universities or attend such workshops elsewhere. In addition, there are a variety of helpful resources analyzing psychobiological data with multilevel models (e.g., Blackwell, Mendes de Leon, & Miller, 2006; Hruschka, Kohrt, & Worthman, 2005).

Interpretation of Results

What does it mean if a biological parameter is moderately correlated with psychological variable? And what if they don't correlate? To some extent, the answer to this question will depend on the particular theory driving the research. However, alternative biological and methodological explanations are also worth considering and addressing in future research. For example, a review of 49 studies employing a standardized psychosocial laboratory stressor found small to moderate associations between biological and emotional stress responses in approximately 25% of the studies (Campbell & Ehler, 2012). Because synchrony of physiological and psychological measures in response to stress was expected based on theory, the authors concluded that a variety of methodological, psychological, and biological factors may have contributed to the weak convergence in measures (e.g., socially desirable responding, poorly timed assessments, novelty of task). One recommendation stemming from this review is that rather than relying on single post-task or pre-post assessments, researchers might be wise to implement and aggregate across multiple assessments (e.g., across multiple tasks; in response to repeated same task exposure) to reveal trait-like physiological response dispositions.

Concluding Remarks and Future Directions

To date, the majority of research incorporating psychobiological measurement has taken place in research laboratories. Although this approach has led to a variety of important insights, idiographic longitudinal studies are best suited to track how alterations in one system or set of processes may covary with changes in another. For example, daily diary designs with ambulatory biological assessment are ideal to study the dynamic unfolding of stress and coping processes (e.g., stress appraisals, emotions, and concomitant biological changes). Moreover, ambulatory physiological assessments in naturally occurring contexts can allow researchers to more fully understand the experience of people by capturing life as it is lived. Future research that incorporates both laboratory-based and ambulatory assessments could lead to transformational research findings and help address limitations that exist in laboratory-only or field-only designs. Fortunately, a variety of technological advances in high-capacity batteries, compact design and portability, miniaturized sensors, powerful computing, and easy-to-use software will allow for long-term continuous measurement in daily life with minimal disruption. Substantial efforts also are being made to produce biological sensors that are smaller and allow for more continuous wearable sensors that can reliably quantify biomarkers in perspiration (Jia, Chew, Feinstein, Skeath, & Sternberg, 2016). Still other new technologies allow for sensors to be swallowed, printed on skin, or imbedded in already worn accessories (e.g., eye-glasses, jewelry). For a review of a variety of wearable sensors, see the Department of Defense's recent report (Hirschberg, Betts, Emanuel, & Caples, 2014).

As technology continues to advance, so too will the opportunities for innovative psychobiological assessment by social scientists. The marriage of sound psychological theory with deep understanding of physiological systems are necessary for the most insightful and impactful research. Successful scholars of psychobiological assessment will be those who are both well-versed in social psychological theory and have a good understanding for human physiology (both its function and assessment). That said, it is readily acknowledged that any one person will not have full knowledge of all biological systems and processes that may be relevant to the research question at hand. Thus, collaborative science is truly necessary in this realm of research. Is psychobiological assessment worth the trouble? The answer is a resounding yes. Although psychobiological assessment may not be suited for all social psychological research questions or researchers, it has the ability to complement and enhance existing social psychological methods, and lead to new insights to understanding human mind and behavior.

Recommended Further Reading

- Andreassi, J. L. (2007). *Psychophysiology: Human behavior and physiological response* (5th ed.). New York, NY: Taylor & Francis.
- Cacioppo, J. T., Tassinary, L. G., & Bernston, G. G. (Eds.). (2017). *Handbook of psychophysiology* (4th ed.). New York, NY: Cambridge University Press.
- Neave, N. (2008). *Hormones and behaviour: A psychological approach*. New York, NY: Cambridge University Press.
- Segerstrom, S. C. (Ed.). (2012). *The Oxford handbook of psychoneuroimmunology*. New York, NY: Oxford University Press.

References

- Andreassi, J. L. (2007). *Psychophysiology: Human behavior and physiological response* (5th ed.). New York, NY: Taylor & Francis.
- Appelhans, B. M., & Lueken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, 10, 229–240. Retrieved from <https://doi.org/10.1037/1089-2680.10.3.229>
- Blackwell, E., de Leon, C. F. M., & Miller, G. E. (2006). Applying mixed regression models to the analysis of repeated-measures data in psychosomatic medicine. *Psychosomatic Medicine*, 68, 870–878. Retrieved from <https://doi.org/10.1097/01.psy.0000239144.91689.ca>
- Blascovitch, J., Vanman, E. J., Mendes, W. B., & Dickerson, S. (2011). *Social psychophysiology for social and personality psychology*. Los Angeles, CA: Sage Publications.
- Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). β -Adrenergic activation and memory for emotional events. *Nature*, 371, 702–704. Retrieved from <http://dx.doi.org/10.1038/371702a0>
- Campbell, J., & Ehler, U. (2012). Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology*, 37, 1111–1134. Retrieved from <https://doi.org/10.1016/j.psyneuen.2011.12.010>
- Cannon, W. B. (1915). *Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement*. New York, NY: D. Appleton and Company.

- Clements, A. D., & Parker, C. R. (1998). The relationship between salivary cortisol concentrations in frozen versus mailed samples. *Psychoneuroendocrinology*, 23, 613–616. Retrieved from [https://doi.org/10.1016/S0306-4530\(98\)00031-6](https://doi.org/10.1016/S0306-4530(98)00031-6)
- Cohen, S., Doyle, W. J., Turner, R. B., Alper, C. M., & Skoner, D. P. (2003). Emotional style and susceptibility to the common cold. *Psychosomatic Medicine*, 65, 652–657. Retrieved from <http://dx.doi.org/10.1097/01.PSY.0000077508.57784.DA>
- Cohen, S., Tyrrell, D. A., & Smith, A. P. (1991). Psychological stress and susceptibility to the common cold. *New England Journal of Medicine*, 325, 606–612. Retrieved from <https://doi.org/10.1056/NEJM199108293250903>
- Daruna, J. H. (2012). *Introduction to psychoneuroimmunology* (2nd ed.). New York, NY: Academic Press.
- Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130, 355–391. Retrieved from <https://doi.org/10.1037/0033-2909.130.3.35>
- Eisenberger, N. I., Inagaki, T. K., Mashal, N. M., & Irwin, M. R. (2010). Inflammation and social experience: An inflammatory challenge induces feelings of social disconnection in addition to depressed mood. *Brain, Behavior, and Immunity*, 24, 558–563. Retrieved from <https://doi.org/10.1016/j.bbi.2009.12.009>
- Granger, D. A., Kivlighan, K. T., Fortunato, C., Harmon, A. G., Hibell, L. C., Schwartz, E. B., & Whemolua, G. L. (2007). Integration of salivary biomarkers into developmental and behaviorally-oriented research: Problems and solutions for collecting specimens. *Physiology & Behavior*, 92, 583–590. Retrieved from <http://dx.doi.org/10.1016/j.physbeh.2007.05.004>
- Hirschberg, D. L., Betts, K., Emanuel, P., & Caples, M. (2014). *Assessment of wearable sensor technologies for Biosurveillance (ECBC-TR-1275)*. Aberdeen, MD: Edgewood Chemical Biological Center.
- Houtveen, J. H., Hamaker, E. L., & Van Doornen, L. J. P. (2010). Using multilevel path analysis in analyzing 24-h ambulatory physiological recordings applied to medically unexplained symptoms. *Psychophysiology*, 47, 570–578. Retrieved from <https://doi.org/10.1111/j.1469-8986.2009.00951.x>
- Hruschka, D. J., Kohrt, B. A., & Worthman, C. M. (2005). Estimating between-and within-individual variation in cortisol levels using multilevel models. *Psychoneuroendocrinology*, 30, 698–714. Retrieved from <http://dx.doi.org/10.1016/j.psyneuen.2005.03.002>
- James, W. (1884). What is an emotion? *Mind*, 9, 188–205.
- Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: Vanilla or resting baseline. *Psychophysiology*, 29, 742–750. Retrieved from <https://doi.org/10.1111/j.1469-8986.1992.tb02052.x>
- Jia, M., Chew, W. M., Feinstein, Y., Skeath, P., & Sternberg, E. M. (2016). Quantification of cortisol in human eccrine sweat by liquid chromatography—tandem mass spectrometry. *Analyst*, 141, 2053–2060. Retrieved from <https://dx.doi.org/10.1039/C5AN02387D>
- Kiecolt-Glaser, J. K., & Glaser, R. (1995). Measurement of immune response. In S. Cohen, R. Kessler, & L. Gorden (Eds.), *Measuring stress: A guide for health and social sciences* (pp. 213–229). New York, NY: Oxford University Press.
- Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84, 394–421. Retrieved from <https://doi.org/10.1016/j.biopsycho.2010.03.010>
- Kudielka, B. M., Gierens, A., Hellhammer, D. H., Wüst, S., & Schlotz, W. (2012). Salivary cortisol in ambulatory assessment—some dos, some don'ts, and some open questions. *Psychosomatic Medicine*, 74, 418–431. Retrieved from <https://doi.org/10.1097/PSY.0b013e31825434c7>

- Leng, G., & Ludwig, M. (2016). Intranasal oxytocin: Myths and delusions. *Biological Psychiatry*, 79, 243–250. Retrieved from <https://doi.org/10.1016/j.biopsych.2015.05.003>
- Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion Review*, 6, 100–112. Retrieved from <https://doi.org/10.1177/1754073913512003>
- Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology*, 63, 234. Retrieved from <http://dx.doi.org/10.1037/0022-3514.63.2.234>
- Maier, S. F., & Watkins, L. R. (1998). Cytokines for psychologists: Implications of bidirectional immune-to-brain communication for understanding behavior, mood, and cognition. *Psychological Review*, 105, 83–107. Retrieved from <https://dx.doi.org/10.1037/0033-295X.105.1.83>
- Mehta, P. H., & Josephs, R. A. (2010). Testosterone and cortisol jointly regulate dominance: Evidence for a dual-hormone hypothesis. *Hormones and Behavior*, 58, 898–906. Retrieved from <https://doi.org/10.1016/j.yhbeh.2010.08.020>
- Neave, N. (2008). *Hormones and behaviour: A psychological approach*. New York, NY: Cambridge University Press.
- Neumann, E., & Blanton, R. (1970). The early history of electrodermal research. *Psychophysiology*, 6, 453–475. Retrieved from <https://doi.org/10.1111/j.1469-8986.1970.tb01755.x>
- Payne, A. F. H., Schell, A. M., & Dawson, M. E. (2016). Lapses in skin conductance responding across anatomical sites: Comparison of fingers, feet, forehead, and wrist. *Psychophysiology*, 53, 1084–1092. Retrieved from <https://doi.org/10.1111/psyp.12643>
- Petros, N., Opacka-Juffry, J., & Huber, J. H. (2013). Psychometric and neurobiological assessment of resilience in a non-clinical sample of adults. *Psychoneuroendocrinology*, 38, 2099–2108. Retrieved from <https://doi.org/10.1016/j.psyneuen.2013.03.022>
- Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review*, 8, 62–75. Retrieved from <https://doi.org/10.1177/1754073914565517>
- Poh, M. Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 57, 1243–1252. Retrieved from <http://dx.doi.org/10.1109/TBME.2009.2038487>
- Poulin, M. J., Holman, E. A., & Buffone, A. (2012). The neurogenetics of nice: Receptor genes for oxytocin and vasopressin interact with threat to predict prosocial behavior. *Psychological Science*, 23, 446–452.
- Russell, E., Koren, G., Rieder, M., & Van Uum, S. (2012). Hair cortisol as a biological marker of chronic stress: Current status, future directions and unanswered questions. *Psychoneuroendocrinology*, 37, 589–601. Retrieved from <https://doi.org/10.1016/j.psyneuen.2011.09.009>
- Saul, J. P. (1990). Beat-to-beat variations of heart rate reflect modulation of cardiac autonomic outflow. *Physiology*, 5, 32–37.
- Seery, M. D. (2011). Challenge or threat? Cardiovascular indexes of resilience and vulnerability to potential stress in humans. *Neuroscience & Biobehavioral Reviews*, 35, 1603–1610. Retrieved from <http://dx.doi.org/10.1016/j.neubiorev.2011.03.003>
- Segerstrom, S. C., & Miller, G. E. (2004). Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry. *Psychological Bulletin*, 130, 601–630. Retrieved from <https://dx.doi.org/10.1037/0033-2909.130.4.601>
- Segerstrom, S. C., & Smith, G. T. (2012). Methods, variance, and error in psychoneuroimmunology research: The good, the bad, and the ugly. In S. Segerstrom (Ed.), *The Oxford*

- handbook of psychoneuroimmunology* (pp. 421–432). New York, NY: Oxford University Press.
- Sherman, G. D., Lerner, J. S., Josephs, R. A., Renshon, J., & Gross, J. J. (2016). The interaction of testosterone and cortisol is associated with attained status in male executives. *Journal of Personality and Social Psychology*, *110*, 921–929. Retrieved from <http://dx.doi.org/10.1037/pspp0000063>
- Slavish, D. C., Graham-Engeland, J. E., Smyth, J. M., & Engeland, C. G. (2015). Salivary markers of inflammation in response to acute stress. *Brain, Behavior, and Immunity*, *44*, 253–269. Retrieved from <https://doi.org/10.1016/j.bbi.2014.08.008>
- Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. (1996). Heart rate variability: Standards of measurement, physiology interpretation, and clinical use. *Circulation*, *93*, 1043–1065. Retrieved from <https://doi.org/10.1161/01.CIR.93.5.1043>
- Thorsteinsson, E. B., & James, J. E. (1999). A meta-analysis of the effects of experimental manipulations of social support during laboratory stress. *Psychology and Health*, *14*, 869–886. Retrieved from <http://dx.doi.org/10.1080/08870449908407353>
- Walsh, J. A., Topol, E. J., & Steinhubl, S. R. (2014). Novel wireless devices for cardiac monitoring. *Circulation*, *130*, 573–581. Retrieved from <https://doi.org/10.1161/CIRCULATIONAHA.114.009024>
- Weinman, J., Ebrecht, M., Scott, S., Walburn, J., & Dyson, M. (2008). Enhanced wound healing after emotional disclosure intervention. *British Journal of Health Psychology*, *13*, 95–102. Retrieved from <https://doi.org/10.1348/135910707X251207>
- Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A., & Stoll, R. (2010). Comparison of three mobile devices for measuring R—R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. *European Journal of Applied Physiology*, *109*, 779–786. Retrieved from <https://doi.org/10.1007/s00421-010-1415-9>
- Williams, D. P., Jarczok, M. N., Ellis, R. J., Hillecke, T. K., Thayer, J. F., & Koenig, J. (2016). Two-week test—retest reliability of the Polar® RS800CX™ to record heart rate variability. *Clinical Physiology and Functional Imaging*, *n.v.*, *n.p.* Retrieved from <https://doi.org/10.1111/cpf.12321>
- Zoccola, P. M., Figueroa, W. S., Rabideau, E. M., Woody, A., & Benencia, F. (2014). Differential effects of poststressor rumination and distraction on cortisol and C-reactive protein. *Health Psychology*, *33*, 1606–1609. Retrieved from <http://dx.doi.org/10.1037/hea0000019>

5

IT'S ABOUT TIME

Event-Related Brain Potentials and the Temporal Parameters of Mental Events

*Meredith P. Levsen, Hannah I. Volpert-Esmond,
and Bruce D. Bartholow*

Time as a dimension of every mental or behavioral process lends itself to measurement . . . [but] a technical difficulty at once suggests itself. "The speed of thought," we say; but as soon as we set about measuring the time occupied by a thought we find that the beginning and end of any measurable time must be external events. We may be able in the future to use "brain waves" as indicators of the beginning and end of a mental process.

(Woodworth, 1938, p. 298)

The timing of mental events is among the most important and enduring constructs in psychology (see Jensen, 2006). Particularly in social-personality (see Chaiken & Trope, 1999) and cognitive psychology (e.g., Jacoby, 1991), numerous theories posit "dual processes" for understanding thought and action in which a central organizing principle is the idea that some mental processes unfold rapidly and spontaneously, whereas others rely on a slower, more deliberative form of processing. This dichotomy is nicely underscored by the title of Daniel Kahneman's (2011) book, *Thinking, Fast and Slow*. This dichotomy also can be understood in terms of the influence of rapidly occurring processes on slower developing events. For example, impressions of people formed in milliseconds can contribute to thoughts, decisions, and behaviors that affect interpersonal interactions over minutes, days, or years (e.g., Ambady & Rosenthal, 1992).

For centuries, scientists and philosophers believed that thought happened instantaneously, too quickly to be measured (see Glynn, 2010). But in 1850, Hermann von Helmholtz hit upon a method for measuring the speed of thought. Using a device called a galvanometer, von Helmholtz measured the time required for an electrical impulse applied to a sciatic nerve to cause movement in a calf

muscle, inferring that this method emulates the electrical impulses that naturally travel along nerve fibers. Using this procedure, von Helmholtz discovered that neural transmission speed was not instantaneously fast but in fact was relatively sluggish—around 30 meters per second in humans. This discovery, coupled with subsequent extensions to the central nervous system (see Hodgkin, 1964), led to the revolutionary idea that the speed of thought could be quantified, setting the stage for virtually all of experimental psychology.

Despite the enduring attractiveness of this idea, the measures typically used in cognitive and social-personality psychology provide limited information regarding the timing and function of mental events. Most behavioral responses (e.g., accuracy or response time [RT]) represent a single, discrete outcome of the operations of numerous processes with overlapping (and often unknown) temporal parameters operating at different levels (e.g., perceptual, cognitive, motor), some of which may not be of interest to the researcher (see Bartholow, 2010). RTs measured in different experimental conditions generally are assumed to vary because of the content, duration, or temporal sequencing of mental events across conditions (see Donders, 1969; Posner, 1978). To the extent that separating these influences is of theoretical interest, using RTs alone is likely to be insufficient.

In contrast, event-related potentials (ERPs) are uniquely suited to characterizing the temporal properties of specific mental processes. The electroencephalogram (EEG), from which ERPs are derived, can be measured with a temporal resolution of less than a millisecond (up to 2,500 samples per second [Hz]), faster than the native temporal resolution of neural activity (Reed, Vernon, & Johnson, 2004). This allows researchers to assess processes reflecting mental operations that unfold over tens or hundreds of milliseconds (see Amodio, Bartholow, & Ito, 2014). When combined with methodological and theoretical rigor, ERPs allow researchers a way of more directly accessing otherwise unobservable neurocognitive processes that support psychological constructs. The utility of ERPs for characterizing the temporal architecture of the information-processing system (i.e., *mental chronometry*; Posner, 1978) was convincingly demonstrated by Coles and colleagues (1985), who showed that RT in a cognitive control task varied according to three distinct and largely independent processes (response priming, stimulus evaluation, and response competition) that partially overlap in time during stimulus processing. These data challenged the long-held assumption that processing proceeds in serial stages (see Sternberg, 1969), and supported the alternative idea that processes conjointly accumulate information contributing to behavioral responses. Perhaps more importantly, such findings represent a realization of Woodworth's (1938) long-anticipated vision of a better means for timing mental events.

When applied to understanding social-personality processes, ERPs are especially helpful as covert measures of processes that occur too rapidly for assessment via self-report or other behavioral methods (i.e., implicit processes). Additionally, ERPs have considerable promise as markers of individual differences whose variability can signify temperament or other person-level processes (e.g., Bress,

Meyer, & Proudfit, 2015). Despite these advantages, the potential of the temporal specificity of ERPs to advance social-personality theory remains largely untapped. In this chapter, we describe the utility and reliability of several widely studied ERPs within social-personality psychology. Our approach acknowledges that social-personality psychologists are interested in both mental processes and behavior, and thus we emphasize that this technique may be used to complement behavioral measures, not replace them. Note that space limitations preclude a comprehensive review of ERPs and their application to social-personality psychology; additional information can be found in other sources (e.g., Amodio & Bartholow, 2011; Amodio et al., 2014; Von Gunten, Bartholow, & Volpert, 2016).

What Are ERPs?

In simplest terms, ERPs are electrical signals produced by the firing of (mainly cortical) neurons. An ERP waveform (see Figure 5.1) represents a defined segment of ongoing brain electrical activity (i.e., electroencephalogram; EEG) that

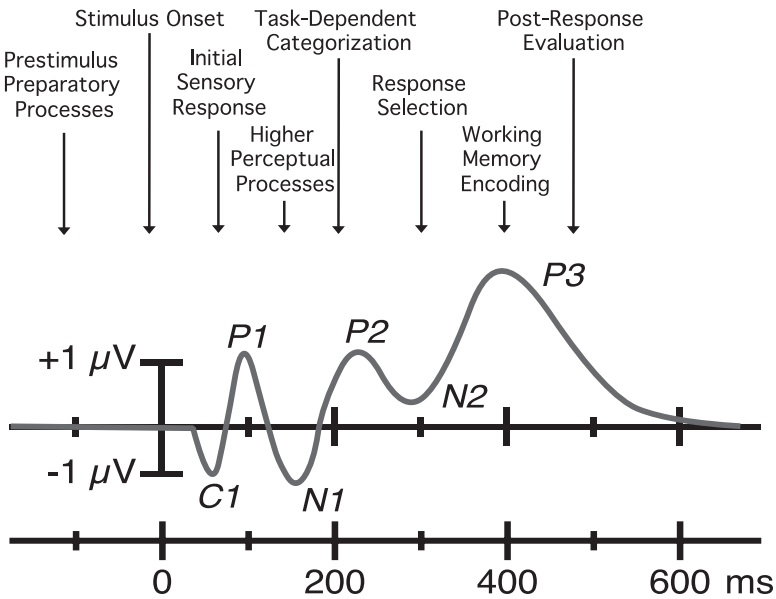


FIGURE 5.1 A Schematic Representation of an ERP Waveform Elicited by a Discrete Visual Stimulus

Note: The x-axis represents time and the y-axis represents amplitude in microvolts. The positive and negative deflections represent typical ERP components named for their polarity (“P” for positive, “N” for negative) and ordinal position in the waveform. Here, positive amplitudes are plotted upward, although ERP waveforms are often plotted with negative values upward according to electrophysiological convention. Image used with permission, © 2016 S. J. Luck.

is time-locked to a discrete event, commonly a stimulus presentation or a participant's behavioral response. Tiny electrical signals produced by post-synaptic potentials in activated neurons propagate through the brain to the surface of the scalp, where they are detectable by electrodes. The sequences of positive and negative voltage fluctuations observed in the EEG signal reflect opposing ends of an electrical dipole (akin to a common battery) created by the summation of these potentials generated in millions of activated neurons that are synchronously active and spatially aligned perpendicular to the scalp (see Allison, Wood, & McCarthy, 1986; Luck, 2014).

Researchers generally are interested in the magnitude and/or timing of specific fluctuations, often referred to as *components*, occurring at particular intervals in the ERP that theory and prior research have associated with psychological processes of interest. As with any measure of an unobservable entity, linking physiological events to psychological processes requires inferences. In ERP research generic inferences assume: (a) that ERP components represent the activity of one or more information-processing operations; (b) that variations in the size (i.e., amplitude) of these components reflect the degree of engagement of those operations; and (c) that variations in the timing (i.e., latency) of the components reflect differences in the temporal parameters of those operations, such as their initiation and duration (see Donchin, Karis, Bashore, Coles, & Gratton, 1986).¹

How Are ERPs Measured?

Recording ERPs requires that stimuli be discrete events presented within a task in which the timing of stimulus onset and offset (and, when appropriate, behavioral responses) can be precisely controlled. Participants in ERP experiments typically sit upright before a video display, often with fingers placed on keys of a response device. While they complete the task EEG is recorded from an array of electrodes placed on the scalp, arranged according to standard placement guidelines (see American Encephalographic Society, 1994).

The most common approach to ERP measurement relies on a signal averaging approach in which stimulus- or response-locked epochs of EEG activity from numerous trials of the same type are averaged. Through this averaging process, EEG activity elicited by events of interest (i.e., signal) increases, whereas activity unrelated to the event (i.e., noise) will vary randomly across epochs and tend to average to zero. These averaged epochs are aligned with reference to a pre-event baseline period—usually 100–200 ms—so that EEG amplitude at the time of event onset will be zero. Trials containing large EEG artifacts (e.g., from muscle movement) are discarded. There are numerous options for quantifying ERP signals (see Gratton & Fabiani, 2017), but the most common involve measuring the average amplitude within a researcher-defined segment of the ERP waveform (generally a component of interest) and/or the post-event latency at which component amplitude peaks. (For a more extensive consideration of the

neurophysiological origins, measurement processes and inferential considerations important for ERPs, see Luck, 2014).

Psychometric Properties of ERPs

Validity

A measure's validity indexes the extent to which it assesses the construct it is intended to assess. With respect to ERPs, validity refers to the psychological significance of a given component or voltage deflection. Given that the ERP waveform represents the summation of a number of different underlying components (see Luck, 2014), each theoretically reflecting a different neurocognitive process or processes, psychophysiolgists must devise various approaches to disentangle the contributions of these processes and their psychological significance. The simplest approach is to experimentally manipulate the engagement of a process and then measure its effects. For example, studies showing that rare stimuli elicit larger amplitude in some component (e.g., the P3) than frequent stimuli provide evidence that the component may index novelty detection (e.g., Friedman, Cycowicz, & Gaeta, 2001). Further manipulations can then determine whether the component responds to novelty per se, or if novel stimuli represent some more general property (e.g., motivational significance; see Nieuwenhuis, Aston-Jones, & Cohen, 2005) responsible for the component's variation.

Sometimes, a known-groups validity approach (Cronbach & Meehl, 1955) is used. If two groups differ along a psychological trait or construct and a particular physiological response is thought to be linked to that construct, then a group difference should be evident in that physiological response. Consider the reward positivity (RewP), an ERP component elicited in response to performance feedback (e.g., winning money during a gambling task) that has been linked to reward sensitivity (Proudfit, 2015). Using a known-groups validity approach, Foti and colleagues (2014) found reduced RewP responses to rewarding feedback among individuals with major depressive disorder who exhibited blunted positive affective reactivity. This finding increases the RewP's validity as a measure of reactivity to reward.

Reliability

Reliability refers to the overall consistency of a measure and represents the upper limit of that measure's validity (e.g., Nunnally & Bernstein, 1994). Two types of reliability are of interest for ERPs: internal and retest reliability. Conceptually, *internal reliability* measures the extent to which responses elicited by trials of the same type are interchangeable within a given task. This is typically assessed in ERPs using split-half reliability (cf., Thigpen, Kappenman, & Keil, 2017), where waveforms recorded on odd and even trials for each subject are separately averaged. The measurement of interest (e.g., component amplitude or latency) is then

computed for each waveform within each subject and their degree of association is tested using the Pearson product-moment correlation (r) and/or the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979).

While the majority of ERP studies examine effects of within-subjects manipulations, interest in individual differences in the psychological processes indexed by ERP components is increasing. Investigation of the suitability of ERP components as measures of trait constructs requires the additional consideration of *test-retest reliability*, the degree to which an ERP measure is stable over time. This idea is closely tied to recent interest in the use of ERPs (and other neurophysiological measures) as “neuromarkers” for various clinical phenotypes (e.g., Kwako, Mommenan, Litten, Koob, & Goldman, 2016; Olvet & Hajcak, 2008; Williams et al., 2005), but the same logic applies to using such measures as markers for personality or trait dimensions (e.g., Pailing & Segalowitz, 2004). (For more extensive discussions of psychometric principles applied to psychophysiology, see Clayson & Miller, 2017; Strube & Newman, 2017; Thigpen et al., 2017).

Applying ERPs to Information Processing

As computer memory and hard-drive space has become less expensive, it has become commonplace for ERP researchers to record EEG continuously throughout experimental tasks. A major advantage of this approach (over recording only during stimulus- or response-defined epochs) is that it permits a researcher to track information processing across multiple events within a given trial and/or examine changes in resting EEG between trials. For example, in addition to the cognitive operations elicited by a stimulus itself, it could be of theoretical interest to understand pre-stimulus, anticipatory processes (e.g., Ruge, Jamadar, Zimmermann, & Karayanidis, 2013), response preparation (e.g., Smolders & Miller, 2012), and post-response processes (e.g., Chang, Chen, Li, & Li, 2014), and, in some paradigms, processes elicited by performance feedback (e.g., Proudfit, 2015). In essence, and in contrast to the relatively impoverished information provided by RT, ERPs make it possible to track the mental processes contributing to behavioral responses from pre-perceptual anticipation through perception and response, and beyond. This point is illustrated in Figure 5.2, which lists five types of processes for which ERPs can be used to elucidate mental events that could be of interest in a given experimental trial. The following sections consider each of these processes in turn.

Anticipatory Processes

Contingent Negative Variation (CNV)

In some paradigms, researchers want to determine whether learning has occurred or expectations concerning an upcoming stimulus have been developed. Research

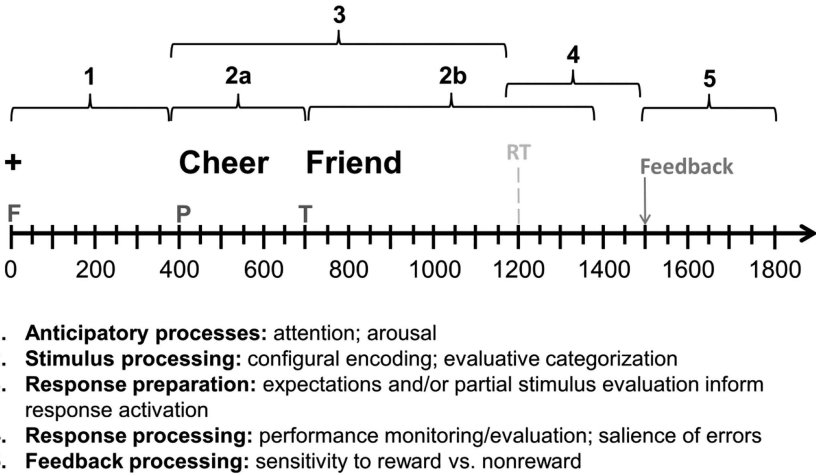


FIGURE 5.2 Mental Events During Affective Priming Plausibly Elucidated by ERPs

Note: Timing of a trial from a hypothetical affective priming task in which a fixation cross (F) signals trial onset and is followed after 400 ms by a prime word (P), which is followed after 300 ms by a target word (T). Participants must classify the target (positive or negative) as quickly as possible by making a right- or left-hand button press. Reaction time (RT; 500 ms on this trial) and accuracy are recorded, and feedback indicating whether the response was accurate and fast enough is provided 300 ms later. Recording EEG provides a temporally precise way to measure each of five processes, indicated by brackets delineating their occurrence, during each trial.

has shown that preparing a movement or waiting for the onset of a stimulus is accompanied by a slowly developing negative voltage in the EEG (see Brunia, van Boxtel, & Böcker, 2012). This negativity reflects one of several processes depending on the context in which it is elicited. For example, the CNV has been characterized as reflecting the successful transition from evaluating the potential for reward (on the basis of a cue) to motivated approach behavior during reward anticipation (Novak & Foti, 2015). RT is reduced as CNV amplitude increases (see Haag & Brunia, 1985), indicating some functional significance of CNV-related brain activity for task performance. Reviews of CNV results across numerous paradigms have led to the conclusion that the CNV reflects a combination of motor preparation and anticipatory attention (Brunia et al., 2012). Attempts to separate these two influences led to the discovery of the slow negativity described next.

Stimulus-Preceding Negativity (SPN)

In some cases imperative stimuli may not require a behavioral response, but if timing of events within a task is predictable an anticipatory negative voltage leading up to stimulus onset—the SPN—can still be observed. Initially, the SPN was

introduced as a way to describe differences between the CNV and a potential strictly reflecting movement preparation (the so-called *Bereitschaftspotential*; Kornhuber & Deecke, 1965). van Boxtel and Böcker (2004) described three types of stimuli likely to be preceded by this SPN: (a) performance feedback; (b) instructions for an upcoming task; and (c) affective stimuli. At the most basic level, in tasks that require no behavioral response measuring the SPN is useful as a way of determining whether or not participants are paying attention. This is especially useful if the stimuli themselves are affective, as such stimuli tend to elicit greater anticipatory attention (see Donkers, Nieuwenhuis, & van Boxtel, 2005). Thus, the SPN is sensitive to experimental manipulations but also can distinguish pre-existing groups on relevant dimensions. For example, Fleming and Bartholow (2017) recruited groups of participants representing high and low risk for alcohol-related problems and measured their EEG while they completed a conditioned learning task. As predicted, the SPN preceding delivery of a predicted alcohol odor (but not a predicted nonalcohol odor) was larger in the high-risk group.

Stimulus Processing

Most commonly, researchers are interested in the neurocognitive processes elicited by stimuli representing constructs of theoretical interest. The components described next have been well utilized for this purpose.

N170

Most social interactions begin with face perception, and much of our social communication—conveying moods, emotions, and reactions—is accomplished through facial expressions. Early psychophysiological studies of face processing suggested that regions of the inferior temporal lobe appear specialized for the processing of human faces (Kanwisher, McDermott, & Chun, 1997). The N170 component is a negative deflection observed over the occipital-temporal region ~170 ms following the onset of a face (Bentin, Allison, Puce, Perez, & McCarthy, 1996) and is known to arise from activity in this area (e.g., Corrigan et al., 2009). Extensive experimentation has shown that the N170 represents the configural encoding of faces (Rossion & Jacques, 2011) and therefore can index the degree to which an object is spontaneously categorized as a human face. Although very few published studies to date have documented the reliability of the N170, current evidence suggests excellent internal reliability ($r_s = .77-.97$, ICCs = $.77-.90$; Cassidy et al., 2012) based on split-half reliability analyses within tasks.² N170 also appears to remain stable over a one-month period ($r_s = .82-.85$, ICCs = $.75-.95$; Cassidy et al., 2012; Huffmeijer, Bakermans-Kranenburg, Alink, & van Ijzendoorn, 2014).

There has been intense debate over whether social and motivational factors can have a top-down influence on perceptual experience, including face perception (see Firestone & Scholl, 2016). Classic models hold that the configural

encoding of faces is a purely stimulus-driven, bottom-up process, occurring too early to be influenced by top-down factors (Bruce & Young, 1986). However, studies have shown that self-reported judgment of faces is affected by top-down variables, such as context (e.g., Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011). Behavioral and hemodynamic neuroimaging measures of face processing are limited in their ability to resolve this issue, but the very early emergence of the N170 makes it a good candidate to weigh-in on this debate. Several recent studies have shown that face encoding, as indicated by the N170, may be moderated by a host of top-down social-motivational factors, including minimal group distinctions (Figure 5.3; Ratner & Amodio, 2013), feelings of power (Schmid & Amodio, 2017), and experimental task demands (Senholzi & Ito, 2013). Such data present the strongest case to date for top-down effects on initial face encoding (for review, see Kawakami, Amodio, & Hugenberg, 2017).

Little research has documented individual differences in the N170. Amodio and colleagues have reported that both implicit prejudice (Ofan, Rubin, & Amodio, 2011) and dispositional social anxiety (Ofan, Rubin, & Amodio, 2014) covary with the degree to which the N170 differentiates White from Black faces, but much more data is needed before strong claims can be made regarding the usefulness of the N170 as an index of individual differences.

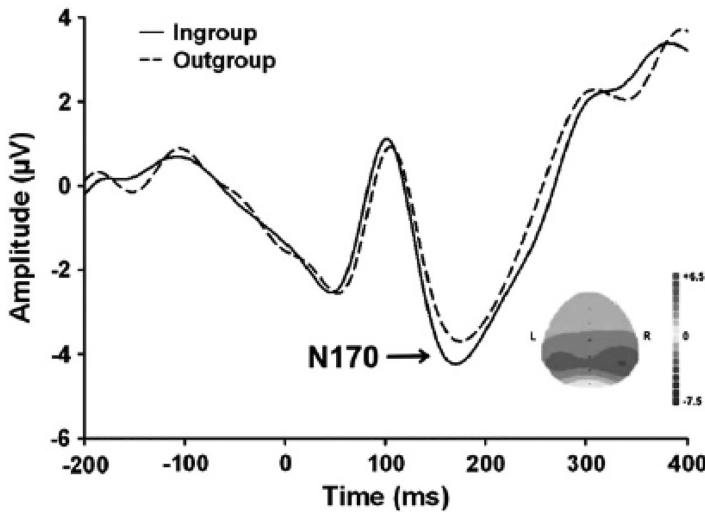


FIGURE 5.3 N170 Amplitude as an Index for Face Encoding

Note: N170 amplitude elicited by faces during a classic minimal groups experiment is larger (more negative) for arbitrarily assigned ingroup than outgroup members. Time = 0 on the x-axis represents face presentation onset. Reprinted with permission from Ratner and Amodio (2013).

P3

First described in the mid-1960s (e.g., Chapman & Bragdon, 1964), the P3 (or P300, or P3b³) is perhaps the most widely studied ERP component in the literature. The P3 is a positive-going deflection maximal at midline parietal scalp locations, which peaks 300–800 ms after the onset of a task-relevant stimulus (see Figure 5.1) (for review, see Nieuwenhuis et al., 2005; Polich, 2007). Because P3 amplitude is enhanced for novel or infrequent stimuli (e.g., Friedman et al., 2001), an early, dominant theory linked P3 amplitude to working memory updating (Donchin, 1981; Donchin & Coles, 1988). That is, when the category of a stimulus differs from that represented by previously attended stimuli, the currently activated mental representation requires updating; this process has been linked to enhanced P3. Of particular interest for social psychologists, this logic also applies when stimulus categories differ only on subjectively determined qualities. In an early demonstration of this property, Cacioppo, Crites, Berntson, and Coles (1993) found enhanced P3s to targets participants had previously indicated they liked (e.g., carrots), compared to targets they did not like (e.g., Brussels sprouts), revealing the utility of the P3 as a tool to understand internally held attitudes and evaluations. Current theory links the P3 with the incentive value (Begleiter, Porjesz, Chou, & Aunon, 1983) or motivational significance (Nieuwenhuis et al., 2005) of an eliciting stimulus. Numerous studies have found that affective or arousing stimuli elicit larger P3 amplitude compared to neutral stimuli (reviewed in Olofsson, Nordin, Sequeira, & Polich, 2008), supporting this general idea.⁴ This theory explains the P3's sensitivity to novelty as a reflection of rare stimuli's motivational significance.

Perhaps of greater interest in the current context, the latency at which the P3 peaks has been shown to reflect the speed or ease with which stimulus evaluation occurs. Considerable research shows that P3 latency increases as stimulus evaluation becomes more difficult (e.g., Kutas, McCarthy, & Donchin, 1977; see also Coles et al., 1995). Critically, P3 latency is largely independent of overt response activation. In a convincing demonstration of this property, McCarthy and Donchin (1981) independently manipulated stimulus discriminability and stimulus-response compatibility in a choice RT task. They found that although reaction time was affected by both discriminability and stimulus-response compatibility, P3 latency was affected only by stimulus discriminability. Thus, not only can P3 latency augment RT and provide insight into pre-response stimulus categorization (see Dien, Spencer, & Donchin, 2004), this measure also can provide such information in paradigms requiring no behavioral response.

P3 amplitude has acceptable internal reliability ($r_s = .54-.93$, ICCs = .50–.53; Cassidy et al., 2012; Fabiani, Gratton, Karis, & Donchin, 1987; Hämmerer, Li, Völkle, Müller, & Lindenberger, 2013; Kinoshita, Inoue, Maeda, Nakamura, & Morita, 1996; Polich, 1986; Walhovd & Fjell, 2002). In adolescents, Segalowitz

and Barnes (1993) found somewhat lower internal reliability ($r = .48$) when based on 40 target trials (though this might not be enough trials for a stable estimate).

While most P3 work has examined effects of experimental manipulations, P3 also has been shown to correlate with individual differences. For example, based on the idea that the P3 is larger when evaluative categorization of a target differs from a preceding context (Cacioppo et al., 1993), Ito and colleagues (2004) measured P3 while White participants completed a task in which faces (White and Black) were shown infrequently amid strings of positive or negative images. Ito et al. found that greater explicit anti-Black attitudes were associated with larger P3s for Black (vs. White) targets when the affective context was positive (i.e., Black faces are more evaluatively divergent from positive images than are White faces); the opposite pattern emerged when the affective context was negative.

Regarding the suitability of the P3 as a trait measure, P3 amplitude has acceptably stable retest reliability in a variety of tasks ($r_s = .53-.85$, ICCs = $.54-.92$) among participants from across the lifespan, including children (Hämmerer et al., 2013), adolescents (Williams et al., 2005), young and middle-aged adults (Fabiani et al., 1987), and elderly adults (Walhovd & Fjell, 2002). However, this appears not to hold in some contexts, including P3 elicited by olfactory stimuli (Thesen & Murphy, 2002), P3s measured during highly complex cognitive tasks (Schall, Catts, Karayanidis, & Ward, 1999), and P3s elicited by targets presented at highly predictable intervals (Sandman & Patterson, 2000). This evidence suggests that stimuli with the greatest motivational significance (i.e., those that are infrequent and unpredictable) elicit P3s that are the most stable over time.

Response Preparation

Lateralized Readiness Potential (LRP)

When participants use one hand to make a behavioral response, a negative potential can be observed from electrodes placed over the motor cortex (central scalp, a few centimeters from midline) contralateral to the responding hand. The source of this *LRP* has been localized to primary motor cortex (Eimer, 1998; Miller & Hackley, 1992), and its onset begins before the response is emitted. Moreover, if participants have information concerning which response (left or right) will be required for an upcoming stimulus, the LRP can be observed even before stimulus onset (e.g., Kutas & Donchin, 1980). These properties suggest that LRP onset reflects the time at which response preparation is initiated in the brain (see Smulders & Miller, 2012). Thus, when combined with simultaneous measures of stimulus evaluation that can be dissociated from response-related processes, such as P3 latency, the LRP can provide millisecond-level resolution of the neural basis of stimulus-response associations.

These properties of the LRP make it very useful for understanding two phenomena that are of particular interest to experimental social psychologists. First,

the emergence of the LRP can establish the point at which response preparation can begin. In this way, the LRP has been used to demonstrate that partial response activation can occur before analysis of a stimulus is complete (see Coles, Gratton, & Donchin, 1988; Miller & Hackley, 1992), contrary to discrete-stage models of processing (e.g., Sanders, 1980; Sternberg, 1969), which hold that contingent stages operate in strict temporal succession, such that each process must finish before the next can begin. The LRP also has been applied to understand the mental operations responsible for affective priming effects. For example, Bartholow, Riordan, Sauls, and Lust (2009) recorded EEG while participants performed an evaluative priming task (Fazio et al., 1986) and found evidence that responses were activated by prime words, prior to target word onset (see also Eder, Leuthold, Rothermund, & Schweinberger, 2012). Moreover, the probability of congruent trials strongly affected response activation as indicated by the LRP: when targets were highly likely to be prime-congruent, preparation of a congruent response was evident prior to target onset; when targets were highly likely to be prime-incongruent, preparation of an incongruent response was evident in the LRP during the prime-to-target interval (see Figure 5.4). These findings helped to establish that response preparation and response conflict (when the prepared response conflicts with the one required by a target) are critical components of the well-known affective congruency effect in evaluative priming (also see Klinger, Burton, & Pitts, 2000).

Response Processing

Error-Related Negativity (ERN)

Cognitive control, or the ability to focus attention on relevant information while ignoring the influence of distraction (see Braver, 2012), is important to many aspects of social behavior (e.g., see Amodio, 2011; Bartholow, 2010). One important aspect of effectively implementing cognitive control is the ability to monitor ongoing performance so that adjustments can be made when cognitive control fails. The ERN, a negative-going deflection generated in the dorsal anterior cingulate cortex (dACC; e.g., van Veen & Carter, 2002), occurs simultaneously with the commission of errors and is thought to play a crucial role in this performance-monitoring process (for review, see Weinberg, Riesel, & Hajcak, 2012). Specifically, the ERN reflects the activation of a *salience network* sensitive to conflict (e.g., between actions and intentions, or between currently implemented and optimal strategies; see Botvinick & Cohen, 2014), which is crucial for instigating performance adjustments when control is threatened (Ham, Leff, de Boissezon, Joffe, & Sharp, 2013; Hoffstaedter et al., 2014). Within this context the ERN can be said to index the degree to which errors are distressing, and therefore, salient (e.g., Bartholow, Henry, Lust, Sauls, & Wood, 2012; Hajcak & Foti, 2008; Inzlicht, Bartholow, & Hirsh, 2015).

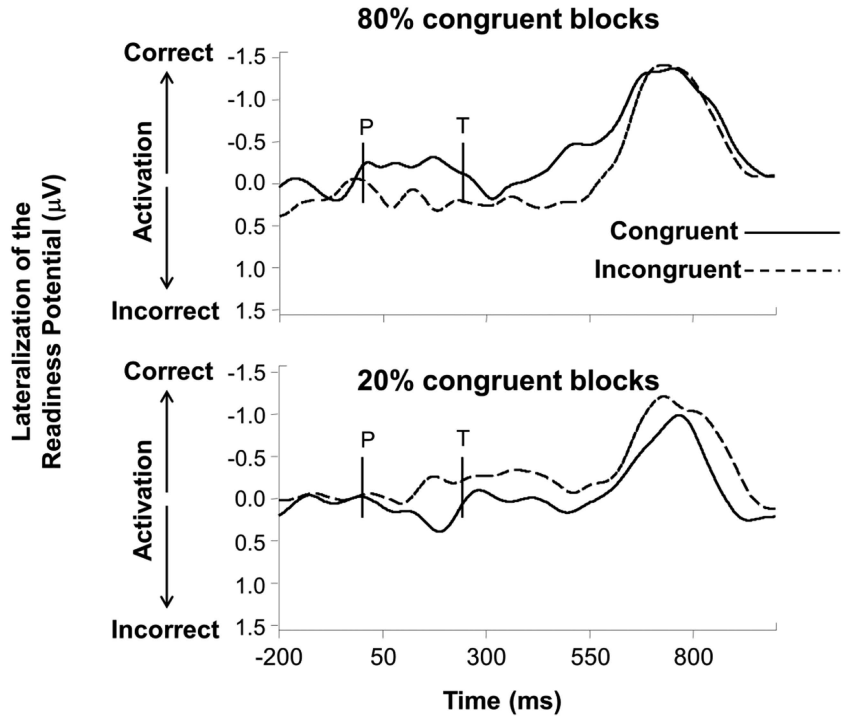


FIGURE 5.4 The Lateralized Readiness Potential (LRP) Measured on Congruent and Incongruent Trials During an Affective Priming Task

Note: The probability of congruent and incongruent trials was manipulated across blocks, such that participants expected congruent trials in the 80% congruent blocks and expected incongruent trials in the 20% congruent blocks. The amplitude and polarity of the LRP between prime onset (P) and target onset (T) indicates relative response activation elicited by the primes, before the target has appeared. The formula used to derive the LRP is applied with reference to the correct response hand in each condition, such that negative voltage deflections indicate that participants were preparing to activate the hand needed to make the correct response, whereas positive voltage deflections indicate that participants were inadvertently preparing to activate the hand that would produce an incorrect response. These LRPs show that motor cortex was preferentially activated to initiate a valence-congruent response prior to target onset when congruent targets were expected (80% congruent blocks), but was preferentially activated to initiate a valence-incongruent response prior to target onset when incongruent targets were expected (20% congruent blocks). Reprinted with permission from Bartholow et al. (2009).

The ERN has proven useful for understanding implicit racial bias. Errors indicative of unconsciously endorsing stereotypes linking Black men with armed violence elicit larger ERNs than errors that are free from biased implications (see Figure 5.5; e.g., Amodio, Devine, Harmon-Jones, 2004; Bartholow et al., 2012). This is particularly the case for individuals who are high in internal motivation to be unbiased (Amodio, Devine, & Harmon-Jones, 2008), suggesting that racially

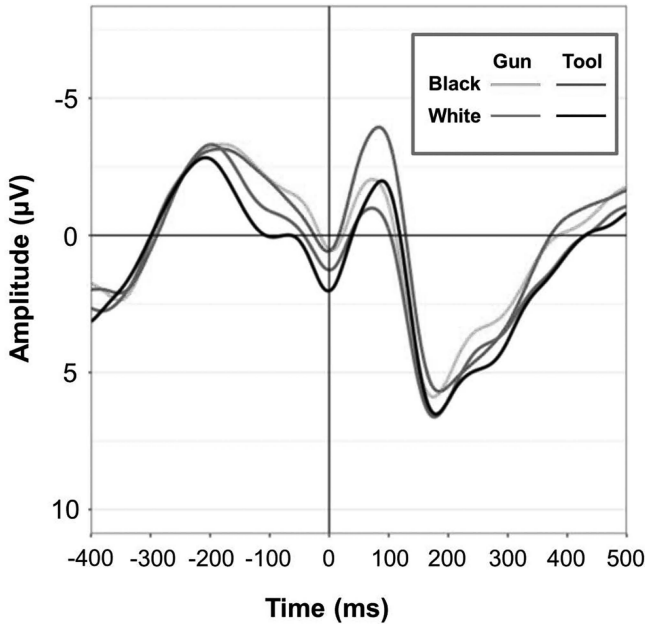


FIGURE 5.5 Response-locked ERPs Recorded on Incorrect Responses During the Weapons Identification Task (WIT; Payne, 2001)

Note: The Weapons Identification Task is a fast-paced, choice RT task where a face (Black or White race) precedes an object that participants must categorize as either a tool or a gun via button press. The ERN is the large, negative-going deflection emerging just after the button press (Time = 0 ms), and is larger when participants accidentally classify a tool as a gun following a Black face. Data used in this figure were first published in Volpert-Esmond et al. (2017).

biased errors are particularly salient to them. Moreover, the larger the ERNs elicited on these bias-related trials, the more control of bias an individual demonstrates overall, consistent with the ERN's role in performance monitoring (see Yeung, Botvinick, & Cohen, 2004).

Considerable effort has been made to demonstrate the psychometric properties of the ERN. Overall, the internal reliability of the ERN can be variable across different tasks and age groups ($r_s = .35-.88$, ICCs = .64–.76; Cassidy et al., 2012; Foti, Kotov, & Hajcak, 2013; Meyer, Bress, & Proudfit, 2014; Olvet & Hajcak, 2009a; Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013). Several researchers have examined internal agreement of the ERN as a function of the number of error trials. The recommended number of trials required to obtain adequate internal agreement (often estimated using Cronbach's $\alpha > .70$) varies widely, from as few as 5–6 errors (Foti et al., 2013; Olvet & Hajcak, 2009b; Pontifex et al., 2010) to as many as 30 or more errors (Baldwin, Larson, & Clayson, 2015; Meyer, Riesel, & Proudfit, 2013; Meyer et al., 2014), largely depending on the type of task (see Riesel et al., 2013).

In addition to being responsive to experimentally manipulated stimuli within subjects, the ERN may help to explain inter-individual variability in the control of racial bias (Amodio et al., 2008), liberal-conservative political orientation (Amodio, Jost, Master, & Yee, 2007), high negative affect (Hajcak, McDonald, & Simons, 2004), and worry (Hajcak, McDonald, & Simons, 2003). Additionally, the ERN has been associated with anxiety (for meta-analysis, see Moser, Moran, Schroder, Donnellan, & Yeung, 2013) and obsessive compulsive disorder (e.g., Carrasco et al., 2013; Riesel et al., 2014), leading to the suggestion that the ERN could be considered a psychiatric endophenotype (Olvet & Hajcak, 2008; Proudfit, Inzlicht, & Mennin, 2013).

The ERN has demonstrated sufficient retest reliability within individuals over time ($r_s = .57-.75$, ICCs = $.54-.74$; Cassidy et al., 2012; Meyer et al., 2014; Olvet & Hajcak, 2009a; Weinberg & Hajcak, 2011), although not in all studies ($r_s = .49$ and $.40$ in Larson, Baldwin, Good, & Fair, 2010, and Segalowitz et al., 2010, respectively; ICC = $.38$ in Segalowitz et al., 2010). The ERN has shown this trait-like stability over time periods as long as two years.

Feedback Processing

RewP

When participants receive external feedback concerning the outcome of a prior choice, an apparently negative-going deflection maximal at fronto-central electrodes can be observed ~250 ms following feedback onset (Miltner, Braun, & Coles, 1997). Initially dubbed the feedback-related negativity (FRN) given its more negative voltage following negative versus positive feedback (e.g., Hajcak, Moser, Holroyd, & Simons, 2006), this component more recently has been rechristened the *reward positivity* (RewP). Rather than being a negative-going response elicited by negative evaluative feedback, the deflection instead has been shown to represent the absence of a positive-going response when reward-related or positive evaluative information is lacking (Proudfit, 2015). RewP is hypothesized to reflect what is known as the *prediction error*; that is, the degree to which feedback deviates from expectations (e.g., Gehring & Willoughby, 2002; Holroyd & Coles, 2002). When feedback indicates that the result is worse than expected, a more “negative dip” in the positive-going RewP deflection is observed (see Figure 5.6). Relatedly, factors such as valence, magnitude, probability, and type of reinforcement all have been shown to influence RewP amplitude in reinforcement learning paradigms (see Sambrook & Goslin, 2015).

Recent work using the RewP has investigated reward expectancies in the context of social economic decision-making games. One frequently used paradigm is the Ultimatum Game (UG; Güth, Schmittberger, & Schwarze, 1982), which emphasizes judgments of fairness. During this two-player game, one player is given a sum of money to divide between him- or herself and the other player.

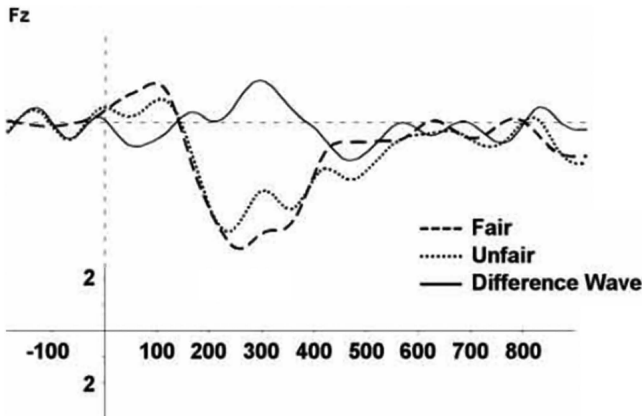


FIGURE 5.6 RewP in Response to Fair and Unfair Offers (Presented at Time 0 on the X-Axis) Made During the Ultimatum Game

Note: Unfair offers (i.e., outcomes that were worse than expected) elicited a negative dip in the positive-going deflection ~300 ms following the offer. Reprinted with permission from Boksem and De Cremer (2010).

Once an offer is made, the other player can either “accept” the offer (in which case, the money is distributed according to the proposal) or “reject” the offer, which results in both players receiving nothing. Previous research using the UG has found that RewP amplitude is more positive when participants receive fair offers compared to unfair offers (e.g., Boksem & De Cremer, 2010). Furthermore, RewP amplitude predicts both the rejection of subsequent offers and the degree of negative affect associated with unfair offers (Hewig et al., 2011).

Additionally, accumulating evidence indicates that RewP amplitude is also sensitive to social expectancies within the context of economic decision-making. For example, Osinsky and colleagues (2014) found that RewP amplitude was affected by both the magnitude of the offer received and the learned reputation of the other players based on offers they had made previously during the game. In another example, Chen and colleagues (2012) tested the effect of facial attractiveness and found a larger difference in RewP amplitude elicited by positive and negative feedback when playing against more attractive partners, consistent with the stereotype that attractive people are more trustworthy and therefore unfair offers elicit a greater prediction error.

Researchers have only recently begun to explore the psychometric properties of RewP amplitude (Bress et al., 2015; Huffmeijer et al., 2014; Levinson, Speed, Infantolino, & Hajcak, 2017; Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011; Segalowitz et al., 2010). Thus far, the RewP has demonstrated good internal reliability in response to monetary losses ($r = .71-.90$) and gains ($r = .79-.89$) in both undergraduates (Levinson et al., 2017) and children

(Bress et al., 2015). Acceptable internal reliability can be achieved with as few as 20 feedback trials in younger participants (Marco-Pallares et al., 2011; Levinson et al., 2017), although as many as 50 trials may be required for older participants (Marco-Pallares et al., 2011).

Individual differences reflected in the RewP are linked to trait measures of reward sensitivity (e.g., Bress, Smith, Foti, Klein, & Hajcak, 2012), and could be a biomarker for low positive affect that leads to depression (see Proudfit, 2015). However, investigations of retest reliability of the RewP have shown somewhat mixed results. In children, RewP amplitude in response to monetary losses and gains during a gambling task was found to have acceptable retest reliability for both gains ($r_s = .45-.67$, ICC = .62) and losses ($r_s = .64-.71$, ICC = .81) over one week to two years (Bress et al., 2015; Levinson et al., 2017). Similar retest reliability has been demonstrated when RewP is elicited during a driving simulation video game (with feedback indicating a crash; Segalowitz et al., 2010). However, when measured in response to feedback indicating response accuracy during a flanker task RewP retest reliability was poor (ICCs = .14-.40; Huffmeijer et al., 2014), possibly because the prediction error signal in tasks like the flanker is generated internally, at the time of the response, and therefore feedback is less informative (Holroyd & Coles, 2002). Some research suggests good retest reliability of RewP across experimentally manipulated temporary states, such as sleep deprivation ($r_s = .52-.84$, ICCs = .55-.82; Segalowitz et al., 2010). Additionally, RewP elicited during a simulated driving task had stable retest reliability ($r_s = .53-.77$) in adolescent boys across different contexts (alone vs. with friends present; Segalowitz et al., 2010).

Conclusion

ERPs represent an extremely powerful tool with unrivaled temporal specificity for examining sociocognitive processes. The rich, multivariate nature of ERP data provide numerous opportunities to address questions on the neurocognitive and affective mechanisms driving phenomena at the heart of many social and personality theories. Although the ERP technique has been used for many decades in hundreds of cognitive and clinical psychology labs, and although the prominence of ERPs—and other neuroimaging techniques—in social and personality psychology has increased dramatically in recent years, presently only a handful of social-personality labs incorporate ERPs into their research programs. In our view, the future of social cognition depends on the ability to validly and precisely probe implicit mental processes and their connections with experience and behavior, and ERPs offer the clearest path forward in this regard. Or, put another way, “Given that cognitive processes are implemented by the brain, it seems to make sense to explore the possibility that measures of brain activity can provide insights into their nature” (Rugg & Coles, 1995, p. 27).

Notes

- 1 Numerous other sources have elaborated the considerations needed to increase the quality of inferences in psychophysiological research (e.g., Amodio, 2010; Cacioppo, Tassinary, & Berntson, 2007; Hutzler, 2014). Because of this, we refrain from discussing it further.
- 2 All split-half reliability estimates of r presented here have been adjusted with the Spearman-Brown prophesy formula.
- 3 Researchers distinguish between two different P3s that occur simultaneously: the more frontally maximal P3a and the posterior P3b (for review, see Polich, 2007).
- 4 A related late-latency, positive-going deflection, the late positive potential (LPP), has been strongly implicated in affective stimulus processing. For a review, see Hajcak, Weinberg, MacNamara, and Foti (2012).

References

- Allison, T., Wood, C. C., & McCarthy, G. M. (1986). The central nervous system. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 5–25). New York, NY: Guilford Press.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256–274.
- American Encephalographic Society. (1994). Guideline thirteen: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology*, 11, 111–113.
- Amodio, D. M. (2010). Can neuroscience advance social psychological theory? Social neuroscience for the behavioral social psychologist. *Social Cognition*, 28, 695.
- Amodio, D. M. (2011). Self-regulation in intergroup relations: A social neuroscience framework. In A. Todorov, S. T. Fiske, & D. Prentice (Eds.), *Social neuroscience: Toward understanding the underpinnings of the social mind* (pp. 101–122). New York, NY: Oxford University Press.
- Amodio, D. M., & Bartholow, B. D. (2011). Event-related potential methods in social cognition. In A. Voss, C. Stahl, & C. Klauer (Eds.), *Cognitive methods in social psychology* (pp. 303–339). New York, NY: Guilford Press.
- Amodio, D. M., Bartholow, B. D., & Ito, T. A. (2014). Tracking the dynamics of the social brain: ERP approaches for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, 9, 385–393.
- Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, 94, 60–74.
- Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, 15, 88–93.
- Amodio, D. M., Jost, J. T., Master, S. L., & Yee, C. M. (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience*, 10, 1246–1247.
- Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. *Psychophysiology*, 52(6), 790–800.
- Bartholow, B. D. (2010). On the role of conflict and control in social cognition: Event-related brain potential investigations. *Psychophysiology*, 47, 201–212.

- Bartholow, B. D., Henry, E. A., Lust, S. A., Saults, J. S., & Wood, P. K. (2012). Alcohol effects on performance monitoring and adjustment: Affect modulation and impairment of evaluative cognitive control. *Journal of Abnormal Psychology, 121*(1), 173.
- Bartholow, B. D., Riordan, M. A., Saults, J. S., & Lust, S. A. (2009). Psychophysiological evidence of response conflict and strategic control of responses in affective priming. *Journal of Experimental Social Psychology, 45*, 655–666.
- Begleiter, H., Porjesz, B., Chou, C. L., & Aunon, J. I. (1983). P3 and stimulus incentive value. *Psychophysiology, 20*(1), 95–101.
- Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*, 551–565.
- Boksem, M., & Cremer, D. (2010). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. *Social Neuroscience, 5*(1), 118–128.
- Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science, 38*, 1249–1285.
- Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106–113.
- Bress, J., Meyer, A., & Proudfit, G. H. (2015). The stability of the feedback negativity and its relationship with depression during childhood and adolescence. *Development and Psychopathology, 27*, 1285–1294.
- Bress, J. N., Smith, E., Foti, D., Klein, D. N., & Hajcak, G. (2012). Neural response to reward and depressive symptoms in late childhood to early adolescence. *Biological Psychology, 89*(1), 156–162.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.
- Brunia, C. H., van Boxtel, G. J., & Böcker, K. B. (2012). Negative slow waves as indices of anticipation: The Bereitschaftspotential, the contingent negative variation, and the stimulus-preceding negativity. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford handbook of event-related potential components* (pp. 189–207). New York: Oxford University Press.
- Cacioppo, J. T., Crites, S. L., Jr., Berntson, G. G., & Coles, M. G. H. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science, 4*, 108–112.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Psychophysiological science: Interdisciplinary approaches to classic questions about the mind. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 1–16). New York, NY: Cambridge University Press.
- Carrasco, M., Harbin, S. M., Nienhuis, J. K., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and unaffected siblings. *Depression and Anxiety, 30*(1), 39–46.
- Cassidy, S. M., Robertson, I. H., & O'Connell, R. G. (2012). Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology, 49*(5), 659–664.
- Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual process theories in social psychology*. New York, NY: Guilford Press.
- Chang, A., Chen, C.-C., Li, H.-H., & Li, C.-S. R. (2014). Event-related potentials for post-error and post-conflict slowing. *PLoS ONE, 9*(6), e99909. Retrieved from <https://doi.org/10.1371/journal.pone.0099909>
- Chapman, R. M., & Bragdon, H. R. (1964). Evoked responses to numerical and non-numerical visual stimuli while problem solving. *Nature, 203*, 1155–1157.

- Chen, J., Zhong, J., Zhang, Y., Li, P., Zhang, A., Tan, Q., & Li, H. (2012). Electrophysiological correlates of processing facial attractiveness and its influence on cooperative behavior. *Neuroscience Letters*, 517(2), 65–70.
- Clayson, P. E., & Miller, G. A. (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, 111, 57–67.
- Coles, M. G., Gratton, G., Bashore, T. R., Eriksen, C. W., & Donchin, E. (1985). A psychophysiological investigation of the continuous flow model of human information processing. *Journal of Experimental Psychology: Human Perception and Performance*, 11(5), 529–553.
- Coles, M. G., Gratton, G., & Donchin, E. (1988). Detecting early communication: Using measures of movement-related potentials to illuminate human information processing. *Biological Psychology*, 26(1), 69–89.
- Coles, M. G., Smid, H. G., Scheffers, M. K., & Otten, L. J. (1995). Mental chronometry and the study of human information processing. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 86–131). New York: Oxford University Press.
- Corrigan, N. M., Richards, T., Webb, S. J., Murias, M., Merkle, K., Kleinhans, N. M., Johnson, L. C., Poliakov, A., Aylward, E., & Dawson, G. (2009). An investigation of the relationship between fMRI and ERP source localized measurements of brain activity during face processing. *Brain Topography*, 22, 83–96.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.
- Dien, J., Spencer, K. M., & Donchin, E. (2004). Parsing the late positive complex: Mental chronometry and the ERP components that inhabit the neighborhood of the P300. *Psychophysiology*, 41, 665–678.
- Donchin, E. (1981). Surprise! . . . surprise? *Psychophysiology*, 18(5), 493–513.
- Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, 11, 354–356.
- Donchin, E., Karis, D., Bashore, T. R., Coles, M. G. H., & Gratton, G. (1986). Cognitive psychophysiology and human information processing. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 244–267). New York, NY: Guilford Press.
- Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, 30, 412–431. [Translation of: Die Schnelligkeit psychischer Prozesse, first published in 1868].
- Donkers, F. C., Nieuwenhuis, S., & Van Boxtel, G. J. (2005). Mediofrontal negativities in the absence of responding. *Cognitive Brain Research*, 25(3), 777–787.
- Eder, A. B., Leuthold, H., Rothermund, K., & Schweinberger, S. R. (2012). Automatic response activation in sequential affective priming: An ERP study. *Social Cognitive and Affective Neuroscience*, 7(4), 436–445.
- Eimer, M. (1998). The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, Instruments, & Computers*, 30(1), 146–156.
- Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. *Advances in Psychophysiology*, 2, 1–78.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50(2), 229–238.

- Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, 20, 1–77.
- Fleming, K. A., & Bartholow, B. D. (2017). *Conditioning the incentive value of alcohol in drinkers at risk for alcohol use disorder*. Manuscript in preparation.
- Foti, D., Carlson, J. M., Sauder, C. L., & Proudfit, G. H. (2014). Reward dysfunction in major depression: Multimodal neuroimaging evidence for refining the melancholic phenotype. *NeuroImage*, 101, 50–58.
- Foti, D., Kotov, R., & Hajcak, G. (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. *Journal of Abnormal Psychology*, 122(2), 520–531.
- Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PLoS ONE*, 6(9), e25107.
- Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential sign of the brain's evaluation of novelty. *Neuroscience & Biobehavioral Reviews*, 25, 355–373.
- Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295(5563), 2279–2282.
- Glynn, I. (2010). *Elegance in science: Beauty in simplicity*. Oxford: Oxford University Press.
- Gratton, G., & Fabiani, M. (2017). Biosignal processing in psychophysiology: Principles and current developments. In J. T. Cacioppo, G. G. Berntson, & L. G. Tassinary (Eds.), *Handbook of psychophysiology* (4th ed., pp. 628–661). New York, NY: Cambridge University Press.
- Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367–388.
- Haagh, S. A. V. M., & Brunia, C. H. M. (1985). Anticipatory response-relevant muscle activity, CNV amplitude and simple reaction time. *Electroencephalography and Clinical Neurophysiology*, 61(1), 30–39.
- Hajcak, G., & Foti, D. (2008). Errors are aversive: Defensive motivation and the error-related negativity. *Psychological Science*, 19(2), 103–108.
- Hajcak, G., McDonald, N., & Simons, R. F. (2003). Anxiety and error-related brain activity. *Biological Psychology*, 64(1), 77–90.
- Hajcak, G., McDonald, N., & Simons, R. F. (2004). Error-related psychophysiology and negative affect. *Brain and Cognition*, 56(2), 189–197.
- Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, 71(2), 148–154.
- Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2012). ERPs and the study of emotion. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potentials* (pp. 441–474). Oxford: Oxford University Press.
- Ham, T., Leff, A., de Boissezon, X., Joffe, A., & Sharp, D. J. (2013). Cognitive control and the salience network: An investigation of error processing and effective connectivity. *The Journal of Neuroscience*, 33, 7091–7098.
- Hämmerer, D., Li, S. C., Völkle, M., Müller, V., & Lindenberger, U. (2013). A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. *Psychophysiology*, 50(1), 111–123.
- Hewig, J., Kretschmer, N., Trippe, R., Hecht, H., Coles, M. G. H., Holroyd, C., & Miltner, W. (2011). Why humans deviate from rational choice. *Psychophysiology*, 48(4), 507–514.

- Hodgkin, A. L. (1964). *The conduction of the nervous impulse*. Springfield, IL: CC Thomas.
- Hoffstaedter, F., Grefkes, C., Caspers, S., Roski, C., Palomero-Gallagher, N., Laird, A. R., . . . Eickhoff, S. B. (2014). The role of anterior midcingulate cortex in cognitive motor control: Evidence from functional connectivity analyses. *Human Brain Mapping, 35*(6), 2741–2753.
- Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679.
- Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & van Ijzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology & Behavior, 130*, 13–22.
- Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *Neuroimage, 84*, 1061–1069.
- Inzlicht, M., Bartholow, B. D., & Hirsh, J. B. (2015). Emotional foundations of cognitive control. *Trends in Cognitive Sciences, 19*(3), 126–132.
- Ito, T. A., Thompson, E., & Cacioppo, J. T. (2004). Tracking the timecourse of social perception: The effects of racial cues on event-related brain potentials. *Personality & Social Psychology Bulletin, 30*(10), 1267–1280.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541.
- Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.
- Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology, 55*, 1–80.
- Kinoshita, S., Inoue, M., Maeda, H., Nakamura, J., & Morita, K. (1996). Long-term patterns of change in ERPs across repeated measurements. *Physiology & Behavior, 60*, 1087–1092.
- Klinger, M. R., Burton, P. C., & Pitts, G. S. (2000). Mechanisms of unconscious priming I: Response competition not spreading activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 441–455.
- Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei willkurbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und refferente potentiale. *Pflügers Archive, 284*, 1–17.
- Kutas, M., & Donchin, E. (1980). Preparation to respond as manifested by movement-related brain potentials. *Brain Research, 202*(1), 95–115.
- Kutas, M., McCarthy, G., & Donchin, E. (1977). Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science, 197*(4305), 792–795.
- Kwako, L. E., Momenan, R., Litten, R. Z., Koob, G. F., & Goldman, D. (2016). Addictions neuroclinical assessment: A neuroscience-based framework for addictive disorders. *Biological Psychiatry, 80*, 179–189.
- Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. *Psychophysiology, 47*(6), 1167–1171.

- Levinson, A. R., Speed, B. C., Infantolino, Z. P., & Hajcak, G. (2017). Reliability of the electrocortical response to gains and losses in the doors task. *Psychophysiology*, 54(4), 601–607.
- Luck, S. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.
- Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, 48(6), 852–860.
- McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science*, 211(4477), 77–80.
- Meyer, A., Bress, J. N., & Proudfit, G. H. (2014). Psychometric properties of the error-related negativity in children and adolescents. *Psychophysiology*, 51(7), 602–610.
- Meyer, A., Riesel, A., & Proudfit, G. (2013). Reliability of the ERN across multiple tasks as a function of increasing errors. *Psychophysiology*, 50(12), 1220–1225.
- Miller, J., & Hackley, S. A. (1992). Electrophysiological evidence for temporal overlap among contingent mental processes. *Journal of Experimental Psychology: General*, 121(2), 195–209.
- Miltner, W. H., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a “generic” neural system for error detection. *Journal of Cognitive Neuroscience*, 9(6), 788–798.
- Moser, J. S., Moran, T. P., Schroder, H. S., Donnellan, M. B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience*, 7, 466.
- Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, 131(4), 510.
- Novak, K. D., & Foti, D. (2015). Teasing apart the anticipatory and consummatory processing of monetary incentives: An event-related potential study of reward dynamics. *Psychophysiology*, 52(11), 1470–1482.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychological theory*. New York, NY: MacGraw-Hill.
- Ofan, R. H., Rubin, N., & Amodio, D. M. (2011). Seeing race: N170 responses to race and their relation to automatic racial attitudes and controlled processing. *Journal of Cognitive Neuroscience*, 23, 3153–3161.
- Ofan, R. H., Rubin, N., & Amodio, D. M. (2014). Situation-based social anxiety enhances the neural processing of faces: Evidence from an intergroup context. *Social Cognitive and Affective Neuroscience*, 9, 1055–1061.
- Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, 77(3), 247–265.
- Olvet, D. M., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical Psychology Review*, 28(8), 1343–1354.
- Olvet, D. M., & Hajcak, G. (2009a). Reliability of error-related brain activity. *Brain Research*, 1284, 89–99.
- Olvet, D. M., & Hajcak, G. (2009b). The stability of error-related brain activity with increasing trials. *Psychophysiology*, 46(5), 957–961.
- Osinsky, R., Mussel, P., Öhrlein, L., & Hewig, J. (2014). A neural signature of the creation of social evaluation. *Social Cognitive and Affective Neuroscience*, 9(6), 731–736.
- Pailing, P. E., & Segalowitz, S. J. (2004). The error-related negativity as a state and trait measure: Motivation, personality, and ERPs in response to errors. *Psychophysiology*, 41(1), 84–95.

- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81(2), 181–192.
- Polich, J. (1986). Normal variation of P300 from auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, 65(3), 236–240.
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10), 2128–2148.
- Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C.-T., Themanson, J. R., & Hillman, C. H. (2010). On the number of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, 47(4), 767–773.
- Posner, M. I. (1978). *Chronometric explorations of mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, 52(4), 449–459.
- Proudfit, G. H., Inzlicht, M., & Mennin, D. S. (2013). Anxiety and error monitoring: The importance of motivation and emotion. *Frontiers in Human Neuroscience*, 7, 636.
- Ratner, K. G., & Amodio, D. M. (2013). Seeing “us vs. them”: Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology*, 49(2), 298–301.
- Reed, T. E., Vernon, P. A., & Johnson, A. M. (2004). Sex difference in brain nerve conduction velocity in normal humans. *Neuropsychologia*, 42(12), 1709–1714.
- Riesel, A., Kathmann, N., & Endrass, T. (2014). Overactive performance monitoring in obsessive—compulsive disorder is independent of symptom expression. *European Archives of Psychiatry and Clinical Neuroscience*, 264(8), 707–717.
- Riesel, A., Weinberg, A., Endrass, T., Meyer, A., & Hajcak, G. (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological Psychology*, 93(3), 377–385.
- Rossion, J., & Jacques, C. (2011). The N170: Understanding the time course of face perception in the human brain. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potentials* (pp. 115–142). Oxford: Oxford University Press.
- Rugg, M. D., & Coles, M. G. H. (1995). The ERP and cognitive psychology: Conceptual issues. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related potentials and cognition* (pp. 27–39). Oxford: Oxford University Press.
- Ruge, H., Jamadar, S., Zimmermann, U., & Karayanidis, F. (2013). The many faces of preparatory control in task switching: Reviewing a decade of fMRI research. *Human Brain Mapping*, 34, 12–35.
- Sambrook, T., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, 141(1), 213.
- Sanders, A. F. (1980). Stage analysis of reaction processes. In G. E. Stelmach and J. Requin (Eds.), *Tutorials in motor behavior* (pp. 331–354). Amsterdam: North-Holland.
- Sandman, C. A., & Patterson, J. V. (2000). The auditory event-related potential is a stable and reliable measure in elderly subjects over a 3-year period. *Clinical Neurophysiology*, 111(8), 1427–1437.
- Schall, U., Catts, S. V., Karayanidis, F., & Ward, P. B. (1999). Auditory event-related potential indices of fronto-temporal information processing in schizophrenia syndromes: Valid outcome prediction of clozapine therapy in a three-year follow-up. *The International Journal of Neuropsychopharmacology*, 2(2), 83–93.
- Schmid, P. C., & Amodio, D. M. (2017). Power effects on implicit prejudice and stereotyping: The role of intergroup face processing. *Social Neuroscience*, 12, 218–231.

- Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, 30(5), 451–459.
- Segalowitz, S. J., Santesso, D. L., Murphy, T. I., Homan, D., Chantziantoniou, D. K., & Khan, S. (2010). Retest reliability of medial frontal negativities during performance monitoring. *Psychophysiology*, 47(2), 260–270.
- Senholzi, K. B., & Ito, T. A. (2013). Structural face encoding: How task affects the N170's sensitivity to race. *Social Cognitive and Affective Neuroscience*, 8(8), 937–942.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.
- Smulders, F. T. Y., & Miller, J. O. (2012). The lateralized readiness potential. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 209–229). New York, NY: Oxford University Press.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 30, 276–315.
- Strube, M. J., & Newman, L. C. (2017). Psychometrics. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *The handbook of psychophysiology* (pp. 789–811). Cambridge: Cambridge University Press.
- Thesen, T., & Murphy, C. (2002). Reliability analysis of event-related brain potentials to olfactory stimuli. *Psychophysiology*, 39(6), 733–738.
- Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, 54(1), 123–138.
- Van Boxtel, G. J., & Böcker, K. B. (2004). Cortical measures of anticipation. *Journal of Psychophysiology*, 18(2–3), 61–76.
- Van Veen, V., & Carter, C. S. (2002). The anterior cingulate as a conflict monitor: fMRI and ERP studies. *Physiology & Behavior*, 77(4), 477–482.
- Volpert-Esmond, H. I., Merkle, E. C., Levens, M. P., Ito, T. A., & Bartholow, B. D. (2017). Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology*, doi:10.1111/psyp.13044
- Von Gunten, C., Bartholow, B. D., & Volpert, H. I. (2016). Perceiving persons: A social cognitive neuroscience perspective. In E. Harmon-Jones & M. Inzlicht (Eds.), *Social neuroscience: Biological approaches to social psychology* (pp. 10–33). New York, NY: Psychology Press.
- Walhovd, K. B., & Fjell, A. M. (2002). One-year test-retest reliability of auditory ERPs in young and old adults. *International Journal of Psychophysiology*, 46(1), 29–40.
- Weinberg, A., & Hajcak, G. (2011). Longer-term test-retest reliability of error-related brain activity. *Psychophysiology*, 48, 1420–1425.
- Weinberg, A., Riesel, A., & Hajcak, G. (2012). Integrating multiple perspectives on error-related brain activity: The ERN as a neural indicator of trait defensive reactivity. *Motivation and Emotion*, 36(1), 84–100.
- Williams, L. M., Simms, E., Clark, C. R., Paul, R. H., Rowe, D., & Gordon, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: “neuromarker”. *The International Journal of Neuroscience*, 115(12), 1605–1630.
- Woodworth, R. S. (1938). *Experimental psychology*. New York, NY: Holt.
- Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, 111, 931–959.

6

USING DAILY DIARY METHODS TO INFORM AND ENRICH SOCIAL PSYCHOLOGICAL RESEARCH

Marcella H. Boynton and Ross E. O'Hara

The foundation of the discipline that we know today as social psychology was, like so many advancements in science, born of war. The field began out of the need to change people's perceptions and behaviors in support of the war effort during World War II, and then, post war, to understand how seemingly decent people could participate in the atrocities of the Holocaust (Ross, Lepper, & Ward, 2010). Since that time social psychology has evolved into a field keenly interested in how individuals respond to one another and their social environments. What began as an area of research focused on issues such as persuasion (Merton, Fiske, & Curtis, 1946), conformity (Asch, 1956), and obedience (Milgram, 1963) has blossomed into a discipline that spans a wide range of topics, including attitudes, prosocial behaviors, personality, interpersonal relationships, self-regulation, discrimination, risk perceptions, and dual-process models of behavior, to name just a few (Fiske, Gilbert, & Lindzey, 2010). This extensive body of work has informed the efforts of researchers and practitioners working in mental health (Tennen, Gillig, Boynton, & O'Hara, 2015), public health (Klein, Shepperd, Suls, Rothman, & Croyle, 2015), medicine (Meuli, 1983; Suls & Wallston, 2008), education (Farrington et al., 2012), and economics (Kahneman, 2003).

Along with the expansion of the topical reach of social psychology, so too has there been an explosion of research methodologies and analytic techniques to accommodate the wide-ranging pursuits of this work. Early social psychological research frequently employed observational and field studies. Over time social psychologists conducted an increasing amount of their work using lab experiments, allowing them to test hypotheses in a highly controlled and systematic fashion (Wilson, Aronson, & Carlsmith, 2010). Such lab studies still figure prominently in the work of social psychologists, and there is much to recommend this approach. Lab experiments allow for the simulation of scenarios that may happen

infrequently in day-to-day life, such as job interviews, interactions with people of other races, creative brainstorming sessions, and experiences with aggression/conflict. And under most circumstances, lab experiments can be run efficiently and inexpensively, enabling researchers to test many implications of a theory in a short span of time. Perhaps most importantly, experimental methods permit researchers to ascribe causal explanations to complex psychological processes.

Although scientifically rigorous and inventive, the benefits of these studies come at a cost. Laboratory social science experiments have been criticized as artificial and overly reliant on small convenience samples (Glasgow et al., 2006; Arnett, 2008). If using an entirely between-person experimental design (i.e., no repeated measures), these studies also require a larger number of participants to achieve sufficient statistical power for detecting an effect of the experimental manipulation. Although such lab experiments have done much to further social psychological theory, there is now an increasing movement to revive the earlier approach of examining social phenomena in the “real world” and beyond a single moment in time. Indeed, it is only when social psychologists test their theories in the context of people’s daily lives that they can develop a fully accurate picture of human cognition and behavior.

Intensive Repeated Measures Designs

Given these limitations, some social science researchers have turned to alternative methods for collecting data on people’s day-to-day experiences. One approach that has yielded numerous insights into psychosocial processes and behavior is the *intensive repeated measures* (IRM) design. In its simplest form, an IRM design is a longitudinal or repeated measures study that collects data not at the yearly or monthly level, but at the daily, hourly, or moment-to-moment level. IRM techniques have been variously called naturalistic sampling, experience sampling, event-sampling, thought-sampling, diary methods, or ecological momentary assessment (Conner & Bliss-Moreau, 2006). Because some IRM designs are employed in a laboratory setting, as is common in neuroscience studies using fMRI or EEG, some writers have even used the more specific term “intensive repeated measures in naturalistic settings” (Moskowitz, Russell, Sadikaj, & Sutton, 2009). Although certain of these terms signify nuanced differences in methodologies, some of which we will describe in this chapter, they are often used interchangeably and it is worth recognizing them as close cousins. In this chapter we use the term IRM to refer to intensive repeated measures studies conducted in non-laboratory settings. This chapter will primarily focus on the IRM method of daily diary research, where respondents answer a brief set of questions each day. Such studies typically serve as passive measures of people’s daily cognitions and behaviors. Although the language used in this chapter will largely reference daily diary methods, it is important to note that many of the principles and practices that will be discussed equally apply to a wide range of IRM designs.

Daily Diary Methods in Practice: Testing Ego Depletion Theory

Many theories in social psychology make assumptions about changes in a person's behavior occurring as a result of the environments that person enters, with whom they interact, and the thoughts, feelings, and moods that they experience. These momentary deviations from a person's baseline cognitions and behaviors are often referred to in the daily diary context as *within-person* or *daily level* effects—that is, changes at the repeated measures level such as self-reported mood at the end of a day, daily alcohol use, etc. A person's more stable characteristics and their typical way of thinking and behaving are referred to as *between-person* or *person-level* variables; examples include gender, responses on a baseline survey, and mental health status.

Although laboratory studies help establish the causal mechanisms that drive behavior, these observations frequently examine phenomena at the between-person level and, as such, do little to inform our knowledge about fluctuations in cognitions and behavior over time. Diary studies, on the other hand, offer social psychologists a tool for assessing both stable and ephemeral behaviors, emotions, and cognitions. Specifically, this approach allows for statistical tests that simultaneously examine within-and between-person variables as well as potential interactions between these factors. To be clear, our contention is not that daily diary and other IRM designs are superior to laboratory experiments. Rather, daily diary research is informed and enriched by theoretical insights tested in the lab, and, in turn, daily diary research is able to put to the test the real world relevance of social psychological theories of behavior.

Let's turn to one example—self-control—to explore the question of how real-world research allows us to understand the accuracy and relevance of social psychological theory generated in a laboratory environment. Self-control came to the forefront of social psychology with Walter Mischel's observations on children's general capacity to self-regulate their impulses, often in the form of delay of gratification (for a review, see Mischel, Shoda, & Rodriguez, 1989). Typically, these studies involved a test examining how long children could resist consuming a treat placed before them when promised even more or better treats later if they succeeded in self-restraint (Mischel & Ebbesen, 1970). Not only did children who resisted temptation for longer show higher intelligence, greater achievement striving, and more social responsibility at that age (Mischel et al., 1989), but as adolescents, they scored higher on the Scholastic Aptitude Test (SAT) and demonstrated more planfulness and stronger coping skills (Shoda, Mischel, & Peake, 1990).

Mischel largely conceptualized the ability of children to delay gratification (and for adults to engage in delay discounting) as a fairly stable trait (Mischel et al., 1989). However, Baumeister and his colleagues have conducted a series of experiments exploring whether there are also situational or social components that

influence people's capacity for self-control. These researchers put forth the ego depletion or muscle theory of self-control, which, in short, argues that self-control is a limited resource that is expended when trying to control one's thoughts or actions (Baumeister, Vohs, & Tice, 2007; Muraven & Baumeister, 2000). This idea implies that regeneration of self-control takes time, and during that time we are highly susceptible to failures of self-control. Although numerous lab studies have replicated the basic premise of ego depletion (Hagger, Wood, Stiff, & Chatzisarantis, 2010), this theory has come under scrutiny of late, with both a meta-analysis (Carter, Koffler, Forster, & McCullough, 2015) and a registered replication project (Hagger et al., 2016) failing to find convincing support for ego depletion. The importance of ego depletion theory to understanding impulsive behavior as well as the mixed lab findings calls for research testing this theory in a real world context.

There are a number of scenarios in the real world where loss of self-control could presumably be in part due to earlier self-regulatory demands. For example, young adults have many opportunities and incentives to drink alcohol, yet often have reason to refrain from doing so, such as needing to drive, having to go to work or class the next day, trying to lose weight, having a fear of legal sanctions, and many others. Muraven and colleagues identified this issue as an ideal way to explore the relevance of ego depletion in a natural setting. Muraven devised a daily measures study (specifically, an *ecological momentary assessment*, or EMA, study) to determine whether a contextually induced lack of self-control in the real world could increase young adults' alcohol consumption, an effect that had been persuasively demonstrated by multiple controlled laboratory experiments (Christiansen, Cole, & Field, 2012; Otten et al., 2013; Muraven, Collins, & Nienhaus, 2002). If ego depletion theory indeed offers insights into human behavior, then it should be possible to witness depletion and regeneration of self-control resources in naturalistic settings. The research question that they hoped to answer was this: If young people are forced to self-regulate during the day, will they have any resources left to resist the urge to drink that evening?

In this study, adolescents used a handheld digital device, a Palm Pilot personal organizer with calendar, to provide information on their behaviors, thoughts, and moods across a three-week span (Muraven, Collins, Shiffman, & Paty, 2005). Study participants were asked to complete a survey upon waking each day, at 6:30 p.m. each evening, at the beginning and end of any drinking episode, and whenever the device randomly prompted them to respond. In support of ego depletion theory, individuals who experienced more self-control challenges than usual during the day were more likely to drink more than they intended to that evening. Additionally, those who reported stronger intentions to limit their drinking were the most susceptible to overindulging that evening when they were left depleted by greater than typical self-control challenges earlier in the day. Although by no means conclusive, conducting a daily measures study allowed Muraven and colleagues to gather real-world evidence in support of ego depletion theory. Specifically, they

were able to demonstrate that, at least in the context of young adult alcohol use, days exerting greater demands on regulatory capacity were associated with a higher likelihood of self-regulatory failure later that evening.

Daily Diary Studies

Muraven and colleagues' study is an example of diary measures research that adroitly leverages digital technology and multiple IRM techniques to explore dynamic daily processes, and it remains a relevant example today. The study methodology is an extension of what is commonly called a *daily diary study*, wherein a respondent is asked to answer a brief set of survey items each day for a certain number of days. Before the digital era, the administration of daily diary measures was far less controlled than in the Muraven study. Respondents in the early years of diary research were given a set of paper surveys, with one to be completed each day and returned periodically during the study or at completion of the study (e.g., Bolger, Zuckerman, & Kessler, 2000; Swim, Hyers, Cohen, & Ferguson, 2001). This format was fraught with methodological issues, one of the biggest being that timing of survey administration could not be controlled (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002). The most extreme form of this issue was when some respondents would complete multiple paper surveys on the same day to compensate for missed surveys. The introduction of the digital personal organizer and calendar in the late 1990s allowed for daily diary surveys to be administered digitally. Muraven and colleagues (2005) relied on a Hewlett-Packard Palm Pilot, which was a commonly used option at that time (Barrett & Barrett, 2001; Green, Rafaeli, Bolger, Shrout, & Reis, 2006). Roughly the size of a large cellular phone, the device could present text and images, randomly order the presentation of survey items, restrict the times respondents could complete their daily surveys, time-stamp survey responses, and prompt respondents, either randomly or at preset times, to respond to survey requests (Green et al., 2006)—such design elements are now frequently used in daily diary research. The major drawback of the Palm Pilot approach was that it required participants to charge and return the devices in order for the data to be maintained and downloaded.

The increasing ubiquity of mobile phones in the early 2000s made it possible for daily diary researchers to largely transition to phone-based daily survey data collection. Automated telephone system technology, often referred to as an interactive voice response (IVR) system, has been used to administer daily diary measures using pre-recorded questions that individuals can answer using their telephone keypad (e.g., Perrine, Mundt, Searles, & Lester, 1995). A main advantage of using an IVR system is that it does not require a respondent be able to read or to possess a dedicated mobile or landline phone, making this an optimal approach for daily diary survey studies with certain populations. In many contexts, however, having respondents enter their daily survey responses into their mobile phone in a manner analogous to using a Palm Pilot has proven to be an optimal method

of collecting daily survey data. This approach has all of the advantages of using a Palm Pilot with none of its disadvantages. Most people already possess their own device (recent estimates put mobile phone ownership among U.S. adults at greater than 95% and smartphone ownership at 77%; Pew Research Center, 2017) and data are transmitted to the researcher as soon as they are sent from the phone. Survey measures can be administered by text message or, in the case of an internet-capable mobile phone (i.e., smartphone), administered online or via a cell phone application (app; Runyan et al., 2013). Daily diary studies can also be administered online via computer (e.g., Boynton & Richman, 2014; Pond et al., 2012), although this method can be burdensome to participants who do not regularly use a computer. Online surveys administered by programs such as Qualtrics can be dually optimized for both computer and smartphone display (Vannette, 2015), making it a highly flexible approach for obtaining survey data.

Advantages of Daily Diary Studies

Naturalistic

Daily diary studies collect data in naturalistic settings, that is, wherever it is that people go in their day-to-day lives (Moskowitz et al., 2009). Although some lab researchers have gone to commendable lengths to improve external validity by transforming their laboratories into more realistic settings like living rooms (Koordeman, Anschutz, van Baaren, & Engels, 2010) or cocktail lounges (Collins, Parks, & Marlatt, 1985), the staid environment of a university building will never completely replicate the complexity of real life. As we discuss later in the chapter, wearable technology is further enhancing the naturalism of daily diary methods by allowing for passive data collection, including physiological data, which does not disconnect people from their environment in order to answer survey questions.

Real-Time

Measurements in daily diary designs are collected close to real-time, anywhere from hours to minutes after a thought or behavior occurs. These techniques are ideal for establishing temporal order, although not causality, between two events, whether those are behaviors, cognitive states, emotions, social experiences, or some combination thereof. The diary approach is assumed to collect data largely free of the fallibilities of human memory, such as mood-congruent recall, cognitive schemas, heuristics, biases, and simple forgetting (Eich, 1995; Gigerenzer & Gaissmaier, 2011; Stangor & McMillan, 1992). In the case of wearable technology, data can be recorded nearly constantly without human intervention for as long as memory and battery resources last. Daily diary studies capture people's experiences as they happen and provide researchers with a more accurate picture than

retrospective reports of what occurred in someone's daily life and how they felt about it at the time.

It is important to note, however, that "in real-time" does not necessarily equate with a "gold standard" for self-report (Conner & Bliss-Moreau, 2006). Depending on the research questions at hand, daily measures are often most informative when combined with more traditional, retrospective measures. For example, people's perspectives can change over time with self-reflection, and their later behavior may become a response to their reconstructed memory of an event (Robinson & Clore, 2002), a phenomenon that can only be studied with retrospective reports. To that end, daily diary researchers frequently administer a comprehensive baseline survey at the beginning of a daily survey study as well as a follow-up survey at the end of a study. Baseline data are frequently used as between-person variables in analyses of daily diary data and can provide important insights above and beyond those generated from daily survey data (e.g., O'Hara et al., 2014).

Rich Data Sets

Daily diary designs accrue complex datasets that can be used to generate nuanced models of behavior (Conner & Bliss-Moreau, 2006). Data analysis will be discussed in more detail later in the chapter, but to explore the potential of daily diary data, let us examine the example of a 30-day diary study to examine marital satisfaction among newlywed couples (Gadassi et al., 2016). The researchers provided smartphones to 68 spouses so they could report daily on their marital satisfaction, sexual satisfaction, and their perception of how much their spouse cares about them (i.e., perceived partner responsiveness). By administering daily measures assessing these three constructs over the course of a month, the researchers were able to show that spouses' general level of sexual satisfaction (between-person effect), as well as their sexual satisfaction on a given day (within-person effect), predicted daily levels of marital satisfaction. Additionally, the researchers found perceived partner responsiveness mediated these associations. The dyadic nature of this study further allowed for examination of *partner effects*, in this case whether a spouse's sexual satisfaction influenced the other spouse's marital satisfaction—it did not. This example demonstrates that even fairly simple daily diary designs can explore many theoretically informed questions within a single study.

Limitations of Daily Diary Methods

As with all social psychological research methods, there are drawbacks to using daily diary designs. In the case of daily measures studies that rely on self-report, people may answer carelessly, fraudulently, or in a socially desirable manner. Moreover, even delays of mere hours, though preferable to retrospective reports with even longer lags, may introduce memory biases and heuristics into participants' responding. And, like any survey design, the findings from daily diary studies are

correlational (unless these methods are incorporated into an experimental design, which is currently uncommon). Even with sound temporal ordering, third-variable explanations can still be at play. In Gadassi et al.'s study of newlyweds, for example, the association between daily sexual and marital satisfaction could have been entirely due to the influence daily mood, wherein on days when people were happier, their more positive mood led to more positive sexual interactions as well as higher levels of perceived marital satisfaction. This possibility is one reason why, as alluded to earlier, daily diary studies may be best suited as an extension of or prelude to experimental studies that can pin down causal mechanisms.

Another concern related to daily diary designs is the potential of undue burden placed on participants. Asking people to answer survey questions perhaps one or more times per day over the course of several weeks can create fatigue and even resentment. Although skip logic can be used to shorten the time it takes to complete a survey, researchers must be careful to structure the design of the survey such that participants do not learn the shortest path through a survey and answer dishonestly in order to finish faster. For this reason, it is often advisable to focus on creating a very short survey that is consistently the same length rather than allowing survey length to vary as a function of participant responses. Incentives may help maintain motivation to complete surveys, but doing so with a large sample may go beyond one's budget. Daily surveys, therefore, must be brief, making these designs poorly suited for measuring a wide array of variables or for using multi-item scale measures.

Finally, it is possible that daily diary designs could have a minor influence on constructs that are repeatedly assessed, primarily by engendering a greater awareness of certain cognitions and behavior (Tennen & Affleck, 1996; Tennen, Affleck, Armeli, & Carney, 2000). Caution may be especially important for researchers interested in sensitive or socially undesirable behaviors that participants would be reticent to report. Although there has been little evidence in daily process studies of temporal changes in within-person associations, which would suggest possible measurement reactivity (e.g., Armeli, Todd, & Mohr, 2005), researchers may wish to implement strategies to help keep participants blind to their hypotheses, such as incorporating items of a secondary research interest that help diffuse the focus of the measures. Such an approach, however, would come at the expense of replacing other items of primary interest or increasing participant burden, so like all methodological decisions the gains must be weighed against the costs.

Methodological Considerations of Daily Diary Research

As with many types of complex research methods, daily diary research requires thoughtful planning and implementation that takes into account a host of potential pitfalls. Major considerations include issues of participant recruitment, study duration, measures development, measures timing, non-response, incentive structures, programming, and database management.

Participant Recruitment

A key decision when recruiting participants is sample size. One of the greatest strengths of daily diary studies is that the repeated measures greatly boost the statistical power of the sample as compared to a standard between-subjects design. The level of increased power is related to a statistic called the intraclass correlation (ICC). In a random effects model this statistic specifies the proportion of the variance in the repeated measures outcome variable that is explained by the differences between people (Snijders & Bosker, 2012). Consider a daily diary study wherein undergraduates report their daily aggressive tendencies over the course of 25 days (Pond et al., 2012). The ICC for this daily outcome measure is found to be .68, indicating that 68% of the variability in aggressive tendencies is due to between-subjects factors and 32% is due to within-person factors. Such a large ICC is fairly common in daily diary research, as person-level factors are often as important in predicting daily behaviors as are daily factors. In other instances, such as when people are nested within organizations, the ICC tends to be much smaller, resulting in relatively greater statistical power. Power calculations must take the ICC statistic into account either by incorporating it as part of a calculation (Eldridge, Ashby, & Kerry, 2006; Rutterford, Copas, & Eldridge, 2015) or as part of a simulation study (Hoyle & Gottfredson, 2015).

Another key decision, which should be informed by the theory being tested and the value of external validity, is deciding on a population and sampling approach. A great deal of daily diary research has been conducted with non-probability based convenience samples, which are often comprised of college students. This research has yielded a great deal of knowledge about daily diary processes in general, thus allowing researchers to refine their methodological and data analytic techniques, and about college student risk behaviors and mental health, specifically. However, reliance on such samples has several drawbacks. First, consistently conducting research with samples that are predominantly young, white, educated, and technologically savvy does not always provide insights on how to conduct daily diary research with other, more complex populations such as clinical or older adult populations. Second, many of the social psychological phenomena currently studied through daily diary research are presumed to be reflective of a wide range of populations; however, the reality is that a reliance on non-representative samples limits our ability to generalize the findings of our research.

There are increasingly more options available to daily diary researchers to broaden the representativeness of their participant pools. Some daily diary researchers have increased the diversity of their samples by employing online convenience or opt-in research samples (e.g., Boynton & Richman, 2014). There are also multiple companies and organizations offering paid access to existing research panels, which are large groups of individuals available to participate in research. There are nationally representative panels, such as the University of Chicago's NORC AmeriSpeak® panel and the GfK KnowledgePanel®. There are

also specialized panels that offer access to specific groups of individuals, such as the SmartPoint Research® panel of healthcare professionals. The implementation of daily diary studies using online participant pools and research panels is an approach that is currently in its early stages; however, where resources allow, it is an attractive option for increasing the diversity and representativeness of a sample. With the use of probability-based sampling becoming more feasible and common, social psychology researchers should consider expanding their expertise with respect to survey sampling techniques and weighting of survey data.

Study Duration

There is a wide range of daily diary study durations. Some have been as short as a week (e.g., Conner, Fitter, & Fletcher, 1999) whereas others have been as long as four months (Johannes et al., 1995). There is also variability in terms of the continuity of data collection. Often daily designs, as the name implies, collect data every day, but researchers interested in studying participants over a longer period of time without creating undue burden have had participants complete surveys in alternating chunks of time, such as completing diary surveys on alternating weeks (e.g., Carson, Carson, Olsen, Sanders, & Porter, 2017; Riordan, Conner, Flett, & Scarf, 2015) or on random days over the course of multiple weeks (e.g., Claessens, van Eerde, Rutte, & Roe, 2009). Another approach called a *measurement-burst design* combines longitudinal and daily diary methods by having the same participants complete a daily diary study multiple times over the course of many years (e.g., Cullum, O'Grady, Armeli, & Tennen, 2012; Patrick, Maggs, & Lefkowitz, 2015). This approach keeps participant burden low, while also allowing researchers to examine trajectories of change at both micro and macro levels.

A commonly used period of time to collect daily surveys is 30 days. As a general principle, daily diary studies of longer duration have the advantage of statistical estimates being more precise and less affected by missing data. As detailed later in this chapter, the disaggregation of between-person versus within-person effects, as well as the estimation of random effects, are more meaningful in longer studies. However, longer studies have the disadvantage of being more burdensome to participants and more costly to implement. Selection of a study time period must balance these multiple theoretical and logistical considerations. When the primary outcome of interest occurs on a relatively infrequent basis and the added resources required for a longer data collection period are available, then a longer study period may be appropriate. Choice of study duration is usually determined based on the goals of the study, the conventions of the research area, and the resources of the research team. Pilot studies can generally afford to be relatively short whereas daily diary studies looking to examine questions that require complex and robust data benefit from being longer in duration.

Measures Development

Daily diary approaches, by necessity, must use a small number of briefly worded measures. Many social psychological measures, however, are multi-item scales using long sentences with complex syntax. Quite a few popular social psychological scales are at least 10 items (e.g., Gray-Little, Williams, & Hancock, 1997; Gibbons & Buunk, 1999), with a number of others using 20 (e.g., Fenigstein, Scheier, & Buss, 1975), 30 (e.g., Buss & Perry, 1992), or even 40 or more items (e.g., Costa & McCrae, 1992) to measure a single construct. Many of these measures use Likert-type response scales that are a minimum of 5 points and can be upward of 10 points. Use of such measures is obviously not feasible in the daily diary context. Thus, daily diary researchers themselves must often craft brief and psychometrically valid measures to fit within the bounds of this research design. This can be a challenging task; however, there are both qualitative and quantitative methods available for developing such scales.

Cognitive interviewing is a technique where a researcher conducts individual semi-structured interviews with a small sample of people (often about 9), asking questions assessing understanding and interpretation of a set of measures (Willis, 2004). Results from these studies allow researchers to identify items that do not convey their intended meaning in a reliable or clear manner as well as wording schemes and response options that are both brief and psychometrically valid. Another approach is to administer an original scale as well as analogous brief items to an appropriate convenience sample. Using the resulting data, a daily diary researcher can quantitatively validate a shorter version of the original scale. As a first step, the researcher can identify one or more brief items that correlate highly with associated parent items or factors and that exhibit low to moderate correlations with other relevant constructs, thereby establishing convergent validity and discriminant validity, respectively. Exploratory and confirmatory factor analyses can be used to examine whether brief measures load onto the same factors as their original parent items (DeVellis, 2016), and whether such models exhibit good fit. Item response theory (IRT), a method originally developed by researchers from the educational testing field, permits identification of the item or items that provide the maximum information for the underlying latent construct of interest (Embretson & Reise, 2000; Zanon, Hutz, Yoo, & Hambleton, 2016). It is usually not possible, or even necessary, to implement all of the methods described above, but the use of one or more of these tools can do much to bolster a daily diary researcher's argument that their brief measures are reliable and valid.

Measures Timing

As demonstrated by the Muraven et al. study of alcohol use, a survey completion schedule can be structured in a number of manners. A common approach in daily diary research is to set a survey completion "time window" where respondents are

able to complete a survey once per day during a certain period of time, say any-time between 6 p.m. and 10 p.m. As recall of even daily thoughts and behaviors can be imprecise, researchers using this approach often ask about behaviors for two separate time periods. In the 6 p.m. to 10 p.m. example, respondents could be asked to report on behaviors between 6 a.m. and 6 p.m. for that day as well as from 6 p.m. the day before to 6 a.m. of that day. This approach enhances recall accuracy and decreases the possibility of double counting behaviors. A signal-contingent design uses a digital device to repeatedly prompt respondents to answer surveys over the course a day, at either regularly set or randomly chosen time windows (Mehl & Conner, 2012). As discussed in greater detail later in the chapter, measurement timing can also be event contingent, meaning that respondents initiate survey completion when a certain event or behavior occurs.

A key consideration in selecting the timing and type of measures is how fleeting and/or frequent cognitions and behaviors of interest are likely to be. For example, EMA-type approaches have been used extensively to study bulimia, as binge-purge episodes have been posited to be caused by transient episodes of negative affect that can occur multiple times over the course of the day (Haedt-Matt & Keel, 2011). By assessing affect and binge-purge behaviors multiple times per day, researchers can directly examine whether periods of the day where negative affect is higher than average are more likely to be followed by a binge-purge episode. It is also possible to examine whether the period subsequent to the binge-purge episode is associated with a lower than average level of negative affect. In the end, the tool must be fit for purpose. A daily diary study is well-suited for studying phenomena that are believed to generally vary at the daily level (e.g., alcohol consumption in non-clinical samples). This method is not as suitable for studying behaviors with little daily variability (e.g., number of cigarettes smoked per day by regular smokers) or experiences that wax and wane multiple times over the course of a day.

Non-Response and Incentive Structures

A perennial challenge to daily diary researchers is the issue of non-response. Although modern analytic methods such as mixed modeling are able to include data from individuals who do not complete all of their daily measures, low response rates in a daily diary study can nevertheless detrimentally impact results in multiple ways. Non-response decreases the representativeness of findings, as the responses of the especially diligent and intrinsically motivated will predominate the results. Non-response also hampers a researcher's ability to disentangle within-person versus between-person effects, as the low number of repeated measures decreases the distinguishability of these two variables. Finally, because the optimal analytic methods for daily diary data use an iterative computational approach, low response rates can lead to model non-convergence.

Non-response can be the result of multiple design choices that frustrate participants or reward meaningless responding. Surveys should have a logical flow that

preferably starts with items most prone to bias introduced from earlier responses. The flow should also take into account the priorities of a study; if there are certain measures that are of critical importance to achieving the aims of the research, then a researcher should consider presenting those toward the beginning of the daily survey. Measures assessing sensitive issues or behaviors may be more likely to be answered when placed toward the middle or end of a survey as respondents are “eased in” by the less sensitive preceding measures. As previously noted, surveys ideally should be structured in such a way that there are an equal number of items in the survey, no matter the skip patterns. Additionally, the act of skipping a measure should be as effortful for the respondent as entering a valid response, such as requiring a refusal for each question rather than allowing a respondent to skip through a set of items by clicking a single button. The user interface should be as appealing and user-friendly as possible—for example, when designing a survey to be completed using a smartphone, make selecting a response as easy as touching one button and not as hard as opening a screen’s keyboard and entering a number. This may seem like a small difference, but even a tiny amount of increased effort for multiple measures over many days can quickly become obnoxious to a respondent. Sending at least one daily reminder to complete the day’s survey also increases response rates. Perhaps most critically when making decisions about survey design, researchers should strive to keep the daily surveys as short as possible. Every question added to a daily survey decreases the perceived value of the incentives associated with participation.

As one might expect, participant non-response is substantially lower when an appealing incentive payment structure is used. The incentive amount must be sufficiently high for people to feel that their time investment is well spent, but not so high as to be coercive. What constitutes a coercive payment is highly dependent on context and is therefore fairly subjective. Besides amount, the nature of the incentive must be considered—cash, gift cards, and items appealing to the sample (e.g., earbuds for a young adult sample) are all potentially suitable options. Norms within the field as well as institutional review board guidelines typically dictate what constitutes appropriate incentive amounts and types. The remuneration system should be structured to optimize completion of multiple surveys. Most commonly this involves creating a “bonus” system where participants receive one or more extra payments for completing a certain number of consecutive or cumulative daily surveys.

Survey Programming

A challenging component of daily diary research that is often unanticipated by novice daily diary researchers is the programming of the daily surveys. Because daily diary studies are now so reliant on technology for survey administration, some amount of programming is generally required. Over the years the level of programming expertise required has diminished; however, the need for features

such as skip patterns, reminders, randomization, experimental elements, multiple language options, etc. all add to the programming burden. There are multiple IVR systems, computer-assisted telephone interviewing (CATI) software programs, and phone apps designed specifically for daily diary research. No matter what system is chosen, expect to spend at least double the amount of time anticipated for programming and debugging your survey.

Piloting the survey methodology and programming is of utmost importance. Preferably, researchers should first conduct an internal pilot recruiting friends and colleagues as mock participants charged with the task of trying to identify survey item flaws or programing errors. These respondents should answer items in seemingly non-logical and inconsistent ways. They should also try to discover patterns of responding that decrease survey completion time or lead to errors in survey administration or the dataset. After fixing the issues identified from the internal pilot, researchers ideally should conduct an external pilot with a small set of individuals recruited from or similar to the targeted population. Implementing a daily diary study without at least an internal pilot will, at a minimum, lead to loss of data and, at its worst, prove catastrophic to a study.

Data Management

An oft-neglected aspect of daily diary research is the thoughtful design and implementation of the database management system. Daily diary studies generate multiple waves of data, with each wave typically contained in an individual data file. These files must all be accurately merged and then restructured for analytic purposes. It is advisable to devise a system that has a minimum of two means of linking the daily survey data. Possible options include the use of unique survey links that internally identify each participant or a short identification number that a participant must enter for each survey. Such system redundancies help to ensure that no data are lost because there was a failure of one data linkage method.

Another data management issue that has profound implications for the results of your study is the rubric chosen for excluding participants based on their amount of missing data. Individuals who complete only one daily survey will, by default, be excluded from inferential statistical models because such models require at least two consecutive surveys per person. And, as previously mentioned, including data from only those completing all surveys leads to large amounts of lost data and biases study results. Practices vary within the literature, with some studies including only those participants with the majority of daily surveys completed (Roth et al., 2014), those with half of their surveys completed (O'Hara et al., 2014), and those with a minority of surveys completed (Boynton & Richman, 2014). We are not aware of any currently published simulation or similar methodological studies that provide guidance on this issue. Hence, we suggest that researchers find a middle ground that accounts for the difficulty of obtaining completed surveys as well as study context, duration, and analytic goals.

Analyzing Daily Diary Data

After all of the hard work to design, implement, and complete a daily diary research project comes the data analysis. As with any research study, choices about how to analyze the data should be made *a priori* in a way that both informs, and is informed by, the research questions and design. An important consideration here is that diary data violate a key assumption of a generalized linear modeling approach—the assumption of data independence. This violation occurs because outcome data are clustered in a meaningful way and are therefore non-independent. These types of data are fairly common within education research, for example, where students are clustered within classrooms, schools, and districts (Raudenbush & Bryk, 1988). In these cases, students' responses are non-independent because they are undergoing similar experiences as other students in their "cluster," such as having the same teacher or attending the same school. Because diary methods collect multiple measures from the same people over time, these data are similarly non-independent, with each person acting as their own "cluster." For example, a person's level of self-esteem today is likely related to their level of self-esteem both yesterday and tomorrow.

Not accounting for data clustering greatly increases the likelihood of a Type 1 error (Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014). Some researchers have dealt with the issue of non-independence by simply collapsing their repeated measures outcome into a single measure by averaging or otherwise collapsing all of the daily responses into a composite score (Walker, Garber, Smith, Van Slyke, & Claar, 2001). We strenuously advise against this approach as it negates almost all of the advantages of conducting a daily diary study in the first place. In other words, this approach does not allow for the examination of day-to-day variation in your outcomes of interest. Fortunately, specialized data analytic techniques specifically designed for analyzing non-independent data are available. These approaches not only appropriately adjust the *p* value to reflect the true likelihood that your analytic results are simply due to chance, but can leverage the non-independence of the data to explore more nuanced questions of human behavior.

Multivariate Analysis of Variance

One of the most basic analytic approaches for analyzing daily diary data is the multivariate analysis of variance (MANOVA) and the multivariate analysis of covariance (MANCOVA) models, which are extensions of the analysis of variance (ANOVA) and analysis of covariance (ANCOVA) models, respectively. In the early years of daily diary and other IRM studies MANOVA/MANCOVA were one of the few readily available options for analyzing repeated measures data (e.g., Morin & Gramling, 1989). Given that social psychologists frequently use analysis of variance-type models to analyze their experimental data this method is likely appealing; however, the approach is limited in that it only allows for

basic comparisons between different groups or, in the case of daily diary research, comparisons between people. In the 1990s and early 2000s daily diary researchers slowly transitioned away from this analytic method to something called a mixed modeling framework (Bolger & Schilling, 1991; Singer, 1998). Although the MANOVA/MANCOVA approach is still sometimes used (Tasca et al., 2009), daily diary researchers now typically employ mixed modeling to analyze their data, as it is far more efficient, flexible, and nuanced.

Mixed Modeling

Mixed modeling is an iterative modeling approach that allows for flexible modeling of variance components. Some types, such as generalized estimating equation (GEE) models, treat random effects in the model, which in the case of daily diary data are unexplained person effects, as a nuisance, simply adjusting the p values to account for clustering. There are many situations where this approach is perfectly acceptable; however, another method, *multilevel modeling* (MLM; also referred to as hierarchical linear modeling), allows researchers to model both random effects and fixed effects, the latter of which are typically the parameter estimates of interest (Singer, 1998; Hox, 2010). By being able to simultaneously model both types of effects, it is possible to broaden the range of research questions that can be answered (Mehl & Conner, 2012). MLM can be applied to a range of outcome variable types, including standard normal, binary, poisson, and negative binomial, and can be conducted using a variety of statistical software packages such as *Mplus*, SPSS, STATA, SAS, and R.

In any discussion of MLM, it is important to understand that clustering can occur at multiple levels. Let us look to a study of student performance over time by Bryk and Raudenbush (1988) to elaborate on this point. The study took into account repeated measures nested within each student as well as students nested within schools. In multilevel modeling parlance, the most granular unit of data (in this case, repeated measures within each student) is referred to as Level 1 data, the next highest level of data clustering (student-level) as Level 2 data, and the next highest (school-level) as Level 3 data. The number of levels is theoretically limitless (clustering of schools within districts would be Level 4, clustering of districts within city would be Level 5, and so on); however, at some juncture the level of variance explained by higher order units becomes so small as to be effectively nil. More practically, because MLMs use an iterative computational approach, models with more than two levels are much less likely to converge on a computational solution. In the case of daily diary designs, Level 1 data are typically the repeated measures variables and Level 2 data are the person-level variables. In the case of the Gadassi et al. daily diary study of marital satisfaction referenced earlier in this chapter, Level 1 variables are the repeated daily measures, Level 2 are the person-level variables, and Level 3 are the dyadic variables. Note, though, that although

predictors in a MLM can represent all levels of clustering, MLM analyses are only appropriate when the outcome is a Level 1 variable.

A study conducted with a sample of African American college students by Tennen and colleagues serves as a good example of how MLM can be effectively leveraged to examine nuanced aspects of daily behavior. In this study researchers examined how a number of psychosocial factors and experiences predicted number of daily alcoholic drinks consumed (Kranzler et al., 2012). In one analysis (O'Hara et al., 2014) multiple drinking motives were assessed, with one of the most theoretically important being drinking to cope (i.e., respondents stating that they drank that day because they were depressed, nervous, etc.; Cooper, 1994). For days when respondents drank alcohol they were asked several items assessing how much they drank for coping purposes, a Level 1 variable. Age, gender, and past experiences of discrimination and trauma were a few of the many Level 2 variables assessed at baseline. Results showed that on days when people's coping motives were higher than normal they consumed more drinks, especially when drinking alone.

When specifying a multilevel model, variable centering is an important analytic decision (Curran & Bauer, 2011). Centering of Level 1 variables has different theoretical implications than the centering of Level 2 variables. To take the above daily coping motives analysis as an example, each person's average daily coping score across the 30 day period was first computed; this variable served as a measure of overall drinking to cope during the study (Level 2 effect). Next, each person's average coping score across the 30 day period was subtracted from his or her coping score reported for each day, a technique known as person mean centering (referred to as group mean centering when individuals are nested within larger units). This person mean centered variable served as a measure of episodic drinking to cope (Level 1 effect) because it indicated whether a person's coping motives on a particular day were higher or lower than his or her average level of drinking to cope across the 30 day study period. By simultaneously entering both the Level 1 and Level 2 daily survey drinking to cope variables into the model it was then possible to disentangle the within-person effect of drinking to cope more than usual on a given day from the between-person effect of drinking to cope in general.

In this case, both the Level 1 and Level 2 drinking to cope variables predicted the number of daily drinks consumed, particularly in non-social contexts. As this example demonstrates, person mean centering has meaningful theoretical implications and so should be used only when theory warrants (Paccagnella, 2006). Level 2 variables in the study, such as age, were grand mean centered. This simply means that the average value across all participants, which in the drinking to cope example was 20.1 years, was subtracted from each person's individual value. The resulting variable indicates how much older or younger each person was compared to the overall sample average. Results indicated that higher age was

associated with a greater number of drinks consumed in a social drinking context, but not in a non-social context. There are no theoretical implications to grand mean centering—it simply facilitates model convergence and the interpretation of model estimates.

Ecological Momentary Assessment Studies

As previously mentioned, EMA studies differ from diary studies in that they typically include measurement periods throughout the day for multiple days (Shiffman, Stone, & Hufford, 2008). These studies harken back to the Palm Pilot alcohol use study mentioned earlier, but have become even more desirable with the advent of mobile and wearable technologies. There are three main types of experience sampling made possible by EMA studies (Moskowitz et al., 2009). Time-contingent measures are taken at regular intervals, usually every so many hours apart. This approach ensures complete coverage of people's daily experiences and makes easier the temporal ordering of events for data analysis. Signal-contingent measures, however, alert participants at random times of day to complete surveys. Both time-contingent and signal-contingent approaches typically include constraints for when people are asleep or unavailable, as well as the ability for participants to impose a short-term delay in completing a set of measures if they are temporarily indisposed. Signal-contingent measurements have the advantage of covering varying periods of experience when methodological constraints prevent frequent repeated measures. The signal-contingent approach can also help maintain novelty for participants, thereby mitigating fatigue to some degree. However, events of interest may be missed when they fall within a randomly created gap in measurement. Event-contingent measures avoid this particular pitfall by being collected in response to a particular behavior. With this method, participants are asked to activate their collection device whenever a specific event occurs, like waking up, smoking a cigarette, or feeling angry. This approach is often used to assess phenomena that are particularly ephemeral or rare. These assessments have the advantage of capturing behaviors of interest right as they occur, thereby reducing the chance that these important behaviors are missed; however, they have the disadvantage of being reliant upon the participant to self-initiate the data collection protocol.

EMA has a number of advantages over daily diary designs. First, the temporal gap between behavior and measurement is smaller, thereby minimizing recall bias. Second, temporal ordering of events can be estimated more precisely, which, as we saw in the bulimia research example (Haedt-Matt & Keel, 2011), makes it possible to make stronger inferences about causality. Third, the precision of the data allow for more complex modeling of psychological processes over time, while still allowing for aggregation at higher levels of temporality (e.g., daily measures).

EMA also has its drawbacks. Because participants may be asked to interrupt their day multiple times to respond to surveys, the number of survey items must

be even fewer than in a daily design. Second, by increasing the number of response windows per day, both costs and participant burden often necessitate a reduction in the overall length of the study; most EMA studies last no more than 14 days. Finally, EMA studies may not be conducive to studying certain phenomena. For example, a one-week EMA study of sexual behavior may fail to capture any sexual behavior because the overall duration is too short. Moreover, something like sexual behavior may be unlikely to occur more than once per day, making EMA superfluous. The point to note is, as has been illustrated throughout this chapter, one's research questions should drive the choice of methodology.

Future Directions in Daily Diary Research

Daily Diary Methods as an Intervention Tool

Earlier we cautioned that poorly designed daily diary studies can, under certain circumstances, unintentionally influence participants. When the goal of a research study is to accurately describe what happens to people in natural settings and how they respond, such reactivity would be a serious flaw in the design. But many researchers, especially in applied fields like health psychology, want to change people's behavior through proactive intervention, and IRM methods such as daily self-monitoring studies are a promising technique for doing just that (for a review, see Heron & Smyth, 2010). For example, researchers at the University of Otago sent daily text messages to college students during Orientation Week to try to reduce their drinking (Riordan et al., 2015). These texts provided a short message focused on either the health-related or social consequences of excessive drinking. This intervention successfully reduced women's alcohol consumption both during Orientation Week and across the entire semester. Such studies are fairly novel and more research is needed to explore how best to use daily messages to influence behavior.

Wearables and Daily Diary Research

With the advent of wearable technologies such as pedometers and activity trackers, daily diary research focused on bodily processes is increasingly feasible and affordable to accomplish. These "wearables" offer the ability to collect data passively, thereby reducing participant burden and freeing up precious survey space for variables reliant on self-report. The most ubiquitous "wearable" is the smartphone, which has the ability to track people's location, pace, voice, surrounding environment, and activity both online and via social media. Being granted access to participants' mobile devices provides a wealth of data for researchers interested in the minutiae of people's daily lives. An increasingly common wearable are activity trackers, the most popular of which is the Fitbit. Researchers have now validated that Fitbits accurately measure, among other variables, the number of

steps people take in their daily lives (Dontje, de Groot, Lengton, van der Schans, & Krijnen, 2015), and can be used as a means of data collection and behavioral intervention in the context of physical activity and health (Cadmus-Bertram, Marcus, Patterson, Parker, & Morey, 2015).

Some wearables have the ability to augment the potential of more standard daily diary self-reports. For example, wearable heart rate monitors can not only collect data on heart rate but can also be used to trigger an electronic diary entry whenever the person's heart rate passes a certain threshold (Intille, 2007). As another example, the Electronically Activated Recorder (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001), a lightweight wearable that passively records audio in the natural environment, has been used to enhance the richness of daily experience sampling data. Not only are advances in digital recording and wireless technology increasing the capacity of such devices to record data over longer periods of time, but advances in computing capacity now also allow for near-instant parsing of these recordings for keywords that could activate electronic questionnaires (Intille, 2007). As with cell phones before them, wearable technologies continually advance the possible accessibility, reach, and precision of IRM methods.

Summary

Over 70 years of research in social psychology has greatly advanced our understanding of how people behave in response to their environment and those around them. Rigorous experimental methods have been the backbone of this progress; however, IRM methods, including daily diary studies, offer social psychologists the opportunity to test their theories in the real world. These approaches allow for the collection of rich data that can answer a great many questions about how people act in their day-to-day lives. But to leverage the power of these designs, researchers must make a number of important decisions from participant recruitment to study protocol to data analysis that have far-reaching implications for the quality and utility of their data. While we hope that this chapter has provided ample guidance to get readers started toward designing an IRM study, it is important to note that this chapter is a very broad overview of a vast and complex methodology. Even seasoned daily diary researchers must continue to educate themselves on best practices, as many aspects of daily diary research continue to evolve. There are several resources that offer a much more in depth treatment of this subject (Mehl & Conner, 2012; Nezlek, 2012), and we encourage those with a serious interest in conducting daily diary research to seek out these and other valuable sources of theoretical and practical information.

Daily diary designs are just one approach of many available to social psychologists that allows for the further elucidation of the complexities of human behavior. Our ability to collect more and different types of data continues to advance with the introduction of new and more powerful technologies. However, the value of daily diary research is greatest when informed by social psychological theory

and conducted in iterative conjunction with other types of social psychological research. Whether a daily diary study sets out to test a longstanding or controversial theory or is the impetus for a novel theory, capturing people's experiences in the places they go every day, with the people they see every day, is essential to furthering the field of social psychology.

References

- Aarts, E., Verhage, M., Veenliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, 17(4), 491–496. doi:10.1038/nn.3648
- Armeli, S., Todd, M., & Mohr, C. (2005). A daily process approach to individual differences in stress-related alcohol use. *Journal of Personality*, 73(6), 1657–1686.
- Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63(7), 602–614. doi:10.1037/0003-0066x.63.7.602
- Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1–70.
- Barrett, L. F., & Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, 19(2), 175–185.
- Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, 16(6), 351–355.
- Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality*, 59(3), 355–386.
- Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology*, 79(6), 953–961.
- Boynton, M. H., & Richman, L. S. (2014). An online daily diary study of alcohol use using Amazon's Mechanical Turk. *Drug and Alcohol Review*, 33(4), 456–461.
- Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97(1), 65–108.
- Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, 63(3), 452–459.
- Cadmus-Bertram, L. A., Marcus, B. H., Patterson, R. E., Parker, B. A., & Morey, B. L. (2015). Randomized trial of a Fitbit-based physical activity intervention for women. *American Journal of Preventive Medicine*, 49(3), 414–418.
- Carson, J. W., Carson, K. M., Olsen, M. K., Sanders, L., & Porter, L. S. (2017). Mindful Yoga for women with metastatic breast cancer: Design of a randomized controlled trial. *BMC Complementary and Alternative Medicine*, 17(1), 153. doi:10.1186/s12906-017-1672-9
- Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, 144(4), 796–815.
- Christiansen, P., Cole, J. C., & Field, M. (2012). Ego depletion increases ad-lib alcohol consumption: Investigating cognitive mediators and moderators. *Experimental and Clinical Psychopharmacology*, 20(2), 118–128.
- Claessens, B. J. C., van Eerde, W., Rutte, C. G., & Roe, R. A. (2009). Things to do today . . . : A daily diary study on task completion at work. *Applied Psychology*, 59(2), 273–295.

- Collins, R. L., Parks, G. A., & Marlatt, G. A. (1985). Social determinants of alcohol consumption: The effects of social interaction and model status on the self-administration of alcohol. *Journal of Consulting and Clinical Psychology*, 53(2), 189–200.
- Conner, M., Fitter, M., & Fletcher, W. (1999). Stress and snacking: A diary study of daily hassles and between-meal snacking. *Psychology & Health*, 14(1), 51–63.
- Conner, T., & Bliss-Moreau, E. (2006). Sampling human experience in naturalistic settings. In S. N. Hesse-Biber & P. Leavy (Eds.), *Emergent methods in social research* (pp. 109–129). Thousand Oaks, CA: Sage Publications.
- Cooper, M. L. (1994). Motivations for alcohol use among adolescents: Development and validation of a four-factor model. *Psychological Assessment*, 4, 117–128.
- Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4(1), 5–13.
- Cullum, J., O'Grady, M., Armeli, S., & Tennen, H. (2012). Change and stability in active and passive social influence dynamics during natural drinking events: A longitudinal measurement-burst study. *Journal of Social and Clinical Psychology*, 31(1), 51–80.
- Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, 62, 583–619.
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Sage Publications.
- Dontje, M. L., de Groot, M., Lengton, R. R., van der Schans, C. P., & Krijnen, W. P. (2015). Measuring steps with the Fitbit activity tracker: An inter-device reliability study. *Journal of Medical Engineering & Technology*, 39(5), 286–290.
- Eich, E. (1995). Searching for mood dependent memory. *Psychological Science*, 6(2), 67–75.
- Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5), 1292–1300. doi:10.1093/ije/dyl129
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners*. Retrieved from <http://files.eric.ed.gov/fulltext/ED542543.pdf>
- Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43(4), 522–527.
- Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). *Handbook of social psychology* (5th ed, chapter 1). Hoboken, NJ: John Wiley & Sons.
- Gadassi, R., Bar-Nahum, L. E., Newhouse, S., Anderson, R., Heiman, J. R., Rafaeli, E., & Janssen, E. (2016). Perceived partner responsiveness mediates the association between sexual and marital satisfaction: A daily diary study in newlywed couples. *Archives of Sexual Behavior*, 45, 109–120.
- Gibbons, F. X., & Buunk, B. P. (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. *Journal of Personality and Social Psychology*, 76(1), 129–142.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.
- Glasgow, R. E., Green, L. W., Klesges, L. M., Abrams, D. B., Fisher, E. B., Goldstein, M. G., . . . Orleans, C. T. (2006). External validity: We need to do more. *Annals of Behavioral Medicine*, 31(2), 105–108. doi:10.1207/s15324796abm3102_1
- Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 23(5), 443–451.

- Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, 11(1), 87–105.
- Haedt-Matt, A. A., & Keel, P. K. (2011). Revisiting the affect regulation model of binge eating: A meta-analysis of studies using ecological momentary assessment. *Psychological Bulletin*, 137(4), 660–681.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwieneberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573.
- Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, 136(4), 495–525.
- Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behavior treatments. *British Journal of Health Psychology*, 15, 1–39.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Taylor & Francis.
- Hoyle, R. H., & Gottfredson, N. C. (2015). Sample size considerations in prevention research Applications of multilevel modeling and structural equation modeling. *Prevention Science*, 16(7), 987–996. doi:10.1007/s11121-014-0489-8
- Intille, S. S. (2007). Technological innovations enabling automatic, context-sensitive ecological momentary assessment. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 308–337). New York: Oxford University Press.
- Johannes, C. B., Linet, M. S., Stewart, W. F., Celentano, D. D., Lipton, R. B., & Szklo, M. (1995). Relationship of headache to phase of the menstrual cycle among young women: A daily diary study. *Neurology*, 45(6), 1076–1082.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5), 1449–1475.
- Klein, W. M. P., Shepperd, J. A., Suls, J., Rothman, A. J., & Croyle, R. T. (2015). Realizing the promise of social psychology in improving public health. *Personality and Social Psychology Review*, 19(1), 77–92. doi:10.1177/1088868314539852
- Koordeman, R., Anschutz, D. J., van Baaren, R. B., & Engels, R. C. M. E. (2010). Effects of alcohol portrayals in movies on actual alcohol consumption: An observational experimental study. *Addiction*, 106, 547–554.
- Kranzler, H. R., Scott, D., Tennen, H., Feinn, R., Williams, C . . . Covault, J. (2012). The 5-HTTLPR polymorphism moderates the effect of stressful life events on drinking behavior in college students of African descent. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159(5), 484–490.
- Mehl, M. R., & Conner, T. (2012). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.
- Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4), 517–523.
- Merton, R. K., Fiske, M., & Curtis, A. (1946). *Mass persuasion; the social psychology of a war bond drive*. Oxford: Harper.
- Meuli, C. (1983). Social psychology and medicine. *JAMA*, 249(18), 2543–2544. doi:10.1001/jama.1983.03330420087044
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Mischel, W., & Ebbesen, E. B. (1970). Attention and delay of gratification. *Journal of Personality and Social Psychology*, 16(2), 329–337.

- Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, 244(4907), 933–938.
- Morin, C. M., & Gramling, S. E. (1989). Sleep patterns and aging: Comparison of older adults with and without insomnia complaints. *Psychology & Aging*, 4(3), 290–294.
- Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology*, 50(3), 131–140.
- Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, 126(2), 247–259.
- Muraven, M., Collins, R. L., & Nienhaus, K. (2002). Self control and alcohol restraint: An initial application of the self-control strength model. *Psychology of Addictive Behaviors*, 16(2), 113–120.
- Muraven, M., Collins, R. L., Shifman, S., & Paty, J. A. (2005). Daily fluctuations in self-control demands and alcohol intake. *Psychology of Addictive Behaviors*, 19(2), 140–147.
- Nezlek, J. B. (2012). *Diary methods for social and personality psychology*. Los Angeles, CA: Sage Publications.
- O'Hara, R. E., Boynton, M. H., Scott, D., Armeli, S., Tennen, H., Williams, C., & Covault, J. (2014). Drinking to cope among African-American college students: An assessment of episode-specific motives. *Psychology of Addictive Behaviors*, 28(3), 671–681.
- Otten, R., Cladder-Micus, M. B., Pouwels, J. L., Hennig, M., Schuurmans, A. A. T., & Hermans, R. C. J. (2013). Facing temptation in the bar: Counteracting the effects of self-control failure on young adults' ad libitum alcohol intake. *Addiction*, 109, 746–753.
- Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, 30(1), 66–85. doi:10.1177/0193841x05275649
- Patrick, M. E., Maggs, J. L., & Lefkowitz, E. S. (2015). Daily associations between drinking and sex among college students: A longitudinal measurement burst design. *Journal of Research on Adolescence*, 25(2), 377–386.
- Perrine, M. W., Mundt, J. C., Searles, J. S., & Lester, J. S. (1995). Validation of daily self-reported alcohol consumption using interactive voice response (IVR) technology. *Journal of Studies on Alcohol*, 56(5), 487–490.
- Pew Research Center. (2017, January 12). *Mobile fact sheet*. Retrieved from www.pewinternet.org/fact-sheet/mobile
- Pond, R. S., Jr., Kashdan, T. B., DeWall, C. N., Savostyanova, A., Lambert, N. M., & Fincham, F. D. (2012). Emotion differentiation moderates aggressive tendencies in angry people: A daily diary analysis. *Emotion*, 12(2), 326–337.
- Riordan, B. C., Conner, T. S., Flett, J. A. M., & Scarf, D. (2015). A brief Orientation Week ecological momentary intervention to reduce university student alcohol consumption. *Journal of Studies on Alcohol and Drugs*, 76, 525–529.
- Raudenbush, S. W., & Bryk, A. S. (1988). Chapter 10: Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education*, 15(1), 423–475. doi:10.3102/0091732X015001423
- Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6), 934–960.
- Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, pp. 51–81). Hoboken, NJ: John Wiley & Sons.
- Roth, A. M., Hensel, D. J., Fortenberry, J. D., Garfein, R. S., Gunn, J. K. L., & Wiehe, S. E. (2014). Feasibility and acceptability of cell phone diaries to measure HIV risk behavior

- among female sex workers. *AIDS and Behavior*, 18(12), 2314–2324. doi:10.1007/s10461-010014-10718-y
- Runyan, J. D., Steenbergh, T. A., Bainbridge, C., Daugherty, D. A., Oke, L., & Fry, B. N. (2013). A smartphone ecological momentary assessment/intervention “app” for collecting real-time data and promoting self-awareness. *PLoS ONE*, 8(8). doi:10.1371/journal.pone.0071325
- Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3), 1051–1067. doi:10.1093/ije/dyv113
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*, 26(6), 978–986.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4), 323–355. doi:10.3102/10769986023004323
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA, London, New Delhi, Singapore, and Washington, DC: Sage Publications.
- Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, 111(1), 42–61.
- Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *BMJ*, 324(7347), 1193–1194.
- Suls, J., & Wallston, K. A. (2008). *Social psychological foundations of health and illness*. Malden, MA: Wiley-Blackwell.
- Swim, J. K., Hyers, L. L., Cohen, L. L., & Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues*, 57(1), 31–53.
- Tasca, G. A., Illing, V., Balfour, L., Krysan, V., Demidenko, N., Nowakowski, J., & Bissada, H. (2009). Psychometric properties of self-monitoring of eating disorder urges among treatment seeking women: Ecological momentary assessment using a daily diary method. *Eating Behaviors*, 10(1), 59–61. doi:10.1016/j.eatbeh.2008.10.004
- Tennen, H., & Affleck, G. (1996). Daily processes in coping with chronic pain: Methods and analytic strategies. In M. Zeidner & N. S. Endler (Eds.), *Handbook of coping: Theory, research, applications* (pp. 151–177). Oxford: John Wiley & Sons.
- Tennen, H., Affleck, G., Armeli, S., & Carney, M. A. (2000). A daily process approach to coping: Linking theory, research, and practice. *American Psychologist*, 55(6), 626–636.
- Tennen, H., Gillig, P. M., Boynton, M. H., & O'Hara, R. E. (2015). Social psychology: Theory, research, and mental health implications. In A. Tasman, J. Kay, J. A. Lieberman, M. B. First, & M. B. Riba (Eds.), *Psychiatry* (4th ed., pp. 453–462). Hoboken, NJ: John Wiley & Sons.
- Vannette, D. (2015). Four easy ways to optimize your survey for mobile devices. *Qualtrics*. Retrieved October 3, 2017, from www.qualtrics.com/blog/4-easy-ways-to-optimize-your-survey-for-mobile-devices/
- Walker, L. S., Garber, J., Smith, C. A., Van Slyke, D. A., & Claar, R. L. (2001). The relation of daily stressors to somatic and emotional symptoms in children with and without current abdominal pain. *Journal of Consulting and Clinical Psychology*, 69(1), 85–91.

- Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.
- Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). *The art of laboratory experimentation*. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, pp. 51–81). Hoboken, NJ: John Wiley & Sons.
- Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29(1), 18. doi:10.1186/s41155-41016-40040-x

7

TEXTUAL ANALYSIS

Cindy K. Chung and James W. Pennebaker

Introduction

Language is a defining feature of human culture. Although social scientists have long agreed about the profound nature of language, they have been reticent to study it. And there is good reason: it is exceedingly difficult to record, amass, extract, and to code or decode what people are saying. Only recently have social scientists made giant strides in measuring and analyzing the words people use in their everyday lives. Twenty-first-century technological advances have given birth to computer-mediated communication (CMC), which has made our lives considerably easier. For the first time in history, we have the tools to track human interaction through the spoken and written word quickly and efficiently, and at a scale that was unimaginable even a decade ago.

CMC is providing social psychologists with new questions to answer about micro-interactions, emotional tone, group dynamics, and cultural shifts that may change in seconds or centuries. The analysis of CMC and other types of language offers a means to understand how we are influenced by the actual, implied, or imagined presence of others. We can now analyze millions of books and manuscripts dating back centuries. We can also quickly track societal changes in thinking and communication through the analysis of language in Snapchat, Facebook, Tinder, or one of hundreds of new apps that appear each year both within and across cultures.

Even the term “CMC” seems rather antiquated. Most communication now occurs over some type of digitally connected device, with hand-written letters considered a dying art, an unscheduled phone call too obtrusive, and commitments that don’t really exist unless by email or SMS confirmation. The majority of humans in developed countries own a personal smartphone for text messaging,

manage several cluttered email inboxes with thousands of unread messages, and are similarly guilty for having an ever-growing email outbox. It is customary for individuals who have never met or spoken in person to interact long term daily and digitally. For example, these might be primary work collaborators, online acquaintances explicitly looking for love, 20 hours per week video game allies, opposing lawsuit parties, devoted customer and merchant, or daily content makers and subscribers.

Communication, and by extension, social interactions, have changed. Although the medium has changed—as have the tools to record, amass, extract, code or decode, and to assess their meaning or style—the words, especially those that are most revealing of social dynamics, have largely stayed the same.

Content vs. Function Words

A helpful way to look at measuring language in social psychology is to consider two broad categories: content words and function words (see Pennebaker, 2011). Content words are made up of nouns, regular verbs, adjectives, and adverbs. Content words tell us what people are thinking about. Function words are made up articles, auxiliary verbs, conjunctions, negations, pronouns, and prepositions. Function words tell us how people are thinking and connecting with others.

Both categories of words are revealing of our thoughts, feelings, and behaviors, with function words being more reliable markers of psychological states and social dynamics across topics (Chung & Pennebaker, 2007; Mehler, 2006; Pennebaker, Mehler, & Niederhoffer, 2003), which are typically represented by content words. For example, the topic of the statement

My Facebook post had a lot of comments.

is gleaned from the words “Facebook,” “post,” and “comments.”

How someone is thinking about the topic is understood from “My,” “had,” and “a lot of.” Specifically, “My” represents self-focus: a personal share or ownership of the topic. Had the speaker used “The” instead of “My,” the Facebook post could’ve been written by anyone, including a more personally distanced way of referring to the speaker’s own Facebook post. “Had” represents past-tense focus. Together, “my” and “had” assume a shared reference between the speaker and the audience as to which of the speaker’s past Facebook posts the speaker is referencing.

“A lot of” represents some comparison to a quantity which is unknown unless the speaker and audience have a shared reference point to how many comments represent a relatively large quantity. Had the speaker used “a lot of” knowing that both the speaker and audience thinks that say, over 50 comments is a large quantity, but the speaker had in fact received 2 comments, this statement might be viewed as sarcastic or funny as opposed to a casual and intentionally accurate

statement about the large (i.e., over 50) quantity of comments. Had the speaker used “a lot of” thinking that the audience was interested in frequent reports on the number of comments received on each of the speaker’s Facebook posts, but the audience was, in fact, not interested in said reports, this statement might be viewed as annoying or gratuitously boastful.

Note that it is not very interesting to think about the meaning of each word as in the laborious example above. (Note also, that linguists may disagree.) However, a few takeaways from the exercise are that (a) speakers and listeners process function words automatically without going through the steps above. (b) Function words are inherently social, drawing on shared references between a speaker or writer and their audience to use and to understand. (c) Different psychological states and social dynamics are associated with different categories of function word use. (d) Function words make up over 50% of the words we use in our everyday speech and writing. Together, these make function words excellent observable behaviors to understand how individuals are influenced by the actual, implied, or imagined presence of others. In other words, function words are the stuff of social psychologists’ dreams.

Linguistic Inquiry and Word Count (LIWC)

Admittedly, most social psychologists don’t, in fact, dream about function words as gateways into the inner workings of social dynamics. It is understandable why this may be the case. As mentioned in the introduction, language has not always been easy to record, amass, extract, code or decode, and to assess its meaning or style. However, several innovations have made the analysis of function words much more accessible to social psychologists. The primary tool that turned widespread attention in social psychology to function words was the advent of Linguistic Inquiry and Word Count (LIWC 2001; Pennebaker, Francis, & Booth, 2001). LIWC, pronounced “Luke,” is a software made up of a processor and a dictionary. The processor counts words in the category entries listed in the dictionary, and reports on the percentage of words in each text file that represents each dictionary category.

The standard LIWC dictionary is made up of over 80 categories, including function word categories (e.g., articles, negations, pronouns, etc.), and content word categories (e.g., positive and negative emotion words, cognitive mechanisms, social words, biological words, achievement, religion, etc.). Each of the words in the standard LIWC dictionary having been judged by four judges, and agreed on by at least three of those judges, as belonging to its category. It is possible to have the processor count words in custom dictionaries: this function makes it possible for users to create their own “dictionary,” made up of words of their choosing in categories of their choosing.

LIWC is relatively easier to use than other natural language processing (NLP) techniques as LIWC is a cross-platform application, and no programming is

required. We refer the reader to the LIWC website (www.liwc.net) for a manual on its use but provide a brief overview here. For any given project, a corpus (a body of text files) is collected into a folder, where each text file in the corpus represents an observation (i.e. all the typed or transcribed words of an individual, or, a single message from an individual). Each text file should have a minimum number of words specified by the researcher. Note that LIWC reports on the percentages of words, and so a minimum cutoff around 100 words may seem reasonable for many studies, although there may be reasons to decrease this cutoff, and there are no hard and fast rules on what should be the minimum cutoff. In short, more words are associated with greater reliability.

Within LIWC, the first step is to identify the location of the text files to process. The default dictionary is referenced by default, although, as previously mentioned, it is possible to have LIWC point to a custom dictionary. Once the files are selected, LIWC automatically processes all files, counting the percentage of words in each category of the LIWC dictionary for each text file.

The LIWC output, which is a matrix of text files in rows, LIWC categories in columns, and percentages of use in each cell, is saved as an output file in TXT, CSV, or Excel format. This output file can be opened with any statistical package to conduct analyses on the relative rates of word use by different groups of text files. Accordingly, it is essential to design an empirical study with statistical tests that will answer one's research questions in advance of preparing the text files, just as any survey or observational study would format data collection to have observations in rows, measurements in columns, and values in cells. The statistical analyses to be conducted are entirely dependent on the research questions posed. Ultimately, having a large collection of words is not enough; having an appropriate understanding of experimental design and a statistical analytic strategy, just as in any empirical social psychological assessment, are required.

Several updates, including the product's commercialization, have been made to both the processor (Pennebaker, Booth, & Francis, 2007) and the standard LIWC dictionary in 2007 (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) and in 2015 (Pennebaker, Booth, Boyd, & Francis, 2015). The current 2015 processor has the ability to process text in various file formats such as .xlsx, .csv, .pdf, and others, beyond plain ASCII text files. Currently, the standard LIWC dictionary (from both earlier and the current versions) has been translated into Arabic, Chinese, Dutch, French, German, Korean, Russian, Spanish, and Turkish (see www.liwc.net), with several other languages under development. The latest standard LIWC dictionary in English includes several super categories (e.g., analytic thinking, authenticity, etc.) which are constructs derived from LIWC categories and based on past research that has used LIWC (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

Note that there are a myriad of other theoretically based computerized word counting programs, including The General Inquirer for psychological topics (Stone, Dunphy, Smith, & Ogilvie, 1966), DICTION for political texts (Hart & Carrol, 2014), and TAS/C for psychotherapeutic transcripts (Mergenthaler, 1996; Mergenthaler & Bucci, 1999). There are also an increasing number of text analysis

methods from the broader field of NLP that can be used to measure words. There are an overwhelming number of text analysis program to list. So, in order to avoid overwhelming a beginner to text analysis, we provide a table below on just a few of the popular, well-maintained, and simplest tools for which no programming is required. These tools can easily be used by social scientists to begin to incorporate some text analysis in their multi-methods toolkits.

Many of these text analytic software packages count and categorize words; they are measurement tools. However, to gain descriptive and inferential insight into what those words signal, descriptive and inferential statistics must be applied in a subsequent step, in a statistical software package.

TABLE 7.1 Examples of Text Analysis Software

<i>Tool</i>	<i>Purpose</i>	<i>Reference</i>
Linguistic Inquiry and Word Count (LIWC)	To derive a matrix of words that indicates the percentages of words in a text belonging to validated categories (grammatical, psychological, content); custom dictionaries can be uploaded to assess a corpus for specific words; comparisons in word use across and between groups can be conducted when the matrix is uploaded to a statistical software package.	www.liwc.net
Meaning Extraction Method (MEH)	To derive a matrix of terms that indicates the percentages of words in a text belonging to the most frequently referenced terms in a corpus; facilitates the Meaning Extraction Method (MEM; Chung & Pennebaker (2008); a method to inductively derive topics in a corpus) when the matrix is uploaded to a statistical software package.	http://meh.ryanb.cc
QDA Miner Provalis	To conduct frequency and percentage counts of words in a corpus that are not necessarily dictionary driven; some descriptives and basic topic modeling capabilities for descriptive purposes.	https://provalisresearch.com
WordSmith	To conduct frequency and percentage counts of words in a corpus that are not necessarily dictionary driven, but more descriptive of a text; co-occurrences of words can also be assessed.	http://lexically.net
Coh-Metrix	To assess over a hundred features of text using a variety of statistically derived indices, including cohesion, and readability.	www.cohmetrix.com

In the case of LIWC, recall that one of the requirements for the inclusion of words in the LIWC dictionary was that judges agreed that a word belonged in a category. Since function words are widely agreed upon fixed lists, these were added to the standard LIWC dictionary. It wasn't until after using LIWC for a wide variety of studies that function words were often found to be more reliable markers of psychological state than content words (Pennebaker, 2011). With word counting as a basic foundation of quantitative text analysis across many disciplines (see O'Connor, Banman, & Smith, 2011), a growing complexity of statistical techniques have been applied to word counts including the output of LIWC's categories to derive new insights into human behaviors.

Note that there is software to crawl the web, process text, and compute complex statistical procedures on the text, such as TACIT (e.g., Deghani et al., 2016). However, for the text analysis beginner, starting with simple word counts, and applying familiar statistical techniques typically used in social psychological studies might ease the understanding of text analysis as a tool in one's larger social science toolkit.

In the next section, we provide an overview of the studies that have used LIWC or other text analysis approaches to examine social psychological research questions. Then, we review the growth of text analysis across disciplines. Finally, we touch on some of the issues that the future of text analysis applications in social psychology faces, and future directions.

Social Psychological Applications

LIWC and other quantitative text analytic tools and strategies have been applied across a variety of topic areas within social psychology, including the assessment of status, romantic relationships, health, persuasion, forensics, and culture. As with any other measure in psychology, language should be assessed for its reliability and validity as a measure for any given construct. In addition, as with any other measure in psychology, one must consider the degree to which language is studied under naturalistic conditions, is appropriately powered, and has been investigated with multiple methods. The resources—including platforms, tools, applications, and statistical techniques—to study language are growing. Hand in hand with these technical resources, and our theoretical knowledge and empirical literature on human behaviors, social psychologists are able to make more reliable, generalizable, or nuanced statements, as well as new insights about individuals, groups, and culture.

Below, we provide an overview of select areas in which quantitative text analyses have been applied to social psychological research questions. For a review of quantitative text analysis of personality research questions, please see Chung and Pennebaker (forthcoming); Ireland and Mehl (2014).) Although we draw on research using other quantitative text analysis tools, we place a strong focus on the applications of LIWC, because it is the tool with which we are most familiar, and

because it is the most widely applied quantitative text analysis tool in the social sciences.

Status

It's interesting to watch two strangers interacting from a distance. Even though one may not be able to hear what they are saying, it is possible to get a sense of their emotional states, how well they know one another, and which one is more in control of the relationship. Interestingly, the analysis of function words can reveal some of these same dynamics. Whatever the context, people's thoughts, feelings, and behaviors tend to systematically change in response to different situations. With quantitative text analysis, we can find clues to their thoughts, feelings, and behaviors about each other in the language they use.

These discoveries have been facilitated by records of language between interactants in real-time (e.g., text messages, social media posts and comments, closed captioning and advances in transcriptions, etc.). Language clues are apparent in both the content of speech but also in pronouns and other function words. Specifically, the topic of conversation may reflect the type of relationship one has with another person. For example, words relating to work collaborations may include "analysis," "deadlines," "document," "funding," "presentation," "report," and "review." These content words are likely to appear in professional discussions and relatively unlikely to bubble up in a heated romantic encounter.

How interactants are speaking with one another via function words provides a different view into relationships that can be relatively independent from the topic. For example, higher status individuals tend to use more "we," while lower status individuals use more "I" (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Using LIWC, this status differential has been observed across various types of relationships, including in military documents (Hancock et al., 2010), terrorist group member statements and interviews (Pennebaker & Chung, 2008), U.S. President Richard Nixon's Watergate tapes (Chung & Pennebaker, 2007), quarterly earnings transcripts of incoming and outgoing CEOs (Kacewicz, 2013), and even between IMs of randomly paired, unacquainted college students asked to talk or message one another (Kacewicz et al., 2013).

Kacewicz et al. found reliable language markers of status across five studies. They also found reliable effects for higher word count and more second person pronouns by higher status interactants. Effect sizes for language as markers of personality traits such as the Big Five and the Dark Triad typically tend to be lower (for a review of language markers of personality, please see Chung & Pennebaker, forthcoming; Ireland & Mehl, 2014), while effects for demographics such as age, gender, and status tend to be much stronger (for a review, see Tausczik & Pennebaker, 2010). Sir Francis Galton hypothesized why this might be so in his Lexical Hypothesis of Personality (1884).

The Lexical Hypothesis

According to the Lexical Hypothesis of Personality (see also Goldberg, 1993):

Postulate 1: Traits that are important to our lives will be encoded in language.

Postulate 2: The most important traits are likely to be represented as a single word in language.

If we were to extend Galton's Hypothesis beyond personality to social dynamics, we might expect a Lexical Hypothesis of Social Life:

Postulate 1: Dynamics that are important to our social lives will be encoded in language

Postulate 2: The most important dynamics to attend to in our social lives are likely to be represented as a single word in language.

Postulate 3: The most important dynamics to attend to in our social lives are likely to be represented as the shortest, quickest-to-utter words in language.

The last postulate refers to function words, which tend to have a shorter number of letters than most words in our vernacular. In an interaction, function words quickly distinguish whom we should treat as the holder of power, resources, and tribal knowledge; whom we should attract as potential mates or hunting buddies; how far we should pitch our camp from them; how fast we should run from them if necessary. Even sighs and fillers, which aren't typically considered as full words in conversations, but transcribed in a few letters, can be indicative of well-being (Robbins et al., 2011) and demographics (Laserna, Seih, & Pennebaker, 2014).

Luckily, for many of us living thousands of years from the inception of formal language, function words are processed automatically in our frontal lobes, and so are read and spoken automatically. It is possible to infer these relationship attributes from the ways that people speak or write to each other, even when we ourselves are not a part of the conversation. If it's not obvious upon regular human observation, fear not, there are computerized text analysis tools such as LIWC to help decode relationships by examining pronoun use.

Relationship Dynamics

Given the intimate links between function words and social behaviors, it is not surprising that some of the most powerful and mysterious social psychological phenomena can be studied by looking at the ways people talk. In recent years, the computerized analysis of language has revealed new insights into our thinking about romantic attraction, persuasion, and emotional contagion.

We all have an intuitive sense when an interaction goes well or "clicks." We feel that we understand the other person and can practically finish each other's

sentences. These close connections are sometimes common among old friendships and, on other occasions, appear out of nowhere between two strangers. In recent years, several studies have analyzed the language of a wide range of social interactions and have identified the quality of people's relationships and, as mentioned above, their relative status. Some studies have examined how the words in a speed-dating interaction may be predictive of going out on a subsequent date (Ireland et al., 2011; Ranganath, Jurafsky, & McFarland, 2009), or staying in a relationship (Slatcher & Pennebaker, 2006).

A measure of how two people are using function word categories at the same rates, or are mirroring one another in their non-conscious word use is more predictive of mutual attraction than is a measure of how two people use content word categories at the same rates (Ireland & Pennebaker, 2010). This measure has been termed language style matching (LSM). Higher LSM and greater positive emotion word use in relationships are seen in longer lasting relationships (Slatcher & Pennebaker, 2006).

What is fascinating about this simple measure of LSM is that it is not only telling of relationship longevity, but it signals coordination on a much larger scale. LSM has been found to be higher in Wikipedia discussions for articles that have higher ratings (Pennebaker, 2011). That is, Wikipedia articles were judged to be better if editors communicated similarly. The degree to which community members use function words similarly has also been found to be higher in Craigslist ads for mid-sized cities with a gini coefficient that indicates that wealth is more evenly distributed (Pennebaker, 2011). These studies suggest that function word analyses or LSM can be used as a remote sensor of a dyad or group's internal dynamics.

Persuasion

LSM also plays an important role in the social dynamics of persuasion. Matching with an opponent's language style in a political debate has been shown to influence viewers who are watching the debate. In a study of U.S. presidential debates, Romero et al. (2015) found that candidates who matched to the style of their opponent fared better in the election polls, presumably because it demonstrates perspective taking and greater fluency. These are particularly interesting effects since previous research has shown that lower status interactants match more to their higher status interactants (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012).

Romero et al. (2015) replicated the presidential debate findings in a study of business student negotiations. Those who matched to their interactant were seen by third-party observers as having performed better or won the negotiation, and were more likely to be picked to negotiate for the third party observer. LSM, then, not only influences the dynamics and outcomes of an interaction, but it also affects the perception of an interaction by those who are mere observers.

Another study on persuasion and function words examined the Reddit forum “Change My View,” where users post a position statement on any topic, provide supporting reasons, and then receive counterarguments from other users. Tan, Niculae, Danescu-Niculescu-Mizil, and Lee (2016) found that pronouns held predictive power for malleable positions over specific topics (e.g., food, government, etc.). Specifically, individuals who were more likely to change their minds used more first person singular pronouns in their posts; individuals who were less likely to change their minds used more first person plural pronouns in their posts. These results suggest that it was easier to change a viewpoint that was presented as being held by the self as opposed to a viewpoint that was presented as being held by many. Note, that some content words indicating power and success (e.g., completion, smile) were predictive of resistant posts, but topics (e.g., food, government, etc.) did not add predictive power to which posts led to changing the user’s mind.

Another way in which influence has been studied using LIWC has been through the examination of word use on millions of posts on Facebook News Feeds (Kramer, Guillory, & Hancock, 2014). Specifically, an experimental study that systematically reduced the presentation of posts with LIWC’s positive and negative emotion word categories showed that emotional contagion propagates via words in the absence of non-verbal cues. That is, a reduction in the proportion of posts seen with positive or negative emotion word use led to a significant reduction in corresponding positive and negative emotion word expression respectively by connections that were exposed to the experimental manipulations relative to controls. In addition, there was a significant increase in expressing the opposite emotion by connections that were exposed to the experimental manipulations relative to controls.

While the effect sizes were small, the fact that social networks are by definition interconnected suggests that the effects can be wide reaching. The words we use can have profound effects on how the people around us experience their worlds, and in turn, how they influence those around them. This is increasingly important to attend to as we increase our interactions over CMC, and in increasingly connected media.

Forensics

Language markers for forensic analyses have been identified and applied to open-ended statements, emails, papers, and conversations. For example, in studies where participants have been asked to lie in laboratory studies (e.g., Hancock, Curry, Goorha, & Woodworth, 2008; Newman, Pennebaker, Berry, and Richards, 2003), in courtroom transcripts of those found guilty of committing a crime and convicted of perjury (Huddle & Pennebaker, 2009), or in online dating profiles (Toma & Hancock, 2012), language analyses have shown that there is less use of first person singular pronouns in deceptive statements. Presumably, this is due to the lack of ownership of deceptive statements or psychological distancing.

However, in verified fake hotel reviews, it has been found that first person singular pronouns appear at higher rates, relative to genuine hotel reviews (Ott, Cardie, & Hancock, 2012). The authors theorize that placing one's self in the hotel setting is an important feature of this particular type of deception, suggesting that context is important when considering certain types of lies.

Word use has been tracked at the individual level to predict violent crimes by the Boston Bombers (Norman-Cummings & Pennebaker, 2013), by extremist groups in the Middle East (Pennebaker, 2011b), and by leaders intending on going to war (Chung & Pennebaker, 2011). In each of these cases, a drop in first person singular pronouns preceded violent acts. Since an attack involves hiding or concealing one's intentions to surprise "the enemy," it makes sense that language markers of deception are found in the language of attackers leading up to their violent acts.

From a forensics perspective, the ability to spot betrayal and secret-keeping has long been of interest to language scientists. For example, Niculae and colleagues (2015) found that it was possible to identify whether someone would betray their online gaming partner through more positive words, more politeness, and fewer words indicating future plans. Some cues for betrayal came from changes in language use by the victim; victims tended to increase in their use of planning words before betrayal. What was particularly intriguing was that the linguistic shifts in betrayal were apparent in both the betrayer and the betrayed through an increasing imbalance in the use of specific word categories between the interactants.

In another study, our research team tracked the emails of 62 people who admitted to keeping a major life secret from others (Tausczik, Chung, & Pennebaker, 2016). Participants were recruited online and were carefully screened in ways that preserved their anonymity, as well as the specific details of their secret. In all cases, those who agreed to release their previous year's emails were keeping secrets that, if discovered, would have been devastating to their lives or to the lives of others around them. As with the Niculae study, both the language of the secret-keepers and the language of the targets of the secrets changed. By examining function words, we were able to uncover the common ways in which people experience secret-keeping, and how it affected their relationships. Together, the results showed how psychological features can still be extracted when the topic of the exchange is not known to researchers. The secrets study in particular was the first to show how the language of a social network changes in response to a devastating life secret.

Even in forensic studies not involving extreme acts or violent crimes, language has provided clues to delinquent or mysterious activity, for example, in the papers of Diederik Stapel, a psychologist who was found to have been fabricating data for several of his published papers. An analysis of Stapel's articles confirmed to be fabricated vs. those not found to be fabricated showed more terms associated with scientific methods and certainty, and fewer adjectives (Markowitz & Hancock, 2014). These indicate that Stapel was emphasizing the novelty and significance of

the results within the scientific literature, but unable to describe more concretely what he was reporting on.

Finally, investigators occasionally seek to learn who may have written a ransom note or even an entire book or play. As early as the 1960s, two statisticians applied early Bayesian analyses on *The Federalist Papers* to identify the authorship of a select group of disputed pamphlets. Mosteller and Wallace (1963) found that differences in a group of function words could serve as fingerprints to identify if the disputed papers were by Alexander Hamilton or James Madison. Similarly, Boyd and Pennebaker (2015) used LIWC and machine learning methods on both function and content words and concluded that a long-disputed play, *Double Falsehood*, was probably written by William Shakespeare.

The field of forensics will undoubtedly expand considerably in the years to come with increasingly sophisticated text analytic methods. Not only will investigators be able to identify authors but they will be able to better detect the intent or ongoing behaviors and personalities of the authors.

Health

When people are sick, uncertain about medical procedures, or dealing with health-related life changes, they inadvertently broadcast their situation online in ways potentially detectable by text analysis. For example, it is now possible to predict when couples might be expecting a child, and whether or not a new mother is likely to experience postpartum depression based on her tweets (de Choudhury, Counts, Horvitz, & Hoff, 2014). It is also possible to identify those at suicide risk from Facebook posts (Wood, Shiffman., Leary, & Coppersmith, 2016), or if individuals who are more or less likely to lose weight based on their diet blogs (Chung, 2009). Through other Twitter analyses, it is possible to isolate particular geographical locations more likely to experience higher rates of HIV (Ireland, Schwartz, Chen, Ungar, & Albarracín, 2015) or heart disease (Eichstaedt et al., 2015). Particularly exciting have been studies that track Wikipedia searches (Tausczik, Fasse, Pennebaker, & Petrie, 2012) or even vaping rates based on searches (Ayers et al., 2016). Note, however, that there have been failures to replicate some patterns gleaned from dynamic big data without corresponding traditional study methods (see Lazer, Kennedy, King, & Vespignani, 2014), such as flu epidemics based on Google searches (Ginsberg et al., 2009), suggesting that text analysis, like any other method in the social sciences, works best as a part of a multi-method toolkit.

Notice that these larger, more sociological or epidemiological questions appear more in content words, while clues to the more psychological questions appear more in function words. For example, in the aftermath of 9/11, livejournal.com blogs were examined for words representing preoccupation with the attacks (e.g., Osama, hijack, World Trade Center, etc.), positive and negative emotion words, and psychological distancing (i.e. a statistically derived index made of articles, first person singular pronouns, and discrepancy terms; Cohn, Mehl, &

Pennebaker, 2004). Not surprisingly, the analysis of content words revealed that communities across America were attending to death, religion, and the attacks much more after 9/11 than before. Mood went back to baseline levels (i.e. pre-9/11 levels) within a couple weeks, even for those who were highly preoccupied with the attacks. Function words, on the other hand, showed that psychological distancing persisted at least six weeks after 9/11. Individuals were talking less about themselves, and using “we” more to refer to their communities (Pennebaker, 2011). The content and function word analyses provided a timeline of topics and how widespread communities were psychologically responding to a massive upheaval in a naturalistic way.

The ability to detect and to predict the symptoms of various diseases, well-being, and community resilience after widespread upheaval is now possible through the text analysis of social media. Given that more and more of our interactions are online, and the ability to find people with similar symptoms and diseases has been made easier, it is possible that treatments, coping, recovery, and prevention strategies can be developed from our online interactions and behavior, although these are not without serious considerations, such as privacy, in their application (de Choudhury, 2013; Resnick, Resnick, & Mitchell, 2014; Wood et al., 2016).

Culture

Words can mark changes over time and place, providing new ways to assess the attentional focus of individuals, groups, and entire societies. By aggregating texts from the historical record, we can begin to track large-scale historical and cultural trends. The largest project of words over time examined keywords across 4 million digitized books (Michel et al., 2011). The authors counted word use over time to assess cultural trends (e.g., sushi, plagues, technology, etc.). The text analysis of cultural products has also allowed for the examination of psychological trends over time, including the examination of various values (Bardi, Calogero, & Mulen, 2008), individualism vs. collectivism (Twenge, Campbell, & Gentile, 2012), and sentiment (de Wall, Pond, Campbell, & Twenge, 2011) across recent history. Custom dictionaries using LIWC have been used to track possible cultural differences in the moral foundations of Liberals and Conservatives (Graham, Haidt, & Nosek, 2009), and in the relationships between prosocial language as predictive of public approval of U.S. Congress (Frimer, Aquino, Gebauer, Zhu, & Oakes, 2015).

Even within smaller geographic regions, it has been possible to track how pronoun use over time is associated with inciting action, for example, rallying for a revolution in the lead up to Iranian elections (Elson, Yeung, Roshan, Bohandy, & Nader, 2012), and within an online community, how pronouns are indicative of community tenure or life stage (i.e., number of posts from joining to leaving; Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013).

Given the access to digital written pieces, collaborative works, and interactions taking place over global platforms, there has been a growth in the assessment of

regional differences in social psychological processes (Rentfrow, 2014). For example, a text analytic study of a corpus of essays on American's beliefs was assessed for values held across various states, and their relationship to state-level statistics published by national agencies (Chung, Rentfrow, & Pennebaker, 2014). Similarly, studies of tweets across major cities in America showed relationships of word use to rates of heart disease (Eichstaedt et al., 2015) and HIV prevalence (Ireland et al., 2015) assessed by the Centers for Disease Control.

Another study of tweets by U.S. counties found that high state level well-being and life satisfaction was associated with words indicating outdoors activities, spiritual meaning, exercise, and good jobs; low state life satisfaction was associated with negative emotion words indicating boredom (Schwartz et al., 2013b). A study of Facebook posts suggested that relative positive and negative emotion word use could form an unobtrusive assessment of gross domestic happiness (Kramer, 2010). Together, these studies suggest that word comparisons across geographies can reveal systematic social processes that indicate relative health or well-being, providing insights into the relationships between sociological forces on psychological processes.

The Promise of Text Analytic Methods

Text Analytic Goals Across Fields

There are special tools, such as TACIT (Dehghani et al., 2016) that amass, extract, and code language. With word counts as the basic foundation of quantitative text analysis, a variety of simple and highly complex statistical techniques have been applied to code and decode text. For example, one software tool to derive psychological insights from word counts of content words is the Meaning Extraction Helper (Boyd, 2016), which facilitates the implementation of the Meaning Extraction Method (Chung & Pennebaker, 2008), a factor analysis of words to inductively extract themes from text. Note that there are a growing number of open-vocabulary approaches to analyze text (Schwartz et al., 2013a).

Text analysis methods span multiple disciplines. The interested reader should explore computer science and linguistics and, within both fields, natural language processing (NLP). NLP has the goal of classifying documents according to their features, and specifically, by the language used within a document. NLP methods are useful for psychological quantitative text analysis, but differ from psychological methods in various ways.

NLP methods typically require more programming skills to implement than are offered in traditional social psychology graduate programs, with much more preprocessing of text involved. There are a wide variety of open-source toolkits (e.g., Natural Language Toolkit [or NLTK], Stanford CoreNLP, etc.) that provide the code to execute the myriad of preprocessing steps (e.g., stemming, lemmatizing, tokenization, etc.) to prepare text for feature extraction, as well as to carry out a variety of analysis.

A very general distinction between social psychologists and NLP researchers is their purpose for studying language. Social psychologists typically use language as a reflection of social, cognitive, or emotional processes of the speaker or writer. NLP scientists, on the other hand, have one of two general underlying motives for their interest in language. Linguists or computational linguists typically analyze language because they tend to be more interested in language, and typically (but not always) are less interested in context or the attributes of the speaker than are psychologists. In contrast, computer scientists use language to categorize attributes into two or more categories using a variety of statistical methods. For example, it is possible to distinguish between genuine emails from spam by analyzing the features of the emails themselves. Like linguists, computer scientists are generally less interested as psychologists are in learning about the social psychological dynamics of the author of the texts themselves.

Technological Enablers of Text Analysis Growth

Particularly exciting is that researchers from psychology, computer science, and linguistics are now beginning to work together as part of a new discipline variously called Cognitive Science, Computational Social Science, and/or Artificial Intelligence. Each of these fields has benefited from the ways in which humans communicate, work, and socialize, resulting in significant scientific steps forward. The increase in CMC has enabled us to record, amass, extract, code and decode, and assess the meaning and style of text. These, in turn, have enabled us to develop more insights into potential applications with which to communicate digitally, or to capture communication digitally. Along with the increase in connectivity and CMC, there has been a surge of work in real-time speech-to-text capabilities, more language based digital art, machine translation, optical character recognition, machine translation, faster computing, multimedia systems, and an internet of things, statistical learning techniques, and cross disciplinary and cross industry collaborations to support CMC, and accordingly, to support the analysis of natural language. The amount of data collected on a person in association with their communication patterns presents many opportunities for research and innovation.

For example, natural language samples from social media can typically be associated with the user or other meta-data available (geolocation, topic, group affiliation, time of post, etc.). When the information is public, such as Reddit posts or Twitter posts, there are APIs for extracting the information, or more user-friendly tools to call on the platform's API (e.g., TACIT, Dehghani et al., 2016). What does this mean for social psychologists?

Social Psychological Applications of Text Analysis

Language is a defining feature of human culture and we are now only beginning to be equipped with the tools to study it on a massive scale. With all the technological advancements, cross-disciplinary collaborations, and our greater reliance

on CMC throughout all areas of our lives, our insights will only continue to grow faster, more creative, and with broader applications.

There are two important caveats. The first is that while the study of the data we create as we go about our daily lives brings great benefits to our understanding of ourselves, our relationships, our health, our work, our deviance, and our culture, there must be limits and controls to ensure these benefits are weighed against the costs. There will be greater focus on the use of data and opt-outs/opt-ins beyond its collection, and terms of service or end-user license agreements (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Mundie, 2014; PCAST, 2014; Verma, 2014).

The second caveat is that the variance that word scientists account for is low. We often publish exciting results because we have access to giant data sets. Enabled by advances in computerized text analysis and access to massive archives of digitized text, we are finding patterns that no one has seen before. But the effects are subtle. These insights into human behavior and social dynamics are illuminating, and are particularly handy when text is the only behavior we are able to objectively observe. However, their reliability and validity across contexts have yet to be assessed.

As our lives become increasingly digital, and the more we are able to extract psychological patterns from text, the more we open up to possibilities of being able to feedback analytics in real-time, or to predict behaviors across our social networks. We have never before been able to grasp what and whom we influence and how we influence to the degree that is possible today, and that possibility is only growing. The future of text analysis is amazingly exciting, with great potential for our understanding of how we are influenced by the actual, implied, or imagined presence of others.

References

- Ayers, J. W., Althouse, B. M., Allem, J-P., Leas, E. C., Dredze, M., & Williams, R. S. (2016). Revisiting the rise of electronic nicotine delivery systems using search query surveillance. *American Journal of Preventive Medicine*, 40(4), 448–453. doi:10.1016/j.amepre.2015.12.008
- Bardi, A., Calogero, R. M., & Mullen, B. (2008). A new archival approach to the study of values and value-behavior relations: Validation of the value lexicon. *Journal of Applied Psychology*, 93, 483–497.
- Boyd, R. L. (2016). *Meaning extraction helper (MEH)*. Software program.
- Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, 26(5), 570–582. doi:10.1177/0956797614566658
- Chung, C. K. (2009). *Predicting weight loss in diet blogs using computerized text analysis* (Unpublished dissertation). Austin, TX: University of Texas.
- Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343–359). New York, NY: Psychology Press.

- Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, 42, 96–132.
- Chung, C. K., & Pennebaker, J. W. (2011). Using computerized text analysis to assess threatening communications and actual behavior. In C. Chauvin (Ed.), *Threatening communication and behavior: Perspectives on the pursuit of public figures* (pp. 3–32). Washington, DC: The National Academies Press.
- Chung, C. K., & Pennebaker, J. W. (forthcoming). What do you know when you LIWC a person? Text analysis as an assessment tool for traits, personal concerns, and life stories. In T. Shackelford & V. Ziegler-Hill (Eds.), *The SAGE handbook of personality and individual differences*. New York, NY: SAGE Publishing.
- Chung, C. K., Rentfrow, P. J., & Pennebaker, J. W. (2014). Finding values in words: Using natural language to detect regional variations in personal concerns. In P. J. Rentfrow (Ed.), *Geographical psychology: Exploring the interaction of environment and behavior* (pp. 195–216). Washington, DC: The American Psychological Association.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15, 687–693.
- Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. L. (2012). *Echoes of power: Language effects and power differences in social interaction*. Proceedings of the 21st international conference on World Wide Web (WWW '12). New York, NY: ACM, pp. 699–708. DOI=<http://dx.doi.org/10.1145/2187836.2187931>
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). *No country for old members: User lifecycle and linguistic changes in online communities*. Proceedings of the International World Wide Web Conference Committee (IW3C2), Rio de Janeiro, Brazil.
- De Choudhury, M. (2013). *Role of social media in tackling challenges in mental health*. Proceedings of the 2nd international workshop on Socially-aware multimedia (SAM '13). New York, NY: ACM, 49–52. DOI=<http://dx.doi.org/10.1145/2509916.2509921>
- De Choudhury, M., Counts, S., Horvitz, E., & Hoff, A. (2014). *Characterizing and Predicting Postpartum Depression from Facebook Data*. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). New York, NY: ACM, 626–638. DOI: <https://doi.org/10.1145/2531602.2531675>.
- Deghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Parmar, N. J. (2016). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, 49(2), 538–547.
- deWall, C. N., Pond, R. S., Jr., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, 5, 200–207.
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2), 159–169.
- Elson, S. B., Yeung, D., Roshan, P., Bohandy, S. R., & Nader, A. (2012, February 29). *Using social media to gauge Iranian public opinion and mood after the 2009 election*. Santa Monica, CA: RAND Corporation, TR-1161-RC, 2012. Retrieved from www.rand.org/pubs/technical_reports/TR1161
- Frimer, J. A., Aquino, K., Gebauer, J. E., Zhu, L., & Oakes, H. (2015). A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences*, 112, 6591–6594. doi:10.1073/pnas.1500355112

- Galton, F. (1884). Measurement of character. *Fortnightly Review*, 36, 179–185.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96, 1029–1046.
- Hancock, J. T., Beaver, D. I., Chung, C. K., Frazee, J., Pennebaker, J. W., Graesser, A. C., & Cai, Z. (2010). Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences in Terrorism and Political Aggression, Special Issue: Memory and Terrorism*, 2, 108–132.
- Hancock, J. T., Curry, L., Goorha, S., & Woodworth, M. T. (2008). On lying and being lied to: A linguistic analysis of deception. *Discourse Processes*, 45, 1–23.
- Hart, R., & Carrol, C. E. (2014). *Diction 7.0* [Computer software] Austin, TX: Digitext, Inc.
- Huddle, D., & Pennebaker, J. W. (2009). *Language analysis of jury testimony from properly and wrongly convicted people*. Unpublished manuscript. University of Texas.
- Ireland, M. E., & Mehl, M. R. (2014). Natural language use as a marker of personality. In T. Holtgraves (Ed.), *Oxford handbook of language and social psychology*. New York, NY: Oxford University Press.
- Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99, 549–571.
- Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L., & Albarracín, D. (2015). Future-oriented Tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, 34, 1252–1260.
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1), 39–44. doi:10.1177/0956797610392928
- Kacewicz, E. (2013). *Language as a marker of CEO transition and company performance* (Unpublished dissertation). Austin, TX: The University of Texas at Austin.
- Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33, 124–143. doi:10.1177/0261927X1350265
- Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. *American Psychologist*, 70(6), 543.
- Kramer, A. D. I. (2010). *An unobtrusive behavioral model of “gross national happiness”*. Proceedings of Computer-Human Interaction (CHI), pp. 287–290.
- Kramer, A. D. I., & Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 8788–8790. doi:10.1073/pnas.1320040111
- Laserna, C. M., Seih, Y., & Pennebaker, J. W. (2014). Um, who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, 33, 328–338. doi:10.1177/0261927X14526993
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014, March 14). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203–1205. doi:10.1126/science.1248506

- Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PLoS ONE*. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0105937>
- Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141–156). Washington, DC: American Psychological Association.
- Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, 64, 1306–1315.
- Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *British Journal of Medical Psychology*, 72, 339–354.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 14, 176–182. doi:10.1126/science.1199644
- Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58, 275–309.
- Mundie, C. (2014). Privacy pragmatism: Focus on data use, not data collection. *Foreign Affairs*, 93(2), 28–38.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29, 665–675.
- Niculae, V., Kumar, S., Boyd-Graber, J., & Danescu-Niculescu-Mizil, C. (2015). *Linguistic harbingers of betrayal: A case study on an online strategy game*. Proceedings of the Association for Computational Linguistics (ACL2015). 1. 10.3115/v1/P15-1159.
- Norman-Cummings, B., & Pennebaker, J. W. (2013). *Tracking the Tweets of the Boston Marathon Bomber: A text analysis strategy for threat detection*. Unpublished manuscript, University of Texas.
- O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. *Public Health*, 41: 43.
- Ott, M., Cardie, C., & Hancock, J. (2012). *Estimating the prevalence of deception in online review communities*. Proceedings of the International World Wide Web Conference Committee (pp. 201–210). ACM.
- Pennebaker, J. W. (2011a). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.
- Pennebaker, J. W. (2011b). Using computer analyses to identify language style and aggressive intent: The secret life of function words. *Dynamics of Asymmetric Conflict*, 4, 92–102. Retrieved from <http://dx.doi.org/10.1080/17467586.2011.627932>
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count (LIWC2015)*. Austin, TX. Retrieved from www.liwc.net
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count (LIWC2007) [Computer software]*. Austin, TX. Retrieved from www.liwc.net
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi:10.15781/T29G6Z
- Pennebaker, J. W., & Chung, C. K. (2008). Computerized text analysis of al-Qaeda statements. In K. Krippendorff & M. Bock (Eds.), *A content analysis reader* (pp. 453–466). Thousand Oaks, CA: Sage Publications.

- Pennebaker, J. W., Chung, C. K., Ireland, M. I., Gonzales, A. L., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX. Retrieved from liwc.net
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC2001)* [Computer software]. Mahwah, NJ: Lawrence Erlbaum Associates.
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, 547–577.
- President's Council of Advisors on Science and Technology (PCAST; 2014). *Report to the President: Big data and privacy: A technological perspective*. Retrieved from www.whitehouse.gov/ostp/pcast
- Ranganath, R., Jurafsky, D., & McFarland, D. (2009). *It's not you, it's me: Detecting flirting and its misperception in speed-dates*. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1–Volume 1. Association for Computational Linguistics, pp. 334–342.
- Rentfrow, P. J. (2014). *Geographical psychology: Exploring the interaction of environment and behavior*. Washington, DC: The American Psychological Association.
- Resnick, P., Resnick, R., & Mitchell, M. (2014). *Workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. Proceedings of the 2014 Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics.
- Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kastle, S. (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, 30, 129–133.
- Romero, D. M., Swaab, R. I., Uzzi, B., & Galinsky, A. D. (2015). Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Personality and Social Psychology Bulletin*, 41(10), 1311–1319.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., & Lucas, R. E. (2013b). *Characterizing geographic variation in well-being using tweets*. In Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013), online. Retrieved from http://www.wbwp.org/papers/icwsml2013_cnty-wb.pdf
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, Shah, A., . . . Ungar, L. H. (2013a). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0073791>
- Slatcher, R. B., & Pennebaker, J. W. (2006). How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science*, 17, 660–664.
- Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.
- Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). *Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions*. Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 613–624.
- Tausczik, Y. R., & Chung, C. K., & Pennebaker, J. W. (2016). *Tracking secret-keeping in emails*. Proceedings of the 2016 International Conference on Weblogs and Social Media, (pp. 388–397).
- Tausczik, Y. R., Fasse, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following the H1N1 outbreak: Blogs, newspaper articles, and Wikipedia visits. *Health Communication*, 27, 179–185.

- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.
- Toma, C. L., & Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication, 62*, 78–97.
- Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in individualistic words and phrases in American books, 1960–2008. *PLoS ONE, 7*(7), e40181.
- Verma, I. M. (2014). PNAS Editorial expression of concern and correction. *PNAS, 111*(29), 10779.
- Wood, A., Shiffman, J., Leary, R., & Coppersmith, G. (2016). *Language signals preceding suicide attempts*. Proceedings of Computer-Human Interaction (CHI 2016), San Jose, CA.

8

DATA TO DIE FOR

Archival Research

Brett W. Pelham

Baseball great Yogi Berra once noted that “you can observe a lot just by watching.” Berra’s point is consistent with a great deal of archival research. Archival research is merely passive observational research that relies on data collected or created by others. Birth, marriage, or death records, for example, can become archival data. Thus, good archival research can be even easier than Yogi Berra’s casual research. In the case of archival research, others have usually done the watching for you.

Berra was also saying that good observations don’t always have to be complicated. The simplicity of archival methods stands in sharp contrast to the sophistication of many other modern research techniques. Consider computer programs that can (a) parse human speech into linguistic categories (Tausczik & Pennebaker, 2010), (b) disentangle participant- and group-level patterns in hierarchically organized data (Raudenbush & Bryk, 2002), or (c) analyze a video of a dance and calculate the degree of synchrony between dancer and music (Brick & Boker, 2011). Archival research usually stands in contrast to such high-tech observations. Nonetheless, archival studies can yield crucial information that cannot be obtained any other way. Archival studies can tell us how well a theory travels to the real world. Further, the best archival studies do not merely tell us *that* something can happen in the real world. They tell us *when* and even *why* (Cook & Campbell, 1979; McGuire, 1989).

Strengths and Weaknesses of Archival Studies

Archival studies have other strengths. A big problem with experiments is the fact that they can make people behave unnaturally. This may happen, for example, because people know they are being watched. In fact, Bateson and colleagues found that people behave better than usual when something merely *reminds them*

of being watched. When posters portray watchful human eyes rather than flowers, people more often (a) pay for coffee on the honor system and (b) clean up after themselves after eating (Bateson, Nettle, & Roberts, 2006; Ernest-Jones, Nettle, & Bateson, 2011). Because human beings are sensitive to being watched, experimenters must usually work very hard to get people to behave naturally (Aronson & Carlsmith, 1968).

Archival research bypasses the dilemma of unnaturalness by studying natural behavior (Pelham & Blanton, 2013). It also bypasses experimenter bias by eliminating the experimenters who unwittingly create this problem. For example, I know that U.S. consumers purchase regular octane gasoline more often than higher octane gasoline. This is because I have noticed that the buttons for regular gasoline typically show excessive signs of wear. This is a very casual archival observation, but it still bypasses concerns that consumers select regular gasoline because they are trying to impress an experimenter. Likewise, archival studies of criminal behavior solve a major ethical problem when they examine public records of murders—rather than trying to induce murder in the laboratory.

This chapter examines strengths and weaknesses of archival research, with an emphasis on some of the poorly appreciated strengths. For example, I argue that archival research has gotten a bit of a bad rap when it comes to establishing internal validity. Next, I suggest two simple recipes for being a solid archival researcher. The first recipe (the OOPS heuristic) is a checklist for critically analyzing and maximizing the external validity of archival research. The second recipe (the GAGES heuristic) tackles internal rather than external validity. It suggests that if researchers address five pesky confounds, they will often have gone a long way toward maximizing internal validity. After sketching out these two heuristics, I discuss a series of archival research studies that focus on topics as diverse as social judgment, the self-concept, and longevity. My goal in so doing is not merely to extoll the virtues of archival research. It is also to show that even novice researchers can conduct highly rigorous research. OOPS and GAGES are recipes for success.

John Stuart Mill and Internal Validity

As the British philosopher John Stuart Mill (1863) would be quick to note, archival researchers often face serious challenges establishing causality (Pelham & Blanton, 2013). If we update Mill's insights to incorporate modern statistical language, Mill argued that there are three requirements for establishing causality.

Covariation

Mill's first requirement is covariation. For Y to cause Z, changes in Y must covary with changes in Z. At a minimum Y and Z must be correlated. Distress appears to cause divorce. It certainly covaries with it. On the other hand, divorce can cause

at least some people to become distressed (Booth & Amato, 1991). Covariation is consistent with either temporal sequence.

Temporal Sequence

Mill also argued that if Y causes Z, changes in Y must *precede* changes in Z. In many passive observational studies (including archival studies) researchers have little or no information about temporal sequence. Consider the Neanderthal diet. Microbiological and isotopic analyses of the teeth, bones, and even the dried-up feces of Neanderthals reveal that the typical Neanderthal diet was about 80% meat. Our extinct cousins ate much more meat than *Homo sapiens* did (Sistiaga, Mallol, Galván, & Summons, 2014; Wißing et al., 2015). At one level it's obvious that specific Neanderthals were Neanderthals *before* they ate so much meat. They were born that way. On the other hand, thinking about temporal sequence on an evolutionary scale, some have argued that Neanderthals took the particular evolutionary route they did because their ancestors migrated to a part of the earth in which (a) a very cold climate and (b) the preponderance of megaherbivores set them on a specific evolutionary path. Neanderthals may have *become* the short, muscular, cold-adapted hominids they were because they were surrounded by so much steak—and so little salad.

But things aren't *always* this tricky. In fact, some archival studies do include information about temporal sequence. Consider archival studies of sports. In a baseball game, innings clearly indicate temporal sequence. Further, even if we ignore innings in baseball, we can often eliminate worries about reverse causality. Archival studies of aggression in baseball show that pitchers are more likely to hit batters with “bean balls” on hotter days than on cooler days (Reifman, Larrick, & Fein, 1991). We do not need to worry that a pitcher's decision to throw a ball at a batter changes the ambient temperature. We also do not need to worry that pitchers simply become less accurate as the temperature rises. On hotter days pitchers actually walk *fewer* batters than usual, and throw slightly fewer wild pitches. An archival study of more than 57,000 baseball games also showed that a pitcher is more likely to hit a batter when one of his own players was just hit by the pitcher from the opposing team (Larrick, Timmerman, Carton, & Abrevaya, 2014). That's payback. Archival research also shows that athletes repay good deeds as well bad ones (Willer, Sharkey, & Frey, 2012). Information about temporal sequence is not embedded in all archival records. But many archival data sets fare surprisingly well when it comes to temporal sequence.

Eliminating Confounds

When it comes Mill's third rule of causality—eliminating all possible confounds—archival studies do *not* usually fare as well as thoughtfully designed surveys. Survey designers usually choose exactly which questions people answer. Thus,

well-informed survey researchers can measure and statistically control for any confounds about which they are worried. In contrast, when researchers rely on data collected by others, they often see that these others were rarely obsessed with eliminating confounds.

Despite this gloomy observation, there are reasons for optimism. First, some sources of archival data do exist that include measures of important confounds. Some archival data sets were created by scientists rather than sports enthusiasts. The standard cross-cultural sample (Murdock & White, 2006) and the World Values Surveys (www.worldvaluessurvey.org) are two invaluable examples. In these rich archival data sets, one researcher's survey question or ethnographic observation is another researcher's confound. Second, some archival studies are natural experiments, which solve the problem of confounds by giving people a manipulation at random (or nearly so). Third, after-the-fact coding in an archival study can yield important insights into why a specific outcome occurred. For example, archival data provide rankings of the market share of many brands of a specific product (e.g., fast food). It is possible to have consumers rate these stimuli (e.g., McDonald's, Wendy's, Arby's) on multiple dimensions (e.g., price, customer service) to see which dimensions do the best job in a statistical footrace that predicts the brand rankings.

External Validity and the OOPS Heuristic

John Stuart Mill would surely love lab experiments. But some features that make lab experiments high in internal validity make them low in external validity. External validity is about generalizability to the real world, and archival studies usually examine real world events. I've argued elsewhere (Pelham, 2017) that concerns about external validity almost always fall into one of four categories. Each letter of the OOPS heuristic represents one of these categories. Archival research usually does a good job of addressing the OOPS concerns. What are these concerns?

Operationalizations

Because testability is a cherished scientific canon, psychologists acknowledge that we can only study things scientifically if we specify operational definitions (Pelham & Blanton, 2013). But there are many ways to operationalize most hypothetical constructs. Consider sexual arousal. In men, a useful operationalization is change in penis size. Thus, researchers in human sexuality developed the plethysmograph, which measures penis volume (Adams, Wright, & Lohr, 1996). But if we wish to study sexual arousal in the other 52% of the earth's population, plethysmographs are useless. For this reason, experts developed a measure of sexual arousal that works for women as well as men. Thermographic stress analysis (TSA) involves assessing changes in genital temperature. A thermography camera can

detect changes of about one fourteenth of a degree Celsius in a very brief period. Thermography works just as well for women as it does for men (Kukkonen, Binik, Amsel, & Carrier, 2010).

We can place greater confidence in any research finding when it holds up well across multiple operationalizations. *Altruism*, for example, could be defined as (a) giving food to another organism or (b) risking your own safety to protect another organism (Dawkins, 1976). Nursing your infant daughter fits the first definition. Saving a drowning stranger fits the second. Along these lines, kin selection (making sacrifices for organisms with whom you share genes) is now widely accepted because there is good evidence for kin selection regardless of which definition of altruism one adopts (Hamilton, 1964a, 1964b; Smith, Kish, & Crawford, 1987). One reason why Pinker's (2010) argument that human violence has declined over the centuries is so convincing is that Pinker's archival studies use *many* operational definitions of violence, from killing or enslaving people to spanking children, burning witches, or hurting animals in films. Across numerous operational definitions, violence has declined, especially in this century.

Occasions

"To everything there is a season." Human behavior varies greatly across the day, across seasons, and across millennia. College students report having sex much more often between 11 p.m. and 1 a.m. than at any other time of day (Refinetti, 2005). People's hormone levels also vary naturally over time. This variation often has important consequences. Welling and colleagues (2008) showed that men rated highly feminine female faces to be more attractive than usual on days when the men's testosterone levels were higher than usual.

Looking at timing over a broader window, both births and deaths vary with the seasons. Consider the archival data in Figure 8.1. They show that Americans more often die in winter than in summer (despite the fact that most deaths by accident are more common in the summer; Rozar, 2012). There is debate about exactly why this seasonal pattern occurs, but the pattern is clearly *seasonal* rather than calendrical. The pattern disappears at the equator and is reversed in the Southern hemisphere. Marriage rates, too, vary over the course of the year. As you already knew, June is the most popular month for American weddings. As you probably did *not* know, people are also more likely to get married during the month of their own birthdays than in other months (Pelham & Carvallo, 2015).

Time also matters century by century. Two thousand years ago, Romans died more often in the summer than in the winter (because diseases like malaria were much more common in summer; Scheidel, 2009). At that time, the entire population of the earth was smaller than the current population of Indonesia. Turn the clock back to 10,000 years ago, when agriculture barely existed, and the earth's human population was smaller than the current population of Chicago.

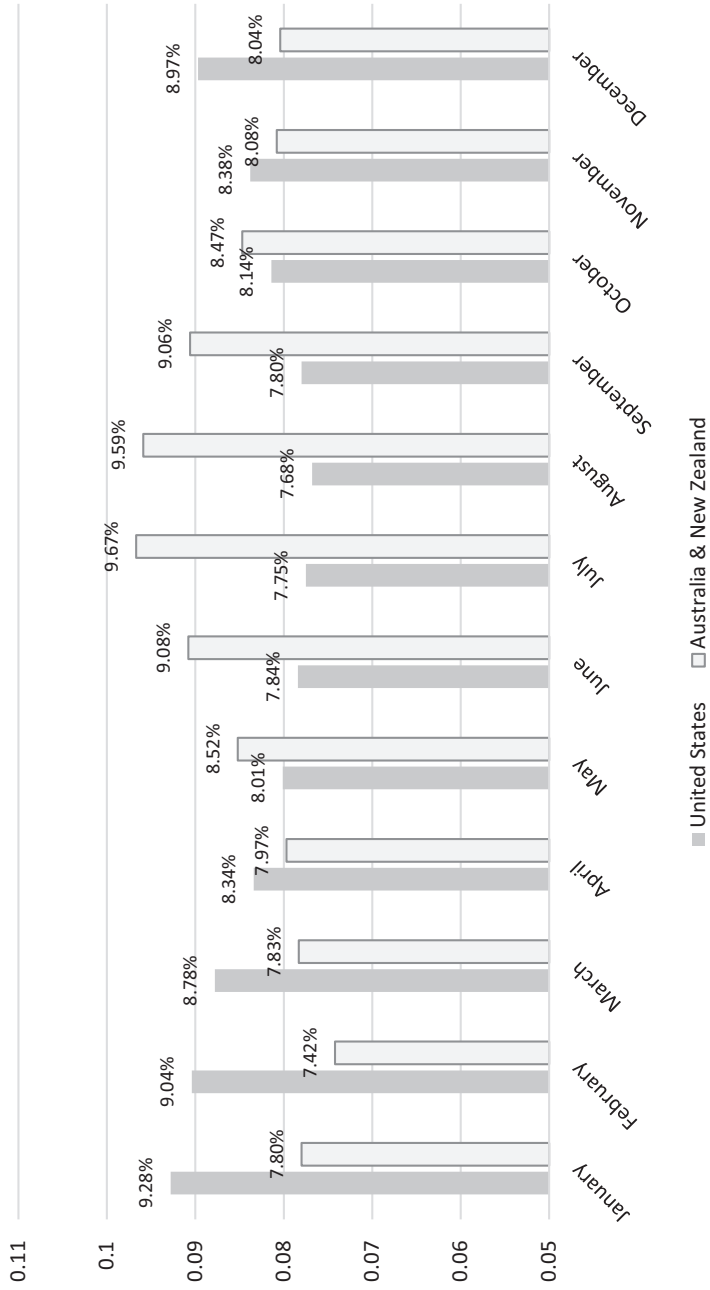


FIGURE 8.1 Likelihood of Dying by Month in the United States versus Australia and New Zealand, 1990–2010 (Adjusted for the Number of Days in a Month)

Note: U.S. Data were harvested from the Social Security Death Index (SSDI). Australia and New Zealand data were harvested from the Australia and New Zealand “Find a Grave Index.” Both archival data sources were accessed at ancestry.com.

So time matters. Thus, when evaluating any research finding, we must ask if that finding would hold true at other times. Showing that something interesting is true is impressive. Showing that it was also true 1,000 years ago is more impressive. Time is particularly important in developmental and evolutionary psychology. What cognitive skills do most toddlers possess that most infants do not? How have hominids changed over the past million years? Does interest in sex vary temporally? Gangestad and colleagues (2010) found that women's degree of sexual opportunism varies across their ovulatory cycle. When women are at the most fertile point in their cycles, they're more likely than usual to endorse attitudes such as "I believe in taking my sexual pleasures where I find them."

Populations

Almost no research finding applies to every imaginable population. People see most colors better than dogs do, and dogs smell things people can barely imagine. European magpies recognize themselves in mirrors but most other birds do not (Prior, Schwartz, & Gunturkun, 2008). Chimps are more sexually promiscuous than people or gorillas. If chimps could talk, they might tell us they believe in taking their sexual pleasures where they find them. Even if we limit ourselves to people, the validity of a specific research finding can also vary dramatically with populations. Presumably very few nuns believe in "taking their sexual pleasures where they find them," even if they are currently ovulating. Many research findings vary by population. Zajonc (1965) documented social facilitation in ants, cockroaches, parakeets, puppies, and monkeys as well as in people. Archival data sets rarely include multiple species, but they do very often include highly diverse human populations. This is a big strength.

Situations

A final aspect of external validity focuses on generalization across situations. All research takes place in a specific context, and that context can influence what researchers observe. The way people think and reason seems to vary based on the way in which an experimenter dresses. When experimenters dress casually people seem to *think* casually (i.e., intuitively; Simon et al., 1997). When people dress more formally, others are more likely to obey them (Bickman, 1974). It is a central tenet of social psychology that the specific situations in which people find themselves (e.g., a synagogue vs. a singles bar) can have a huge impact on how people think, feel, and behave. To know how robust a research finding is, you need to know how well it holds up in a wide variety of situations.

To summarize, to the degree that a specific archival study shows that an effect holds up well using different operational definitions, in multiple temporal windows, for different populations, and in different situations, there can be no doubt that the study has capitalized well on one of the strengths of archival research. But to return

to an earlier point, plenty of external validity in the absence of internal validity is not very informative. Are there any rules of thumb for assessing *internal* validity?

GAGES: Five Common Confounds That Can Undermine Archival Research

There is a reason why census takers, epidemiologists, and marketers have long focused on a handful of regional and demographic variables when doing their jobs. Five of the cardinal ways in which human beings vary include geography, age, gender, ethnicity, and socioeconomic standing (education and/or income). I refer to these five key variables using the acronym GAGES (geography, age, gender, ethnicity, and socioeconomics).

Geography

Geographically speaking, knowing where a person lives can be telling. From red states vs. blue states to latitude vs. altitude, location matters. According to the 2010 Census, the average Maryland resident had almost twice the family income of the average West Virginia resident. New Jersey is about 1,000 times more densely populated than Alaska. Personal beliefs and values also vary widely across U.S. states. Residents of Vermont are more than five times as likely as residents of Mississippi to report that they are not religious (Newport, 2014). In fact, research on cultural evolution suggests that properties of the physical environments in which people live predict variables as different as what kind of language people speak, whether people cook with spices, whether women are allowed to have multiple husbands, and whether people are xenophobic (Billing & Sherman, 1998; Everett, Blasib, & Roberts, 2015; Fincher, Thornhill, Murray, & Schaller, 2008).

Age

Demography can matter just as much as geography. Beginning with age, older Americans worry less than their younger counterparts (Newport & Pelham, 2009). They also eat healthier diets, exercise less frequently, and care more deeply than young people do about nurturing close, established relationships (Carstensen, Isaacowitz, & Charles, 1999; Dugan, 2013). Older Americans are also substantially more likely than their younger counterparts to be religious, and to be wholly unfamiliar with Lil Wayne. Differences such as these are why there is a field called developmental psychology.

Gender

Moving on to gender, across the globe, men are more likely than women to assault or kill others, to commit suicide, to work in dangerous jobs, and to abuse drugs.

Conversely, women are more likely than men to suffer from depression and to serve as caretakers, both at home and at work. The list of ways in which gender matters is so long that there is entire branch of research known as gender studies.

Ethnicity

Ethnicity matters, too. Both Blacks and Latinos are more likely than Whites to suffer from clinical depression (Dunlop, Song, Lyons, Manheim, & Chang, 2003). Relative to Whites, Blacks are also much more likely to lack confidence in the police (Jones, 2015), to vote Democratic, and to be familiar with Lil Wayne. More than 50 years after the passage of the U.S. Civil Rights Act, there are still large ethnic differences in income, unemployment, and education.

SES

Above and beyond ethnicity, one of the best predictors of longevity and well-being is socioeconomic standing (SES; Bosworth, Burtless, & Zhang, 2015). Socioeconomic standing also predicts important attitudes and values (Pelham, 2018; Pelham & Hetts, 1999) as well as serious problems such as suicide risk and automobile accident rates (Sehat, Naieni, Asadi-Lari, Foroushani, & Malek-Afzali, 2012). There is a reason many sociologists study SES. It matters.

Given the importance of the GAGES, researchers who conduct archival research will ideally be able to show that an archival research finding goes above and beyond any effects of the GAGES. Of course, the list of possible confounds about which researchers should worry does not end with GAGES. Specific confounds vary with the specific research question at hand. But the five major worries summarized by the GAGES are a great place to start. Because census takers, public health officials, marketers, and people seeking dates often care about GAGES, there is often good information about GAGES in archival data sets.

Is GAGES WEIRD? Henrich, Heine, and Norenzayan (2010) argued that a great deal of psychological research fails to consider the tremendous cultural diversity of the planet. Specifically they noted that the great majority of past research in psychology focused on WEIRD people, those who come from “Western, Educated, Industrialized, Rich, and Democratic” societies. GAGES is distinct from WEIRD. First, WEIRD expresses a concern about external validity (e.g., would shopkeepers in India behave like students in Indiana?). By contrast, GAGES is all about internal validity. That being said, GAGES does have some overlap with WEIRD. Cultures, after all, have geographies. Cultures also vary in age, ethnicity, SES, and even gender ratios. Furthermore, one could treat GAGES variables as cultural moderators rather than confounds. Conceptually, however, WEIRD overlaps more with OOPS than with GAGES. One key difference here is that WEIRD focuses on cultures whereas OOPS usually focuses on individual people. Further, WEIRD includes populations and situations but is largely silent

regarding operationalizations and occasions. In a sense, then, OOPS means that the WEIRD critique may not go quite far enough. A complete understanding of the strengths and weaknesses of archival research requires more than WEIRD insights.

Moderation and Theory

Researchers who rely on archival data rarely have the luxury of putting their preferred predictor in a footrace with all possible confounds. However, archival researchers do sometimes have access to theoretically derived moderator variables (including WEIRD ones) that ought to predict when an effect grows stronger versus weaker. For example, laboratory experiments on modeling (i.e., social learning) show that people are more likely than usual to imitate targets who resemble them (e.g., see Ariely, 2012; Bandura, 1977). Phillips's archival research on copycat violence capitalizes on exactly this logic. First, people do copy highly publicized suicides and homicides. Second, this copycat effect is stronger than usual when these tragedies get more media attention (Phillips & Carstensen, 1986). Third, consistent with principles discovered in lab experiments, people are more likely than usual to copy the suicidal behavior of others when they belong to the same gender, age, or ethnic group as a target.

The points made thus far suggest that it is possible to conduct archival research that strikes an impressive balance between internal and external validity. In the remainder of this chapter, I briefly summarize archival research on five different topics, ranging from social cognition to health and mortality. The thread that unites these archival studies is the fact that the authors all found creative ways to maximize both internal and external validity. Although I focus mainly on research topics I happen to study, I hope readers will realize that an appreciation of OOPS and GAGES could be applied effectively to almost any topic.

Archival Studies of Social Cognition

False Consensus

One of the first researchers to use archival data to study social cognition was Mullen (1983), who studied the false consensus effect (Ross, Greene, & House, 1977). This is the tendency for people to overestimate the percentage of others who share their beliefs or behaviors. Mullen believed this bias would still appear when avoiding it could help people win thousands of dollars. Mullen also suspected (correctly) that the false consensus effect is larger for people whose attitudes or behavior place them in the statistical minority rather than the majority. To study the false consensus effect, Mullen capitalized on data from a TV game show ("Play the Percentages"). The key data points provided by game show participants were their estimates of the percentage of studio audience members who would be

able to answer specific trivia questions (e.g., “What state did Hubert Humphrey represent in Congress?”) Back when people still remembered Humphrey, 72% of audience members were able to answer this question correctly.

Mullen observed clear evidence of the false consensus effect. Participants overestimated the percentage of others who knew the answers to questions when they themselves had known the answers to the questions. Second, false consensus effects were larger than usual when people’s own answers placed them in the statistical minority. The rare people who knew the answer to a difficult question were especially likely to assume that other people shared their esoteric knowledge.

Mullen’s documented a false consensus effect with a slightly different operational definition than the one usually used in the lab, with a novel population, and in a very different situation than the lab, satisfying three of the four OOPS criteria. Further, because this study included equal numbers of men and women, and because men and women both showed the effect, we cannot attribute Mullen’s effects to a gender confound. Finally, it is hard to imagine that any other GAGES confounds could apply to Mullen’s archival study without also applying to laboratory studies. More highly educated participants *may* have known more of the answers to the trivia questions, but there is no reason to believe that being educated *in and of itself* would make people offer higher *consensus* estimates—or that this would happen in games show but not in laboratories.

Ethnic Stereotyping

A more sobering example of archival research in social cognition is Eberhardt, Davies, Purdie-Vaughns, and Johnson (2006) research on stereotypes and capital punishment. Eberhardt and colleagues identified criminal records from more than 600 men who had been convicted of murder in greater Philadelphia between 1979 and 1999. They then identified all of the cases ($n = 44$) in which a Black defendant had been convicted of killing a White victim. Based on previous work by Blair and colleagues, they suspected that Black men convicted of killing White victims would be more likely to receive a death sentence when they had a more stereotypically Black appearance than when they did not. The researchers showed photographs of all of the selected Black defendants to students who knew nothing of the men’s criminal status. These judges assessed “the stereotypicality of each Black defendant’s appearance and were told they could use any number of features (e.g., lips, nose, hair texture, skin tone) to arrive at their judgments” (Eberhardt et al., 2006). Figure 8.2 shows two Black male volunteers who vary in stereotypicality.

One of the most methodologically impressive aspects of these archival findings is that Eberhardt and colleagues controlled for six potential confounds all known to be important predictors of sentencing decisions. These included “(a) aggravating circumstances, (b) mitigating circumstances, (c) severity of the murder (as determined by blind ratings of the cases once purged of racial information),

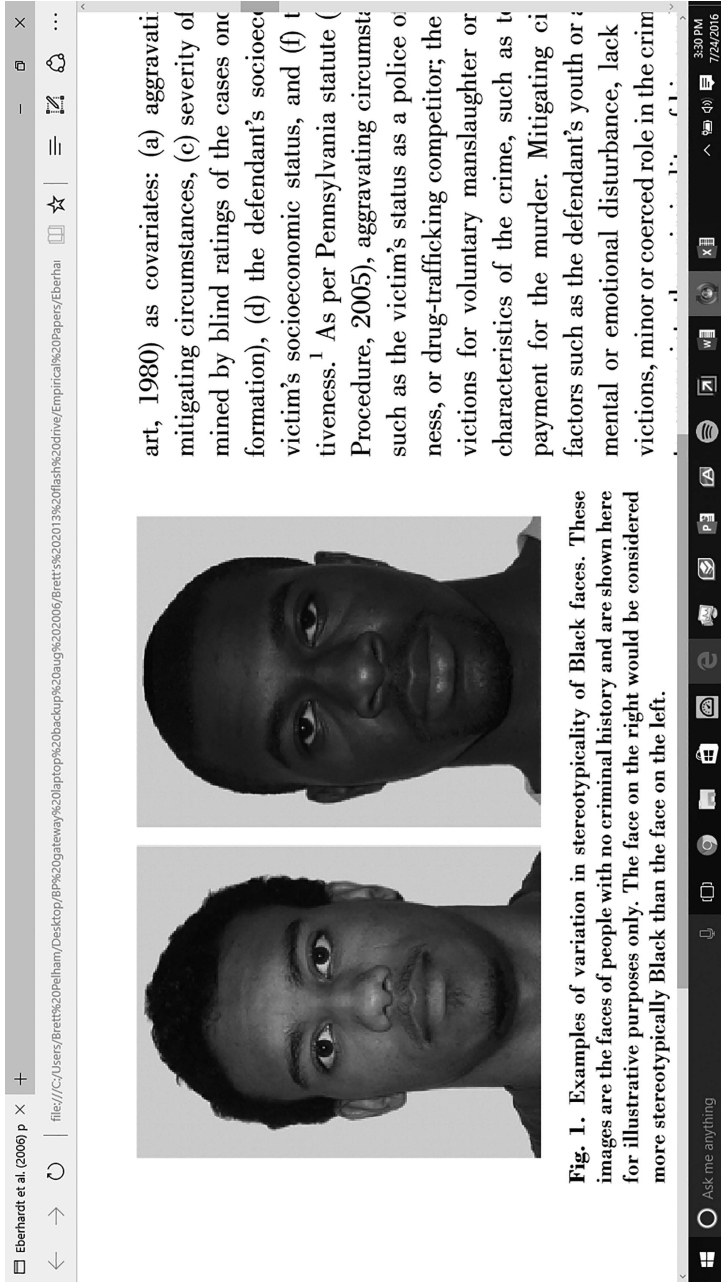


FIGURE 8.2 Two Black men with no criminal records who vary in the stereotypicality of their appearance.

Note: Eberhardt's findings suggest that if both men *were* to commit a crime, the man on the right would be judged more harshly.

(d) the defendant's socioeconomic status, (e) the victim's socioeconomic status, and (f) the defendant's attractiveness" (Eberhardt et al., 2006, p. 384). Further, their operational definitions of constructs a–e were based on established Pennsylvania statutes. Because the archival records did not include information on defendant physical attractiveness, the research team got blind raters to judge this. Even after controlling statistically for all six of these confounds, Eberhardt et al. (2006) found that Black men with a more stereotypically Black appearance were more likely than Black men with a less stereotypic appearance to be given a death sentence. A follow-up study showed that when defendants had been convicted of killing *Black* rather than *White* victims, the stereotypicality of the men's appearance no longer made a difference for their sentences.

A consideration of GAGES reveals that this study controlled for many geographic confounds by staying near Philadelphia. Of course, OOPS dictates that it would have been even better to study more than one region of the United States. But no single study can do everything. The authors also appear to have controlled for the defendants' ages because age is often considered a mitigating factor. They controlled for gender by studying only men. The authors not only controlled for ethnicity but also deconstructed it (ethnic stereotypicality was an independent variable). They also controlled for the SES of both the defendants and the victims. The additional factors for which Eberhardt and colleagues controlled reveals that GAGES is not an exhaustive list. But the fact that these authors left no GAGES stone unturned attests to the importance of these variables, as well as to the sophistication of this study.

Counterfactual Thinking and Emotions

Not all research in social cognition focuses on tragedies. Some of it focuses on triumphs. Medvec, Mavey, and Gilovich (1995) studied athletic triumphs, including triumphs that don't always make people feel very good. Laboratory research on counterfactual thinking shows that when something good or bad happens, people often consider counterfactual (alternative) realities. Counterfactual thoughts sometimes create counterintuitive emotions. For example, missing a flight by two hours usually produce regret. But missing a flight by two minutes usually produces a lot more of it (Roese, 1997). When Medvec and colleagues conducted their archival studies of counterfactual thinking and emotions, most previous studies had been conducted in the lab. Further, many of these studies were based hypothetical scenarios ("How would you feel if . . .?") rather than real outcomes. Medvec and colleagues put the factual into the study of counterfactuals.

They did so by considering the emotional implications of earning gold, silver, and bronze medals in major athletic competitions. Most Olympic gold medalists must surely be on top of the world after their victories. At a bare minimum they end up on top of the medal stand, and their gold medals often bring them fame

and fortune. By contrast, many silver medalists may feel the pain of knowing how close they came to winning. For bronze medalists, however, *two* things would have had to have gone differently for them to have won gold (e.g., Usain *and* Justin). The most salient counterfactual for bronze medalists is probably that they could have easily finished in fourth place, earning no medal at all. This logic suggests that athletes might typically be happier with an inferior outcome (a bronze medal) than with a superior one (a silver medal).

To test this prediction, Medvec et al. recorded NBC's televised coverage of the 1992 Olympics. They then extracted every scene that showed a bronze or silver medalist (in any sport NBC chose to cover) the moment the athletes learned they had finished second or third. They did the same thing for the period when athletes stood on the medal stand. Finally, they showed all of the video clips to a group of raters who were kept blind not only to Medvec et al.'s predictions but also to athletes' order of finish. They also turned the volume to zero for all of the ratings so that raters would not be biased by the comments of any of the NBC sports analysts, especially Bob Costas. The raters simply judged each athlete's expressed happiness on a 10-point scale.

Medvec and colleagues found that, despite finishing third rather than second, Olympic bronze medalists looked happier than their slightly faster, stronger, and more coordinated peers. This was true both immediately after their performances and on the medal stand. Of course, these results alone do not say whether *counterfactual thinking* was responsible for the observed emotions. To address this, Medvec et al. performed a second set of archival analyses from the same Olympic TV coverage. This time they selected all of the available *interviews* with bronze and silver medalists and asked blind raters to judge the "extent to which the athletes seemed preoccupied with thoughts of how they did perform versus how they almost performed." This follow-up study suggested that bronze medalists were more focused on what they "at least" did whereas silver medalists were focused on what they "almost did." A replication study focusing on a state-level athletic competition confirmed this result.

These results seem safe from any obvious GAGES confounds. It is highly unlikely, for example, that men (a) more often finish third than women and (b) are chronically happier than women. One not-so-obvious confound, however, is that in some Olympic events (e.g., wrestling, basketball), bronze medalists have just *won* a competition whereas silver medalists have just *lost* a competition. That's a real confound. In a supplemental analysis, Medvec et al. focused solely on events (e.g., track and field) in which there was no such confound. The bronze medalists still looked happier than the silver medalists. This archival research is also a standout when it comes to OOPS. It used novel operationalizations, it examined behavior in athletic events that took place on two different occasions, the participants came from all over the globe, and the situation in which people were studied was radically different than the lab. In my view the authors of this study struck methodological gold.

Self-Concept and Identity

Implicit Egotism: Early Studies of Career Choice

Archival research can also be a rich source of information about the self-concept. About 15 years ago, my colleagues and I became interested in the idea that people resemble the legendary Narcissus. Laboratory studies had already shown that people have an unconscious preference for the letters in their own names (Nuttin, 1985). Kitayama and Karasawa (1997) extended this to show that people prefer the numbers in their own birthdays (see also Beggan, 1992). So if something is part of the self, it must be good. Inspired by such findings we began to study implicit egotism, an unconscious preference for people, places, and things that resemble the self. We began by using archival data to study careers and street addresses.

In the early days of this work, we did not fully appreciate the risks inherent in archival research. When I began this work, for example, I knew that Carlos was a Latino first name. I also knew that there is a good chance Jeff Goldstein is Jewish. But I did not know, for example, that a person whose last name is Jefferson is about 180 times as likely to be Black as a person whose last name is Carlson. I eventually learned just how strongly people's first and last names can be confounded with GAGES.

This being said, my colleagues and I were not completely oblivious. In one of our first studies of implicit egotism we focused on career choice (Pelham, Mirenberg, & Jones, 2002). Our analyses of professional membership records for dentists and lawyers showed that people with names like Dennis, Denise, Lawrence, and Laura gravitated toward the jobs that resembled their names. We controlled to some degree for geography by limiting our searches (for both occupations) to the eight most populous U.S. states. We controlled for gender by separating male and female names. We further controlled for the frequency of our first names. After completing this initial study, however, we were disappointed to see how difficult it was to locate nationwide records that reliably identified people by name and occupation (but see Abel, 2010).

Implicit Egotism and Choice of a Residence

In contrast to the paucity of national data on names and professions, there are plenty of archival data that identify millions of people by name and place of residence. We have thus conducted numerous studies of implicit egotism and choice of a residence. Relying on 66 million Social Security Death Index records, Pelham et al. (2002) showed that just as people whose first or last name is Thomas are overrepresented in cities named St. Thomas, people whose first or last name is Peter are overrepresented in cities named St. Peter. Although we argued that our focus on cities with "Saint" in their names should have reduced ethnic confounds,

this study was still susceptible to geographic confounds. Although we conducted follow-up studies to address this concern, eliminating all of the GAGES concerns entirely is no easy matter. Some studies of implicit egotism are also open to critiques based on reversed causality (e.g., are there a lot of women in Georgia named Georgia because they moved there or because parents in Georgia prefer this first name?)

One way to address problems such as these is to blend archival methods with more traditional methods. Pelham and Carvallo (2015) did exactly this. First, using archival death records, we showed that men named Cal and Tex had disproportionately lived in the large U.S. states that resembled their names. In a replication study with living participants, we identified more than 800 men named Cal or Tex who lived in either California or Texas. A survey showed that these men, too, disproportionately lived in states closely resembling their names. Further, this was strongly true even when we focused exclusively on men who reported *moving* to the states in which they lived.

In addition to the usual problems of confounds and reverse causality research on implicit egotism is also susceptible to sampling problems. Although we always specified in advance how we had matched names with states, cities, or street addresses, we often faced arbitrary decisions. This problem largely evaporated on April 2, 2013. This is when the 130 million records that made up the entire 1940 U.S. Census became public (Pelham & Carvallo, 2015). At about this same time, the genealogical website ancestry.com also made available a tremendous number of American and international birth and marriage records.

Back to Career Choice

The 1940 U.S. Census data allowed us to see if people preferred occupations that doubled as their exact surnames (Pelham & Carvallo, 2015). To test this idea, we examined the 2,000 most common U.S. surnames. From this list, we identified the 11 surnames that constituted common, traditionally male occupations (e.g., Baker, Carpenter, Farmer, Mason). We then calculated the expected number of men with these exact surnames who *would* have reported working in each of these 11 occupations if there were no association between surname and career choice. We compared each expected frequency with the observed frequency that corresponded to a surname–occupation match. As shown in Figure 8.3, to at least a small degree, all 11 of the surname–occupation pairs yielded support for implicit egotism (as indicated by ratios greater than 1.0).

The score of 1.34 for Porter means that men name Porter were 34% more likely to have worked as porters compared with the entire set of men who had the other 10 surnames. But might these results reflect an ethnic confound? If Black men disproportionately worked as porters, and if Black men were disproportionately *named* Porter, this could create artifactual support for implicit egotism. A similar argument could be made for other GAGES.

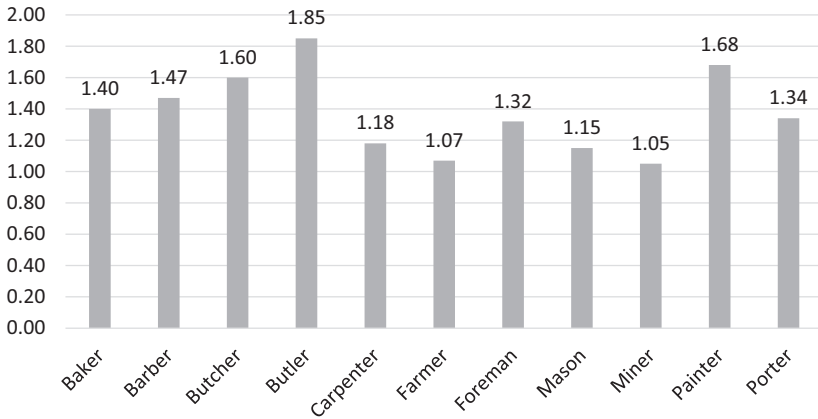


FIGURE 8.3 Ratio of Observed to Expected Surname-Occupation Matches for Men With Common Surnames That Also Serve as Male Occupation Names

Source: 1940 Census.

Because of the sheer size of these census data, we were able to evaluate these results separately for Blacks and Whites, and for men who were equated perfectly for education. Even after we separated men by ethnicity and focused separately on men with exactly 6, 8, 10, or 12 years of education, there was always robust support for implicit egotism. For example, White men named Farmer who had an eighth grade education gravitated toward farming relative to all other White men in the 1940 Census who were not named Farmer but who also had an eighth grade education.

We did not publish this finding, but we also found that implicit egotism in career choice held up for men of different ages. This leaves only geography. If lots of people named Farmer happen to live in Nebraska as compared with Massachusetts, this could lead to artifactual support for implicit egotism. Although we did not consider this geographic confound in our original report, I recently conducted state-by-state analyses to see if the association between surname and occupational choice would hold up at this level. It did. Further evidence that this effect is not likely due to a geographic or ethnic confound is based on the fact that the effect replicated strongly—using exactly the same 11 career-surname pairs—in both the 1880 U.S. Census and the 1911 England Census. Taken together, these three studies—which span a 60-year window—also do a decent job of addressing the OOPS issue of occasions.

Implicit Egotism and Marriage

Perhaps the strongest archival evidence for implicit egotism comes from research that focuses on birthdays rather than names. Most people prefer their birthday

numbers. Unlike a person's first or last name, however, which can easily be confounded with GAGES, a person's birth month or birthday number is much more arbitrary. Birthday numbers also range from 1–31, which greatly simplifies decisions about how to sample specific numbers. If possible, sample them all. If people gravitated toward things that matched their birthday numbers, this would constitute rigorous support for implicit egotism.

People do, and the thing toward which they gravitate is other people. Pelham and Carvallo (2015) found that people were disproportionately likely to marry other people who happened to share either their birthday number or their birth month. As shown in the top panel of Figure 8.4, brides in Summit County, Ohio, were 6.5% more likely than they should have been by chance to marry a groom who shared their exact birthday number. As shown in the bottom panel of Figure 8.4, this bias increased to 37.6% for the subset of brides who *married* on their birthday numbers. Both of these effects, as well as conceptually identical effects for birth *month*, replicated well in a very large set of more recent (1958–2001) Minnesota marriage records.

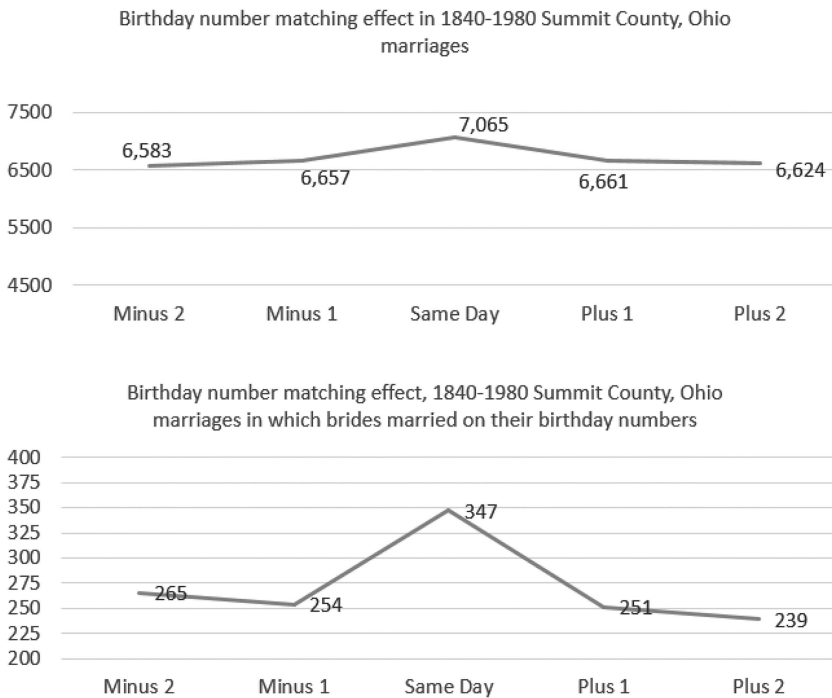


FIGURE 8.4 Implicit Egotism in Marriage

Note: People with the same birthday number were more likely than usual to marry. This effect was much stronger than usual for brides who married on their birthday number. Adapted from Pelham and Carvallo (2015).

I hasten to add that implicit egotism is not the only aspect of the self that has been studied using archival data. Research by Twenge and colleagues—most of it archival—has shown that self-esteem (“I am capable.”) and narcissism (“I will build a great, great wall.”) have been increasing in the United States over the past few decades. Twenge typically examines archival data *collected by psychologists*, which has allowed her to document changes in constructs that would be hard to operationalize using birth or marriage records. That being said, Dewall, Pond, Campbell, and Twenge (2011) did examine changes in egotism over the past few decades without asking anyone to fill out any surveys. They did so by tabulating changes in the relative use of the words “I,” “me,” and “mine” in popular song lyrics between 1980 and 2007. Across this window there was a linear increase in how often popular singers sang about themselves (via these pronouns). If Twenge and colleagues are correct that U.S. culture is becoming more narcissistic over time, this bodes well for the future of research on implicit egotism.

Longevity and Death

In addition to archival research on social cognition and the self-concept, there is also archival research on longevity and dying. This research focuses on how psychological variables influence the age at which people join all the others who have gone before them—to that hallowed place I call archival death records. This archival research reveals that psychological variables have more to do with longevity and dying than one might think. Levity seems to promote longevity.

The Nuns Study

To see if viewing the world favorably helps people live longer, Danner, Snowdon, and Friesen (2001) studied 180 nuns. Around 1930, when these women became nuns, they had to write a brief autobiography—which the church dutifully retained. In the late 1990s, Danner and colleagues got permission to analyze these records. The research team blindly coded each young nun’s life story for how many positive emotions it included. It’s pretty easy to spot happiness in these essays. Sister 1 wrote things like, “With God’s grace, I intend to do my best . . .” Sister 2 wrote things like, “I look forward with eager joy to . . . a life of union with Love Divine.”

Based on a simple count of the number of positive emotion words the nuns used, those in the top quartile of the positive emotion distribution lived 9.4 years longer than those in the bottom quartile! Because the nuns lived in one of only two cities (and because the researchers took city differences into account) we do not have to worry much about a geographical confound. The research team also controlled for the age at which women wrote their essays and their eventual education. It also seems safe to assume that the lifestyle most nuns lead went a

long way toward holding many other important variables constant. I'm guessing that few, if any, of the nuns died because of a lack of food or basic medical care. Because this was a sample of nuns, gender was held constant. The research provided no information on ethnicity, but I am pretty confident that the large majority of them were White (and Catholic, too). In short, this study seems to have controlled well for the GAGES confounds. This is particularly important in a study of longevity because each of the GAGES variables has a well-documented association with longevity.

The Baseball Players Study

Any nun with a decent sense of humor will tell you that she tries to fulfill her calling in life by making sacrifices and saving souls—all while having to wear a funny outfit. Baseball players also care deeply about saving and sacrificing, and they, too, wear funny outfits. Like nuns, baseball players also appear to live longer when they express more positive emotions. Baseball players don't have to write autobiographies to play in the majors, but they do have to get their pictures taken. Abel and Kruger (2010) took advantage of this fact by conducting an archival study of professional baseball players whose photos appeared in the 1952 *Baseball Register*. They blindly rated all 230 of the official player photos for how happy the players looked. Photos in which the men were not smiling were given the lowest score, photos in which the men were smiling a polite but unnatural smile got a middle score, and those in which the men expressed a truly happy ("Duchenne") smile got the highest score.

Abel and Kruger reasoned that these photos would reflect a player's characteristic emotional state. Apparently, they did. By the time, Abel and Kruger wrote their report in 2009, the very large majority of these men had died, and Abel and Kruger extracted these dates of death from archival sources. The men who showed a true smile lived an average of seven years longer (mean age 79.9) than the men who did not smile at all (age 72.9). The men who smiled politely (mean age 75.0) also lived a bit longer than the non-smilers. Impressively, these differences in longevity held up even after controlling for a pretty hefty list of competing predictors of longevity (e.g., body mass index, education, marital status, length of playing career).

The fact that Abel and Kruger used player photos from 1952 means that they came pretty close to controlling for ethnicity. In 1952, 94.4% of professional baseball players were White—with the remaining 5.6% being split pretty evenly between Black and Latino players. Blacks do not live as long as Whites, and Latinos actually live a bit *longer* than Whites. For ethnicity to be a serious confound, then, almost all the Blacks would have to be looking solemn and almost all of the Latinos would have to be smiling very happily. It is unlikely—though not impossible—that an ethnic confound could be responsible for their findings.

Further, if a critic were still worried about this confound, it would be easy to code the 230 photos for ethnicity and add this to a covariance analysis. Finally, just as the nun study controlled for gender by studying only women, this study controlled for gender by studying only men.

In light of the GAGES criteria, Abel and Kruger seem to have knocked this one out of the park. They did not address geography, but any critic who was worried about it could add geographical codes to the data and control for them. If there is any real weakness to this study, it is the small number of (deceased) players with usable data ($n = 23$) who posed with Duchenne smiles (samples sizes were greater than $n = 60$ for the other groups). In contrast, the nun study relied on a continuous rather than a categorical coding scheme for positive emotionality. Taken together, these two archival studies are worth smiling about.

A Distressing Natural Experiment

As strong as these two archival studies are, one could still argue that some unknown confound played a role in the results. Perhaps healthier baseball players smiled more. Perhaps nuns from wealthier family backgrounds used more positive emotions words. To know with greater certainty that psychological variables can influence well-being or longevity, one would have to use random assignment to decide that one group of people experienced a dramatic event that others just like them did not. Ideally this ethically callous experiment would have a huge sample size. Further, the design might be particularly powerful if the experimental manipulation involved some kind of horrific negative experience. Being forced to fight in what many considered an unjust war might fit the methodological bill. During the Vietnam War, the U.S. government created exactly this situation as a natural experiment. As Hearst, Newman, and Hulley (1986) noted, the Vietnam War draft meant randomly assigning a subset of young American men to fight in Vietnam. This draft thus created a huge natural experiment.

Hearst and colleagues identified 14,000 men born in California or Pennsylvania between 1950 and 1952 whose draft numbers did versus did not come up. They then checked to see what happened to the men in the 10-year window after the war (1974–1983). They found, among other things, that suicide rates for the unlucky men whose draft numbers had come were 13% higher than they were for the lucky men whose draft numbers had *not* come up. This sounds like a modest effect of this natural manipulation. In fact, the effect of actually serving in combat was certainly much larger. This is because fully 74% of the drafted men found a way to *avoid* going to war. Further, a patriotic 9% of those who had *not* been drafted enlisted as volunteers. After correcting statistically for these facts, the estimated impact of serving as a soldier in Vietnam was an 86% increase in the risk of suicide. Some archival studies come extremely close to the gold standard of true experiments. In natural experiments, the GAGES, like all other confounding variables, are held constant.

Holidays and Mortality

Archival research also suggests that psychological variables can determine the exact day on which people die. Phillips and colleagues conducted several archival studies that suggested that people can sometimes postpone their own deaths to experience important events. Phillips and Feldman (1973) focused on religious holidays. They found that in cities with many Jewish residents (i.e., New York City and Budapest, Hungary), people were more likely to die shortly after than shortly before Yom Kippur (the Jewish Day of Atonement). I think it's safe to say that Phillips' work was met with great skepticism.

To address the skeptics, Shimizu and Pelham (2008) harvested data from the same Social Security Death Index discussed elsewhere in this chapter. We reasoned that the two major holidays on which most Americans would least want to die would be Thanksgiving and Christmas. On these two major holidays many Americans go to great lengths to be with those they love. We thus reasoned that people might be less likely to die on the exact day of either of these holidays than on any of the immediately surrounding days. If people who are gravely ill try to hold on to make it to a major holiday, we also reasoned that death rates should be a little higher for the two days immediately *after* these two holidays than for the two days immediately before them. For both holidays we focused on the 16-year window between 1987 and 2002. Later data were not yet available, and earlier data did not usually include an exact date of death. Our results were very similar for Thanksgiving and Christmas, with the effect for Christmas being predictably larger.

As shown in Figure 8.5 people did defer their deaths until after Christmas Day. In this 16-year window about 4,000 more Americans died on December 27 than on December 25. Because these data provided no demographic information we

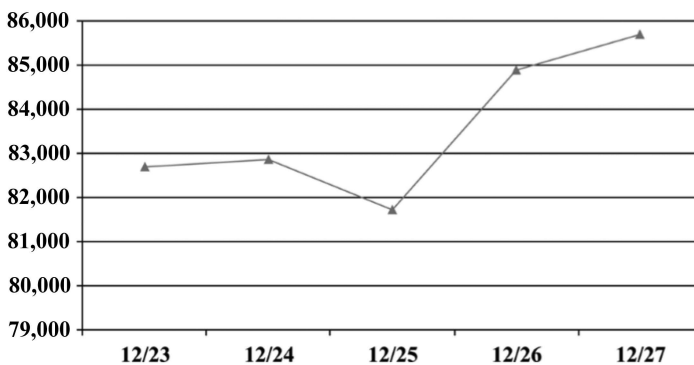


FIGURE 8.5 Number of Americans Dying on Christmas Day versus on any of the Four Neighboring Days

Source: SSDI, 1987–2002. Figure is adapted by permission from Shimizu and Pelham (2008).

were not able to control for GAGES. However, it is worth noting that the exact day on which Christmas occurs varies arbitrarily. Further, poor weather is no more likely to occur on Christmas Day than on any of the days surrounding it. Christmas also comes to women as well as to men. One worrisome confound that applies to both Thanksgiving and Christmas, however, is that people are probably much less likely to take long trips on the exact days of these holidays than on the surrounding days. In a series of supplemental analyses using California mortality data we were able to address this confound. The effects all remained when we removed all of the deaths due to accident.

We were also able to identify two potential moderators of the Christmas Day death-deferral effect. First, using Phillips and King's (1988) list of 106 actuarially Jewish surnames (e.g., Goldberg, Silverstein), we focused on Americans who were very likely to be Jewish. The pattern shown in Figure 8.5 almost completely disappeared in the Jewish surname sample. In contrast to Jews, most American (non-Jewish) children are absolutely enamored of Christmas. If the Christmas death deferral effect is grounded in a desire to experience Christmas, children should show a larger than usual Christmas death deferral effect. Although the SSDI records do not include age, it was possible to derive ages from dates of birth and death. The magnitude of the death deferral effect for children (aged 5–20) was 10 times as large as the effect observed for adults. At least some people appear to exercise at least some control over the exact timing of their own deaths.

Conclusions

Archival data are a rich source of information about human behavior. From birth to death, people leave behind a great deal of evidence of their daily behavior. As the technical barriers that restrict access to archival data become smaller, the potential for new archival discoveries keep growing. In the past year alone, I have explored archival data from the World Values Survey, DHS Surveys, the CIA Factbook, the UNDP data site, Google Correlate, and dozens of birth, census, marriage, divorce, and death records. The possibilities of using archival data extend well beyond social cognition and health psychology. Although I have focused here on three topics that interest me, archival research methods have also shed light on many other topics, from climate change to ethnic biases in birth outcomes.

I hope that as researchers seek new opportunities to conduct archival research they will focus on the crucial role of theory in scientific discovery. The best theory testing often involves testing for statistical moderation, and archival data sets are very well-suited to this—because they often involve huge and diverse samples. Archival data sources also provide opportunities for replication that are almost impossible to match with other data sets. Finally, for researchers who have difficulty with delay of gratification, archival studies can often provide rapid answers to pressing questions. Some archival studies can be done in hours if not minutes. Regardless of how long archival research takes, it can tell us not just whether a

theory holds up outside the lab but also for whom, in what way, when, where, and sometimes even why.

References

- Abel, E. L. (2010). Influence of names on career choices in medicine. *Names: A Journal of Onomastics*, 58, 65–74. doi:10.1179/002777310X12682237914945
- Abel, E. L., & Kruger, M. L. (2010). Smile intensity in photographs predicts longevity. *Psychological Science*, 21, 542–544.
- Adams, H. E., Wright, L. W., & Lohr, B. A. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology*, 105, 440–445.
- Ariely, D. (2012). *The (honest) truth about dishonesty: How we lie to everyone—especially ourselves*. New York, NY: HarperCollins.
- Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (pp. 1–78). Reading, MA: Addison-Wesley.
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.
- Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2, 412–414.
- Beggan, J. K. (1992). On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology*, 62, 229–237. doi:10.1037/0022-3514.62.2.229
- Bickman, I. (1974). The social power of a uniform. *Journal of Applied Social Psychology*, 4, 47–61.
- Billing, J., & Sherman, P. W. (1998). Antimicrobial functions of spices: Why some like it hot. *The Quarterly Review of Biology*, 73, 3–49.
- Booth, A., & Amato, P. (1991). Divorce and psychological stress. *Journal of Health and Social Behavior*, 32(4), 396–407. Retrieved from <http://www.jstor.org/stable/2137106>
- Bosworth, B., Burtless, G., & Zhang, K. (2015). *Later retirement, inequality in old age, and the growing gap in longevity between rich and poor*. White paper, full text. Retrieved from www.brookings.edu
- Brick, T. R., & Boker, S. M. (2011). Correlational methods for analysis of dance movements. *Dance Research, Special Electronic Issue: Dance and Neuroscience: New Partnerships*, 29, 283–304.
- Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist*, 54, 165–181.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Danner, D. D., Snowdon, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: Findings from the nun study. *Journal of Personality and Social Psychology*, 80, 804–813.
- Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.
- DeWall, C. N., Pond, R. S., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of self-focus, loneliness, anger, anti-social behavior, and misery increase over time in popular U.S. song lyrics. *Psychology of Aesthetics, Art, and Creativity*, 5, 200–207.
- Dugan, A. (2013). *Fast food still major part of U.S. diet*. Retrieved from www.gallup.com/poll/163868/fast-food-major-part-diet.aspx
- Dunlop, D. D., Song, J., Lyons, J. S., Manheim, L. M., & Chang, R. C. (2003). Racial/ethnic differences in rates of depression among preretirement adults. *American Journal of Public Health*, 93, 1945–1952.

- Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. J. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, 17, 383–386.
- Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, 32, 172–178.
- Everett, C., Blasib, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *PNAS*, 112, 1322–1327.
- Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society*, 275, 1279–1285.
- Gangestad, S. W., Thornhill, R., & Garver-Apgara, C. E. (2010). Fertility in the cycle predicts women's interest in sexual opportunism. *Evolution and Human Behavior*, 31, 400–411. doi:10.1016/j.evolhumbehav.2010.05.003
- Hamilton, W. D. (1964a). The genetical evolution of social behavior I. *Journal of Theoretical Biology*, 7, 1–16.
- Hamilton, W. D. (1964b). The genetical evolution of social behavior II. *Journal of Theoretical Biology*, 7, 17–52.
- Hearst, N., Newman, T. B., & Hulley, S. B. (1986). Delayed-effects of the military draft on mortality—A randomized natural experiment. *New England Journal of Medicine*, 314, 620–624.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–135.
- Jones, J. M. (2015). In U.S., confidence in police lowest in 22 years. Retrieved from www.gallup.com/poll/183704/confidence-police-lowest-years.aspx
- Kitayama, S., & Karasawa, M. (1997). Implicit self-esteem in Japan: Name letters and birthday numbers. *Personality and Social Psychology Bulletin*, 23, 736–742. doi:10.1177/0146167297237006
- Kukkonen, T., Binik, Y., Amsel, R., & Carrier, S. (2010). An evaluation of the validity of thermography as a physiological measure of sexual arousal in a nonuniversity adult sample. *Archives of Sexual Behavior*, 39, 861–873.
- Larrick, R. P., Timmerman, T. A., Carton, A. M., & Abrevaya, J. (2014). Temper, temperature, and temptation: Heat-related retaliation in baseball. *Psychological Science*, 22, 423–428.
- McGuire, W. J. (1989). A perspectivist approach to the strategic planning of programmatic scientific research. In B. Gholson, W. R. Shadish, Jr., R. A. Neimeyer, & A. C. Houts (Eds.), *Psychology of science: Contributions to metascience* (pp. 214–245). Cambridge: Cambridge University Press.
- Medvec, V. H., Madey, S. F., & Gilovich, T. (1995). When less is more: Counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality and Social Psychology*, 69, 603–610.
- Mill, J. S. (2002/1863). *A system of logic*. Honolulu, HI: University Press of the Pacific.
- Mullen, B. (1983). Egocentric bias in estimates of consensus. *The Journal of Social Psychology*, 121, 31–38.
- Murdock, G. P., & White, D. R. (2006). *Standard cross-cultural sample: On-line edition*. Working Papers Series, Permalink. Retrieved from <http://escholarship.org/uc/item/62c5c02n>
- Newport, F. (2014). *Mississippi most religious state, Vermont least religious*. Retrieved from www.gallup.com/poll/167267/mississippi-religious-vermont-least-religious-state.aspx

- Newport, F., & Pelham, B. W. (2009). *Don't worry, Be 80: Worry and stress decline with age*. Retrieved from www.gallup.com/poll/124655/dont-worry-be-80-worry-stress-decline-age.aspx
- Nuttin, J. M., Jr. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology*, 15, 353–361. doi:10.1002/ejsp.2420150309
- Pelham, B. W. (2018). *Evolutionary psychology: Genes, environments, and time*. London, UK: Palgrave.
- Pelham, B. W., & Blanton, H. (2013). *Conducting research in psychology: Measuring the weight of smoke* (4th ed.). Pacific Grove, CA: Cengage Publishing.
- Pelham, B. W., & Carvallo, M. R. (2015). When Tex and Tess Carpenter build houses in Texas: Moderators of implicit egotism. *Self and Identity*, 14, 692–723.
- Pelham, B. W., Mirenberg, M. C., & Jones, J. T. (2002). Why Susie sells seashells by the seashore: Implicit egotism and major life decisions. *Journal of Personality and Social Psychology*, 82, 469–487. doi:10.1037/0022-3514.82.4.469
- Phillips, D. P., & Carstensen, L. L. (1986). Clustering of teenage suicides after television news stories about suicide. *New England Journal of Medicine*, 315, 685–689.
- Phillips, D. P., & Feldman, K. A. (1973). A dip in deaths before ceremonial occasions: Some new relationships between social integration and mortality. *American Sociological Review*, 38, 678–696.
- Phillips, D. P., & King, E. W. (1988). Death takes a holiday: Mortality surrounding major social occasions. *Lancet*, 2, 728–732.
- Pinker, S. (2010). *Better angels of our nature: Why violence has declined*. New York: Penguin Books.
- Prior, H., Schwarz, A., & Gunturkun, O. (2008). Mirror-induced behavior in the magpie (*Pica pica*): Evidence of self-recognition. *PLoS Biology*, 6, e202.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Refinetti, R. (2005). Time for sex: Nycthemeral distribution of human sexual behavior. *Journal of Circadian Rhythms*, 3. Retrieved from <http://doi.org/10.1186/1740-3391-3-4>
- Reifman, A. S., Larrick, R. P., & Fein, S. (1991). Temper and temperature on the diamond: The heat-aggression relationship in Major League Baseball. *Personality and Social Psychology Bulletin*, 17, 580–585.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121, 133–148.
- Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13, 279–301.
- Rozar, T. L. (2012, May 23). *Impact of seasonality on mortality*. Society of Actuaries Conference. Retrieved from www.rgare.com/knowledgecenter/Pages/MortalitySeasonalityPresentation.aspx
- Scheidel, W. (2009). *Disease and death in the ancient city of Rome Version 2.0*. Working Paper. Retrieved from www.princeton.edu/~pswpc/pdfs/scheidel/040901.pdf
- Sehat, M., Naieni, K. H., Asadi-Lari, M., Foroushani, A. R., & Malek-Afzali, H. (2012). Socioeconomic status and incidence of traffic accidents in metropolitan Tehran: A population-based study. *International Journal of Preventive Medicine*, 3, 181–190.
- Shimizu, M., & Pelham, B. W. (2008). Postponing a date with the grim reaper: Ceremonial events, the will to live, and mortality. *Basic and Applied Social Psychology*, 30, 36–45.

- Simon, L., Greenberg, J., Harmon-Jones, E., Solomon, S., Pyszczynski, T., Arndt, J., & Abend, T. (1997). Terror management and cognitive experiential self-theory: Evidence that terror management occurs in the experiential system. *Journal of Personality & Social Psychology*, 72, 1132–1146.
- Sistiaga, A., Mallol, C., Galván, B., & Summons, R. E. (2014). The Neanderthal meal: A new perspective using faecal biomarkers. *PLoS ONE*, 9, e101045.
- Smith, M. S., Kish, B. L., & Crawford, C. B. (1987). Inheritance of wealth as human kin investment. *Ethology and Sociobiology*, 8, 171–182.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54.
- Welling, L. L. M., Jones, B. C., DeBruine, L. M., Smith, F. G., Feinberg, D. R., Little, A. C., & Al-Dujaili, E. A. S. (2008). Men report stronger attraction to femininity in women's faces when their testosterone levels are high. *Hormones and Behavior*, 54, 703–708.
- Willer, R., Sharkey, A., & Frey, S. (2012). Reciprocity on the hardwood: Passing patterns among professional basketball players. *PLoS ONE*, 7(12), e49807.
- Wißing, C., Rougier, H., Crevecoeur, I., Germonpré, M., Naito, Y. I., Semal, P., & Boche-rens, H. (2015). Isotopic evidence for dietary ecology of late Neandertals in North-Western Europe. *Quaternary International*. doi:10.1016/j.quaint.2015.09.091
- Zajonc, R. B. (1965). Social facilitation. *Science*, 149, 269–274.

9

GEOCODING

Using Space to Enhance Social Psychological Research

Natasza Marrouch and Blair T. Johnson

Acknowledgements: The preparation of this chapter was supported by University of Connecticut grant 6329500. We thank Hart Blanton, Emily Alden Hennessy, Stephen Herzog, and Tania B. Huedo-Medina for their helpful and insightful comments on previous drafts of this chapter.

Introduction

Overview

This chapter provides an orientation to the roles spatial information can—and often should—play in scientific investigations of social and personality phenomena. We argue that spatial and temporal data are ideal sources of information for aiding researchers due to their accessibility and the proven effectiveness of spatiotemporal analyses across multiple scientific domains. To give one example, if a phenomenon exhibits spatial or spatiotemporal similarities (or consistent differences—also referred to as covariance or autocorrelation), the researcher must reject the assumption of the independence of observations. But, beyond merely listing examples that support our perspective, our hope is to provide the reader with an introductory toolkit to aid their scientific pursuits. This toolkit will be of particular relevance when theoretical insights lead researchers to believe that macro, spatially embedded factors have important implications for their study. We present the tools from this kit in the form of programming code—along with links to download data—both of which will allow the reader to replicate and expand on the examples herein, while at the same time helping us to illustrate some of the more useful spatial statistics and mapping techniques.

This chapter proceeds in three sections. We begin by providing descriptions of spatial data and variables, with a focus on their intellectual history and the role that theory now plays in current research applications. We then turn to some of the more technical issues, providing readers with a “gentle” introduction to multilevel modeling in the hopes that we might reveal how geospatial applications represent straightforward extensions of more traditional and possibly more familiar research methods. We then wade into some of the thornier statistical issues that one must consider when analyzing geospatially coded data, with an eye towards providing a set of resources to readers to help them launch research programs that include geospatial components. We close by reviewing some of the potential benefits of spatiotemporal approaches as tools that might contribute to our understanding of a wide range of phenomena in social psychology.

Brief History of Spatiotemporal Variables

Once we recognize that the Earth is a sphere, we need three dimensions to describe it. Enter the science of geography, and with it ever more precise methods to characterize space: The ancient Greeks were the first to use the concepts of *latitude* (location on the north vs. south, or vertical plane) and *longitude* (location on the west vs. east, or horizontal plane). The modern Cartesian coordinate system places 0° latitude at the Earth’s Equator. Placing the Prime Meridian at 0° longitude was a matter of debate—because its location is largely arbitrary. Eventually, in 1851, the English mathematician and astronomer Sir George Biddell Airy won out, assigning the prime (or first) meridian to his home of Greenwich, England. From there, if you follow the Prime Meridian south to the Equator, you reach the point on the Earth’s surface where latitude and longitude are both exactly zero. At their heart, latitude and longitude gauge distances from this point in the Atlantic Ocean, located south of modern Ghana.

Longitude and latitude are typically expressed in degrees, to capture different angles. Longitude refers to the angle between a location and the Prime Meridian, ranging from -180° (west of the Prime Meridian) to +180° (east). Latitude describes the angle between a location and the plane of the Equator, ranging from -90° (south) to +90° (north). The same information can be, and often is, measured using radians or degrees-minutes-seconds (DMS)¹ in lieu of degrees alone. We mention the different units to sensitize the reader to the need to use a consistent unit when comparing multiple spatial data. That said, even if the reader accidentally uses dissimilar units, it is unlikely that they can cause a disaster of similar magnitude to the loss of the Mars Orbiter in 1999, which occurred because its components used inconsistent units (Stephenson, 1999).

Further, due to the complex nature of the Earth’s shape, longitude and latitude coordinates require additional details that place the “ideal” sphere coordinates onto the irregular (and ecliptic) surface of the Earth. Figure 9.1 is a simulated illustration of a sphere with an irregular surface. For this purpose, geospatial data

Wrinkled Sphere



FIGURE 9.1 A Sphere With Surface Irregularities

are linked to specific coordinate reference systems, which provide standardized descriptions. These systems are of special relevance for researchers who may need to combine information from spatial datasets with, for example, varying grid sizes (Figure 9.2). To match the data, the researcher will need to collapse information from one dataset into larger areas (grids) before merging the files.

Most readers have likely encountered similar complications already. For example, when looking at the age of participants using data from different sources: In one dataset, the variable is captured as years since birth, while in another, age is a four-level categorical variable. To meaningfully combine these datasets, we need to transform the former, more detailed, variable into corresponding levels of the latter categorical variable. Among the more common coordinate reference systems are: World Geodetic System—WGS 84 (Department of Defense, 2000), used by the Global Positioning System (GPS); the International Terrestrial Reference Frame (ITRF, 2016); and the European Geodetic Reference System (EUREF, 2016).

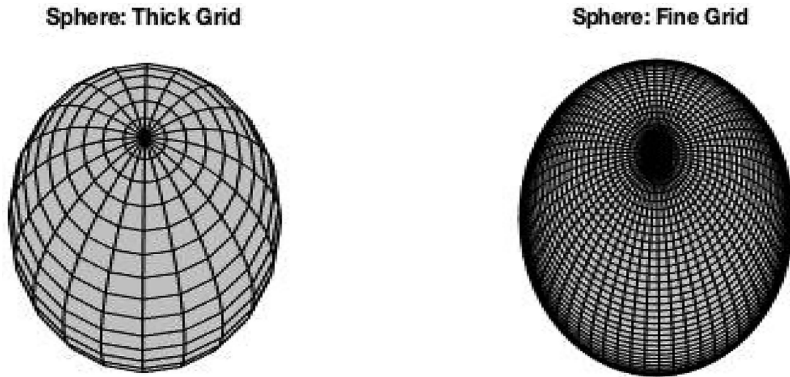


FIGURE 9.2 Examples of Different Size Grids on the Same Sphere

The Importance of Theory

The grid of the geospatial unit and the size of examined space should be driven by theory-based questions. Choosing the appropriate area and dividing it into equal-sized grids is of critical importance. On one hand, an incorrect grid size area may obscure potential relationships, but on the other, it may produce artifacts. Smaller units offer higher resolution and enhanced precision in differentiating specific elements of examined space, but they also generate noisier data due to smaller samples within spatial units (fewer participants within the space). In essence, “precision in the spatial sense does not equate with precision in the statistical sense” (Haining, 2003, p. 68). That said, overly large units will fail to include varying levels of spatially embedded social and environmental factors.

A good example of the importance of the appropriate hypothesis- or theory-driven spatial unit size can be found in developmental and social psychology. For instance, consider tests of the “universality” of a personality theory looking at a finite space in a single country. The researcher will be more likely to inaccurately confirm theoretical “universality” in the single-country design than by using units of the same size within a larger, more heterogeneous space encompassing different countries and cultures. This example is not foreign even to the most prominent psychological theories. For example, Piaget’s (1957) theory of cognitive development assumed in its earliest phases that developmental stages appear in a specific order; thus, reaching a sequential developmental stage is possible only after achievement of the preceding stage. The theory was initially developed through observations of Swiss children and later refined in studies of European and North American children. Geospatially, the size of the space of the studied phenomena soon became large and incorporative of multiple countries. Yet, the study’s limited sampling from only western cultures faced challenges due to insufficient heterogeneity in the study environment. Joint research efforts over

the past few decades have facilitated a more precise model, improving upon this theory by expanding the geographic focus and examining non-western children. It allowed for a more accurate description of cognitive development in children and the eventual realization that the age and order of developmental stages depends heavily on environmental factors and social structures (Dasen, 1994; Maynard, 2008).

Manipulations and measures are *figural* aspects to any investigation in social psychology, but the emphasis on the environment where the study took place, the *ground*, is much rarer. Even those who undertake systematic reviews and meta-analyses of phenomena brush over such factors, as Johnson, Cromley, and Marrouch (2017) documented, despite the fact that such reviews endeavor to take a big picture of a domain of research. Often the reason behind omission of spatial aspects is prosaic, the lack of sufficient details to account for spatial locations. In social psychology and other fields, countless studies include words to this effect: The study took place in a large Midwestern city. But, there are many such cities, why not be more specific? It could be that the cultural milieu in Chicago, Milwaukee, Indianapolis, St. Louis, Cincinnati, St. Paul, and Minneapolis, among others, is equivalent, but even more likely, it is dramatically different.

A Gentle Introduction to Multilevel Models

In theory, changes over time and space—and among the research team conducting the study—should not affect the study's results, or effect sizes, beyond that expected due to sampling error. When such inconsistencies in results do occur, they are often worrisome (e.g., Open Science Collaboration, 2015). And while these concerns may be legitimate, a progressive science requires that we view inconsistencies for what they are: venues to discover additional relationships and tap potential hidden moderators to better understand complexities of the social psychological phenomena in question (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016).

A number of meta-analyses show that, indeed, effect sizes vary over space and/or time (Lee, Pratto, & Johnson, 2011; Twenge, 2000), highlighting the importance of expanding our models to account for these variables. Some meta-analyses go a step further. Hsiang and colleagues (Hsiang, Burke, & Miguel, 2013) combined studies looking at bursts of intergroup and interpersonal violence, applying spatiotemporal data to quantify the degree of climate change at different points of time in various countries. They successfully showed that climate change explained a statistically significant and meaningful amount of variability in effect sizes beyond what individual-level characteristics were able to capture. In another study (Burke, Hsiang, & Miguel, 2015), researchers used a similar approach to investigate political stability as a derivative factor of ongoing changes in climate. Both findings, where possible, were combined with climate data using spatial points—when typical for social sciences measures—to merge the two types of

information: those pertaining to the social psychology of violence and those related to climate change.

In all fairness, given the complexities of topics studied in this domain (e.g., violence, beliefs, stereotypes), experimental manipulation and end measures can rarely account for all observable differences; explanatory variables also do not contain the full range of details required for analysis. These facts highlight the need to expand traditional methodological frontiers in social psychology.

New developments in mathematics, computer science, and open-source software, combined with a growing emphasis on scientific data sharing (e.g., NSF, 2014), provide easy access to anonymous individual responses from large samples. Adding well-documented spatial details can result in successful embedding of individual responses in their corresponding environments; it can also reduce the risks of interpretation based on models that often violate basic assumption of traditional statistical methods.

Example of a Traditional Approach

We begin by using an example of religiosity to show the relevance of spatial techniques to social psychology. Our choice of this particular variable is motivated by both the importance of the topic in the field, as expressed by a considerable base of literature (Brezna, Lykes, Kelley, & Evans, 2011; Kay, Whitson, Gaucher, & Galinsky, 2009) and the availability of large survey data. Our example draws on four Gallup survey databases from 2007 through 2009 (Gallup Organization, 2007, 2008, 2009a, 2009b), with a total of 3,492 participants. The surveys assessed self-reported religiosity and other individual-level predictors that previous studies related to religiosity. Additionally, each individual response contains a variable defining the state where the respondent lives based on the zip code to their place of residence. Using this dataset, we examine individuals' self-reported religiosity as a function of previously linked predictors: gender, income, and political ideology.²

When studying a social psychological phenomenon (Y), we assume that its distribution is a factor of determined variables—predictors (x_1, x_2, \dots, x_n)—and some undetermined, random influences. In essence, we expect Y to be a function of our predictors and some omitted variables. Following commonly used notation, we can express this relationship, for person i as:

$$Y_i = \mu_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni} + \varepsilon_i \quad [\text{Formula 1}]$$

where Y_i is the level of the dependent variable that characterizes a subject with X_{1i} , X_{2i} , and X_{ni} levels of the predictors, μ_0 is the mean of the sample, and ε_i is the error variance. Another way to think about the error variance is the quantity in Y not accounted for by the model: namely, the difference in the observed level of Y and what our equation predicts based on the values of X_1, X_2, \dots, X_n and the sample mean μ_0 . Typically, we assume this error is due to random influences the subject experienced: errors while measuring Y , individual differences not relevant

to Y , etc. Further, we assume that their expression is independent and randomly distributed across subjects and, therefore, conditions (levels of X_1, X_2, \dots, X_n).

We use the predictors of religiosity to fit a model of i -th participant's religiosity. Our model, describes the religiosity level of any i participant as follows:

$$Y_i = \mu_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 income_i + \beta_4 ideology_i + \epsilon_i \quad [\text{Formula 1.1}]$$

Assuming the model is well specified, the observations and predictors are independent, and no hidden moderators confound the results, we expect the model to meet the following criteria (Shayle, George, & Charles, 2006): (a) The residuals (ϵ_i) are randomly distributed, (b) the variance of residuals is constant, and (c) the mean of residuals equals 0. Thus, we can now re-write the model as:

$$E(Y_i) = \mu_0 + \beta_1 sex_i + \beta_2 income_i + \beta_3 age_i + \beta_4 ideology_i \quad [\text{Model 1}]$$

Using a general linear model, we can obtain the expected religiosity for a given participant as a factor of our predictors (see Code, lines 4 and 5, in the code section used).

As Table 9.1 shows, this Gallup database—consistent with previous studies—does indeed suggest that women compared to men, individuals with lower (vs. higher) income, and those who are more (vs. less) conservative are more religious. But do these results tell the full story?

Toward Social Psychological Applications of Geospatial Data

The core assumption of traditional statistical tests in psychology is the independence of observations (Cohen, Cohen, & West, 2013). If each individual is subjected to different and *random* impacts, then the independence assumption is valid and

TABLE 9.1 Results of a General Linear Model Predicting Religiosity as a Function of Age, Gender Income, and Political Ideology

	<i>Beta</i>	<i>SE</i>	<i>t</i>	<i>Confidence intervals</i>	
				2.5%	97.5%
Intercept	1.77	0.05	33.47	1.66	1.88
Sex (Female = 1)	0.15	0.03	9.19	0.18	0.28
Age	0.10	0.01	6.56	0.06	0.11
Income	−0.07	0.01	−3.64	−0.06	−0.03
Liberal versus:	—	—	—	—	—
Moderate	0.19	0.03	9.44	0.26	0.39
Conservative	0.40	0.03	19.79	0.34	0.47

Note: Five states were excluded from the analysis due to a small number of participants: WY ($n = 8$), VT ($n = 10$), SD ($n = 11$), RI ($n = 14$), and ND ($n = 15$). Final sample contains 3,505 observations; $AIC = 7756.16$; $R^2 = 0.15$; adjusted $R^2 = 0.15$.

model is correctly specified. More common, however, is that samples are relatively homogenous (undergraduate students from university *X*, patients in clinic *Z*, etc.). Consequently, theoretically randomly sampled individuals usually share attributes specific to their particular population (e.g., community-based norms, regional weather patterns, etc.). If these external impacts are equal (or rather, are normally distributed) across individuals, then they will result in a distribution of residuals that resembles studies with truly independent observations. Therefore, the researcher will be statistically correct in assuming that the independence assumption is met, ergo: The conclusions can be extended and generalized to the population. Still, because traditional tests used in psychology do not allow differentiation between cases where (a) observations are independent, or, (b) tied by common environmental and/or societal confounding factors, other means are necessary.

In the discussion of this point, we have considered a popular and well-studied phenomenon in social psychology that can be studied without the need to tackle spatial variables. Now, we will compare and contrast the results when spatial variables are introduced. Building on this example, we will introduce definitions necessary for comfortable interactions with spatial data. Finally, we will present examples of useful visualization and statistical techniques for spatial data analyses. As we move from one example to another, we will reference the script used.

Space as a Random-Effects Variable

If a model of the type presented in Table 9.1 is well specified—given the factors impacting the dependent variable—and the assumptions regarding the residual variance are correct, we should expect *no* spatial differences in residual expression or coefficient predictor size beyond the confidence intervals of our model. After all, our model should account for the key predictors of *Y*, and therefore any differences in population composition in various spatial locations (different male to female ratios, high vs. low income participants, etc.) will be incorporated in model estimates for each individual. Space should not affect the remaining residual variance.

One way to test the possibility that spatially embedded variables may be important to our data is to quantify the amount of variance in our dependent variable that results from a proxy for spatial location. Here we use each individual's state as this proxy. Given that our interest is in the relationship between the listed predictors, the state variable should be treated as a random-effects variable. Its ties to religiosity can be captured using a hierarchical—also called multilevel—model and looking at the intraclass coefficient (Shayle et al., 2006; Singer, 1998). Multilevel modeling assumes that data are hierarchically structured, and lower-level data points are clustered within higher-level groups. In the case of our example, the lower-level data points are individual responses. The second level is the geo-spatial location of each individual (her or his state). Hierarchical modeling allows the incorporation of random-effects variables: variables linked to the dependent

variable but only a random subset of all possible levels of which are available in the dataset, and/or of no theoretical interest to the researcher (Shayle et al., 2006).

First, using the simplest case, let's evaluate the random effects of state on the variance of religiosity using a hierarchical model with state as its level-2, random-effects variable. Such a model where the intercept is the only fixed-effects factor is an empty or unconditional means model and can be expressed as follows:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad [\text{Formula 2}]$$

In our example, j refers to the *state* (New York, Ohio, etc.) in which the i^{th} respondent resides. β_{0j} is the mean religiosity score in *state* j , while ε_{ij} is the error³ related to the i^{th} participant's deviation from the average score for the state in which she or he resides. This effect can be further described as:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad [\text{Formula 2a}]$$

Here γ_{00} is the population (grand) mean, and u_{0j} is the amount state j 's mean deviates from the grand mean. Similar to ε_{ij} , which captures an individual's difference from the mean of their state, u_{0j} quantifies the degree of state j 's departure from the sample mean. Both terms are measures of error, and when squared capture the error variance related to the corresponding factor (ε_{ij} —level-1: individual error; u_{0j} —level-2: state error).

By combining Formulas 2 and 2a, we arrive at our model (see Code, lines 6–10):

$$E(Y_{ij}) = \gamma_{00} + u_{0j} \quad [\text{Model 2}]$$

Similar to squaring the error in the case of single-level models to obtain the error variance, we can square u_{0j} to obtain level-2 error variance, referred to using the letter tau: τ_{00} . Its ratio to the total error variance—error variance from levels 1 and 2, is referred to as the intraclass coefficient (rho) and quantifies the percentage of the error variance in religiosity driven by our spatial proxy:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \quad [\text{Formula 3}]$$

The fixed-effect variable in this model is the intercept: the grand mean γ_{00} . In the random-effects section, we see the error variance divided between that related to our level-2 (state-level) random-effects variable and level-1 (participant-level) random-effects variance. The value of ρ in our example suggests that 7% ($0.05 / (0.05 + 0.58)$) of the error variance is tied to the spatial proxy variable we used as the random-effect variable in Model 2. Recall that Model 1, with five individual-level predictors and 3,487 degrees of freedom for ε_{ij} , accounted for about 15% of the variance in religiosity, leaving 85% of variance unexplained.

TABLE 9.2 Results of the Unconditional Means Model

<i>Result</i>	<i>Estimate (SE)</i>
Fixed Effects	
Intercept γ_{00}	2.42** (0.04)
Random Effects	
Intercept τ_{00}	0.05 (0.22)
Residual σ^2	0.58 (0.76)
Model Fit	
AIC	9212

** $p < .01$.

Spatial Variables and Individual-Level Indices

The Gallup databases used in our example above contain a categorical spatial variable of the geometric class area, or the state where the interview occurred. While lacking the precision of coordinate point data, the random-intercept model suggests some conclusions that may suffice to explain some of the variance in religiosity. It is also sufficient, for our purposes, to combine Gallup individual-level responses with macro-level data.

Our example certainly misses one such variable, a measure of spatial “closeness” or proximity between states. This information is critical for evaluation of the probability that the clustering of individuals in states appears due to chance alone. U.S. Census Bureau data offer one way to capture “closeness.” Specifically, we use it to calculate distances between the population-weighted center of each state and all other states (Geography Division, 2011). These distances can be used to calculate various indices of spatial dependency between values of the variable of interest (see as an example Moran’s I statistic, described below). Beyond capturing dependent variable spatial dependencies, mapping individual responses onto macro-level data can help to quantify spatial influences in the relationships between individual-level predictors.

The researcher may decide that differences observed across space or time in the dependent variable may be driven by a societal factor, such as Gross Domestic Product (GDP), inequality, neighborhood health level, community prejudice level, political instability, and so on. Obtaining data that capture such dimensions is not difficult thanks to organizations such as the World Bank that make their data available online, increasingly on a cost-free basis (see <http://gapminder.org> for many other sources of data).

In another relevant example, Nisbett and Cohen's (1996) defined cultures of honor and offered a number of examples. Trying to dissect higher crime levels in southern states, they noticed a number of characteristics differentiating these states from those with comparatively lower crime levels. Perhaps most prominent among these factors were religiosity and conservative ideologies (Leung & Cohen, 2011). Honor cultures were also correlated with lower government (state) intervention due to low state population density (Boski, 2009). For this chapter's example, we tested this spatially embedded societal characteristic as a predictor of religiosity. We used the percentage of population living in urban areas in each of the states (Geography Division, 2011) and merged it with data from the Gallup data used in Models 1 and 2. When the state of the respondent was missing, it was inferred based on his/her zip code.

Mixed-Effects Models With Geocoded Fixed-Effects Variables

In our earlier "gentle" introduction to multilevel modeling, we introduced a random-effects model that evaluated the variance in religiosity around the grand mean, depending on the spatial source of our data:

$$Y_{ij} = \beta_{0j} + \epsilon_{ij} \quad [\text{Formula 2}]$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad [\text{Formula 2a}]$$

It was certainly a good start: It allowed us to determine that between-states differences in religiosity suggest the need for an alternative model to capture estimates of religiosity as a factor of individual-level predictors (sex, age, income, and political ideology). But we can take this approach a step further by obtaining estimates of individual-level characteristics in the context of higher, state-level impacts. Extending Model 2 to achieve this goal requires a slight change in the way we think about the slope of our predictors in a simple regression model.

Regression coefficient β_1 in Model 1 was simply the difference we can expect based on the i^{th} participant's sex (0 if male, and .23 if female) when compared to the grand mean. Using a multilevel approach, this coefficient would contain an account of the between-state variability in the relationship between respondent sex and religiosity:

$$Y_{ij} = \beta_{0j} + \beta_{1j} + \epsilon_{ij}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

γ_{10} reflects the mean difference in the expected religiosity between males and females, and u_{1j} is the amount the j^{th} state is expected to deviate from the mean difference of all states.

Combining this extension in the conceptualization of Model 1 slopes with the approach we used to estimate the intercept in Model 2, we are now ready to

define a multilevel model with both fixed ($\gamma_{00}, \gamma_{10}, \gamma_{20} \dots$) and random effects ($u_{0j}, u_{1j}, u_{2j}, \epsilon_{ij}$). It will account for geospatial location (as a random-effects factor) and help to investigate the fixed effects of our predictors:

$$Y_{ij} = \beta_{0j} + \beta_{1j} + \beta_{2j} + \dots + \epsilon_{ij} \qquad \text{[Model 3⁴]}$$
$$Y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10}sex_i + u_{1j} + \gamma_{20}age_i + u_{2j} + \dots + \epsilon_{ij}$$

Model 3 in Table 9.3 shows the estimated coefficients for the same predictors used in Model 1 after including the random effect of geolocation (see Code, lines 9–11). A likelihood statistic comparing both models suggests that the mixed-effects model fits the data better than Model 1: Likelihood Ratio = 97.19, $p < .01$. The estimates of Model 3 are smaller, and the confidence intervals narrower, suggesting that Model 1 exaggerated some of the differences between varying predictor levels.

TABLE 9.3 Summary of Standardized Coefficients of the Multilevel Models

<i>Fixed Effects</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
Intercept γ_{00}	2.42* (0.04)	1.8* (0.06)	2.45* (0.14)
Sex γ_{10}	—	0.21* (0.02)	0.21* (0.02)
Age γ_{20}	—	0.08* (0.01)	0.08* (0.01)
Income γ_{30}	—	−0.04* (0.01)	−0.04* (0.01)
Ideology γ_{40}			
Moderate		0.30* (0.03)	0.30* (0.03)
Conservative		0.61* (0.03)	0.61* (0.03)
URBAN γ_{01}			−0.01* (0.0)
Random Effects			
Intercept τ_{00}	0.05 (0.16)	0.03 (0.2)	0.01 (0.12)
Residual σ^2	0.58 (0.71)	0.51 (0.71)	0.51 (0.71)
Model Fit			
AIC	9121	7660.98	7655.03
$\rho(p)$	7%	6%	2%

Note: Standard errors are in parentheses. * $p < .01$, $N = 3,504$.

We should not just stop here, however, as we are still interested in the societal/state-level factor(s) that may explain the difference between states in observed religiosity. The previously discussed research on cultures of honor suggests a correlation between low levels/speeds of government service provision and high levels of endorsement of cultures of honor. The percentage of each state's population inhabiting urban areas is thus a macro-level variable that indirectly, by proxy, captures levels of population density.

Below we discuss a mixed-effects model that will allow us to incorporate a level-2 fixed effects variable, namely the percentage of the population inhabiting urban areas. To this end, all we need to do is to expand the first component of Model 3 slightly, as follows:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{URBAN} + u_{0j} \\ Y_{ij} &= \gamma_{00} + \gamma_{01}\text{URBAN} + u_{0j} + \gamma_{10}\text{sex}_i + u_{1j} + \gamma_{20}\text{income}_i + u_{2j} + \dots + \mu_{ij}\end{aligned}\quad [\text{Model 4}]$$

Here, β_{0j} is expressed as the grand mean γ_{00} , the deviation of state j from the grand mean u_{0j} , and γ_{01} multiplied by the percentage of urbanized population in state j (see Code, lines 12–14)

Model 3 reduced the amount of variability in religiosity due to state differences to 5%, down from 7% in Model 2. When the percentage of the state population inhabiting urban areas was added as the level-2 fixed effects variable in Model 4, the amount of variability was further reduced to 2%. The level-2 variable was a statistically significant predictor of a decrease in religiosity with each unit increase in the percentage of the population in urban areas, $t(42) = -4.90$, $p < .01$. But it is also important to note that AIC (Akaike's Information Criterion; for an excellent discussion of the limitations of goodness of fit indices, see Hsiang et al., 2013) was only slightly decreased by introducing the level-2 fixed effect variable. This example is simply meant as an illustration; of course, there may be other social-environmental factors that are more relevant.

Naturally, there is no evidence from our very correlational database that factors such as age, gender, or urban residence causally relate to religiosity. It is entirely possible that third variables explain away any of these patterns. If available, confounding variables can be entered, but it is sometimes the case that aggregate-level variables are so highly correlated that any one of them can replace the others. Moreover, any level-2 or structural variable is not only correlational but also presents potential interpretational difficulties due to the fact that the researcher is typically dealing with averages or percentages assigned to a particular polygon. In our example, urbanicity is literally the degree to which a state has higher population density, and it is likely that population density varies for around different localities within each state. Furthermore, it is entirely conceivable for a structural variable to show reversed trends from the same variable defined at an individual level. Community-level income is not the same as individual-level income;

for example, mental health may be superior for low-income residents of higher income communities than for low-income residents in lower income communities. Thus, analysts should interpret aggregate-level results with caution and not to commit the error of ecological fallacy, which is to assume that aggregate results match individual-level results (e.g., Robinson, 1950).

There are two analytic strategies that afford interesting opportunities to increase aggregate-level variables' substantive value. First, if we had a database that covered many more years, then an argument for *plausible causality* of aggregate-level variables could be made by introducing temporally lagged variables and examining whether associations are most peaked just prior to the measurement of the dependent variable. It is largely through such strategies that researchers have recently claimed that income inequality, an aggregate-level variable, plausibly causes numerous mental health outcomes (Pickett & Wilkinson, 2015). And second, although numerous studies have contrasted the main effects of individual- and aggregate-level variables, to date, few have examined interactions between these factors.

A Closer Look at the Distribution of Residuals Across Space

The value of ρ in the unconditioned means model could have a number of explanations, the most straightforward being that religiosity varies across states due to the different population compositions with respect to the key predictors used in Model 1. Another potential explanation is that Model 1 was improperly specified, and therefore, at least some of our estimates are biased. Optimally, therefore, our model should incorporate both factors to account for the random effects tied to location—as well as the fixed effects of our predictors—to estimate our dependent variable more accurately (as in Model 3). Yet, whereas the precision and accuracy of the estimates is the key determinant of the model's value, spatiotemporal techniques can go beyond simply reducing estimate bias. As we mentioned earlier, they can guide researchers toward important, but overlooked spatially embedded societal and environmental factors.

Before we introduce spatial methods that can provide useful statistical tests to aid researchers in this process, let's use a geospatial technique to visually explore the assumptions behind Model 1. If the model was sufficient and independent of other spatially embedded factors, we would expect a random distribution of residuals and their variance across the United States. After all, Model 1 estimates are based on the individual's sex, age, income, and political ideology; any differences between states in the composition of their populations (e.g., higher average income, fewer females, younger residents) will be reflected in the estimated values and should not impact the difference between estimates and observed data. Figure 9.3 shows the mean *variance* of residuals by state; states with low mean variances are those where the model fits best (lighter shades), whereas those with high mean variances are those where the model fits worst (darker shades). The variance of residuals diagnoses areas where the model fits poorly and, as a result, can lead

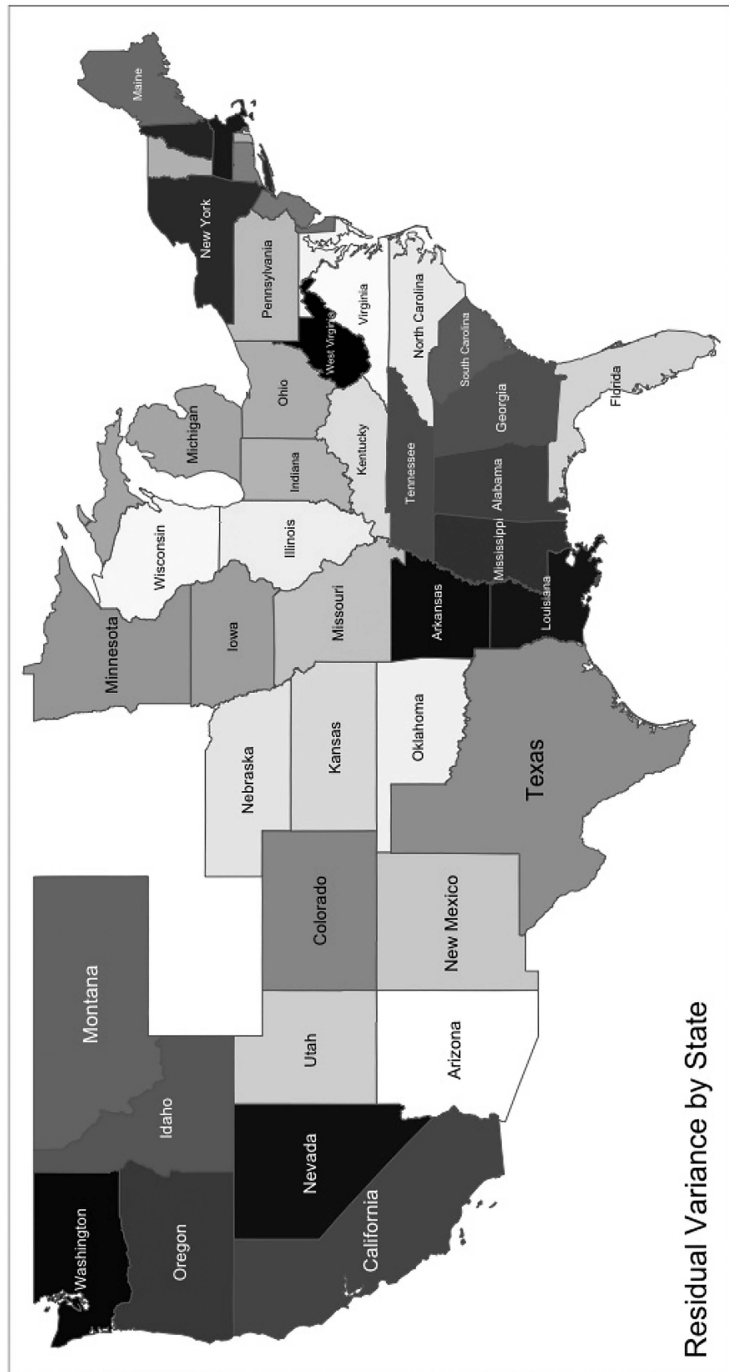


FIGURE 9.3 Map of Mean Residual Variance per State

Note: Darker areas depict states with higher variability, where the model fits less well. Five states were excluded from the analysis and therefore are not graphed.

an investigator to look more closely at the phenomena in question and facilitate a deeper understanding of the phenomenon at-hand. Perhaps important predictor variables are omitted or predictor variables have a restricted range.

Figure 9.3 shows clusters of states where Model 1 is especially inaccurate and displays high variability of residuals. Yet, are these clustered inconsistencies sufficient to generate concerns about the model's accuracy? After all, we should expect some clustering in any phenomena studied in space simply because spatial (and temporal) proximity can lead phenomena to resemble each other (Diggle, 2010; Pearson, 1990). In other words, is the spatial correlation of the predictive inaccuracy of Model 1 different from what we can expect when looking at stochastic expressions of phenomena studied in space? We will return to this question later to introduce geospatial statistics that can help us to evaluate and answer it. Then, using this example as our foundation, we provide the reader with basic terminology needed to interact with spatiotemporal data and variables.

Quantitative Methods and Spatiotemporal Data

Measures of Spatial Clustering

The null hypothesis for spatial tests of clustering assumes that the likelihood of stochastic expressions of any variable or process resembling one another increases based on their relative closeness. The question they address is therefore: Are the observed similarities in values or densities across space larger than what we should expect due to mere proximity? Spatial statistics offer two groups of tests to address this question. One is designed to examine the density of an expression of a phenomenon (e.g., number of schools per county, their location in relation to each other), the other the values of a phenomenon (average SAT scores in each of the schools).

Point-Pattern Analysis

Unlike the example, which focused on the expression of a quantitative variable in space (residuals), point-pattern analysis does not look at the variability in values assigned to spatial units but rather the count or distances between a categorical variable, and evaluates it against a “*random*” expression. It was no accident that we put random in quotation marks—the pattern of a random expression of an object in space may not be random at all. Instead, a deviation from this non-random pattern, in essence a random pattern, may suggest that the expression of the studied variable does not support the null hypothesis of the observed point pattern being the result of chance alone (Upton & Fingleton, 1985b). For example, when looking at the number of people suffering from a contagious disease, we should expect clusters of people infected, rather than singular exemplars randomly distributed in space.

To better ground the point-pattern techniques in social psychology, let's imagine a researcher who developed a hypothesis that specific social, economic, and/or political factors can predict the number and distribution of places of worship. Testing her hypothesis, she chooses local assembly districts across the five boroughs of New York City as her spatial unit. The unit is a polygon of the shape and size equivalent to each district. The researcher can now obtain the count of places of worship and their proximity to one another in each polygon. Next, she may proceed to examine social, economic, and/or political variables as predictors of differences between the geo-administrative units of her choice.

One of the benefits of the point-pattern analysis is the flexibility in choosing spatial units of micro and macro sizes to match the research question. By narrowing the size of the spatial unit, one can look at patterns in the location of various neuron types, or by enlarging the scale, examine global patterns of civil unrest.

The method above is one of many that fall under the umbrella term *pattern identification methods*. This group of methods is applied to extract meaningful patterns in spatial data (Upton & Fingleton, 1985a, 1985b), temporal data, or both (Chan, 1999; Povinelli & Feng, 2003). Applications of these methods have been successful in a variety of domains ranging from machine learning (Begg & Kamruzzaman, 2004), neuroscience (Norman, Polyn, Detre, & Haxby, 2006), ecology (Pearman, Guisan, Broennimann, & Randin, 2008), agent-based modeling (Benenson, 1998), and economics (Desmet & Rossi-Hansberg, 2014). Yet pattern identification remains largely unexplored in social psychology.

Spatial Autocorrelation

The second group is designed to detect clusters in quantitative data. In other words, the correlation of values of the same phenomena in space, also referred to as autocorrelation.

Commonly used indexes of spatial autocorrelation are Moran's (1950) and Geary's C (1954); I is more often available in statistical programs and overcomes some of the shortcomings of C (for a more detailed discussion, see Upton & Fingleton, 1985a, Chapter 3). When either differ significantly from its null, the inference is that the similarity (or difference) in values between proximal data is larger than what we should expect due to an expression of a stochastic process alone. The null hypothesis for spatial autocorrelation, similarly, is that observations are not clustered.

The visualization of Model 1 residuals' variance over space in Figure 9.3 suggested some clustering, but did not answer the question: Are the clustered inconsistencies in the predictive accuracy of Model 1 sufficient to worry about the model? To answer this question, we can use Moran's I (see Code lines number 15–21). Here, we want to point out that the reader is well familiar with a test

that shares the logic behind I , namely when comparing observed counts to those expected in a simple χ^2 test.

Using the inverse Euclidean distance matrix, the results of Moran's I suggest that the null hypotheses of no spatial correlation in our religiosity measure needs to be rejected, based on the comparison between the expected value of $I = .15$, and the observed $I = -.02$, $SD = .03$, $p < .01$. The initial concerns, confirmed also by the significant improvement in model fit when applying a mixed-effects model compared to the traditional model, seem well justified.

Spatiotemporal Meta-Analysis

Meta-analyses are often uniquely poised to incorporate spatiotemporal moderators. Thus, studies manifest not only the potential influence of individual, sample, intervention, and study design characteristics, but also larger influences within neighborhoods, communities, and ecological factors (Berkman, Glass, Brissette, & Seeman, 2000; Johnson et al., 2010; Kaufman, Cornish, Zimmerman, & Johnson, 2014). Although extant meta-analyses have routinely considered study- and participant-level factors to explain heterogeneity (Johnson et al., 2017), they have tended to neglect the environments where studies are conducted, which is ironic, because, as we have noted, cultures and other factors cluster in space over time. The settings in which study participants live, work, and play can be characterized by such environmental factors such as disease, weather, local and broad economic trends, the level of stigmatization of minority groups, and allostatic load due to all causes. For example, a particular social-psychological factor may be more relevant because it takes place at a time when people are most susceptible to its influence, then its effect should be greater, as ecological models predict (Johnson et al., 2010; Kaufman et al., 2014). Similarly, just as strong situations can impel everyone to act the same, overwhelming internal predispositions (Ross & Nisbett, 1991) so too can spatiotemporal forces direct behavior (e.g., extreme weather can force people to stay indoors or to flee to better climes).

Spatiotemporal meta-analysis capture heterogeneity in study environments that vary over time and space (Johnson et al., 2017). The strategy builds on *cross-temporal meta-analysis*, which focuses on how cohorts differ over time (usually within particular nations); for example, Twenge (2000) found that for high schoolers and university students, anxiety levels rose between the mid-1950s and the mid-1990s. Spatiotemporal meta-analysis adds spatial information—and its combination with time—as important factors in the analytic practice, using strategies such as we document in this chapter to describe particular places at particular times and to build models of these factors. Logically, to the extent that relevant spatiotemporal information on environmental conditions is available and varies widely, it can help to explain variability in study results that is not explained by individual, sample, study, or intervention features. Moreover, when enough studies are available, statistics such as we have discussed as well as mapping the residuals of effect sizes,

parallel to Figure 9.3, may detect locations where study results differ systematically and suggest alternative variables to control in subsequent models.

Spatiotemporal Data

Spatial datasets are often arranged using arrays, which can and usually do carry additional information characterizing the spatial unit, such as the size of the grid, its range, and numerous other factors. Additionally, each spatial datum usually carries a temporal dimension and the value of measured variables at each spatial point in time. For instance, we can consider surface temperatures, precipitation, and other relevant variables. To better illustrate, let's look at a database measuring precipitation level in every point on the classic latitude-longitude grid over a 12-months period. We could represent this information using a two-dimensional matrix, which is more commonly used for psychological data. It would need to consist of four columns: latitude, longitude, elevation, time, and the value of the precipitation. Using a two-dimensional data structure would require four columns, with $360 \text{ (longitude)} \times 180 \text{ (latitude)} \times 12 \text{ (time)} \text{ rows} = 777,600$; now imagine 5 years of data on precipitation, temperature, and wind strength and you have a database that might intimidate some (and meteorological databases that use more precise variables are much larger). Below we briefly describe popular alternatives used to store spatial information that allow better storage, processing, and exchange of data, along with references to packages that can aid researchers in unpacking and interacting with this type of data to extract the needed details.

Network Common Data Format (NetCDF)

As an alternative to the large matrix described above, spatiotemporal data are often stored in Network Common Data Format (NetCDF), which uses arrays to contain all information. Figure 9.4 presents a visual abstraction of an array describing precipitation over a 12-month period for every point on the classic latitude-longitude dimension. This format allows the user to access and process sub-sections of the arrays without the need to interact with the complete dataset. NetCDF files, like most spatial data formats, contain several components: a header that carries meta-information concerning the data, such as the grid size, number of variables, their name, type, as well as the actual body of data. For this reason, opening NetCDF files requires additional steps and the use of specialized functions (commands). R offers a number of packages that can help the researcher to complete this task (Michna & Woods, 2013). Once opened, the user can proceed to extract and process the relevant sections of the data. To help readers with no past experience, we include a separate section in the Code. Line 22 opens a NetCDF data file, line 23 extracts the latitude and longitude points, that are then used in line 25 to assign the corresponding country details to each point using

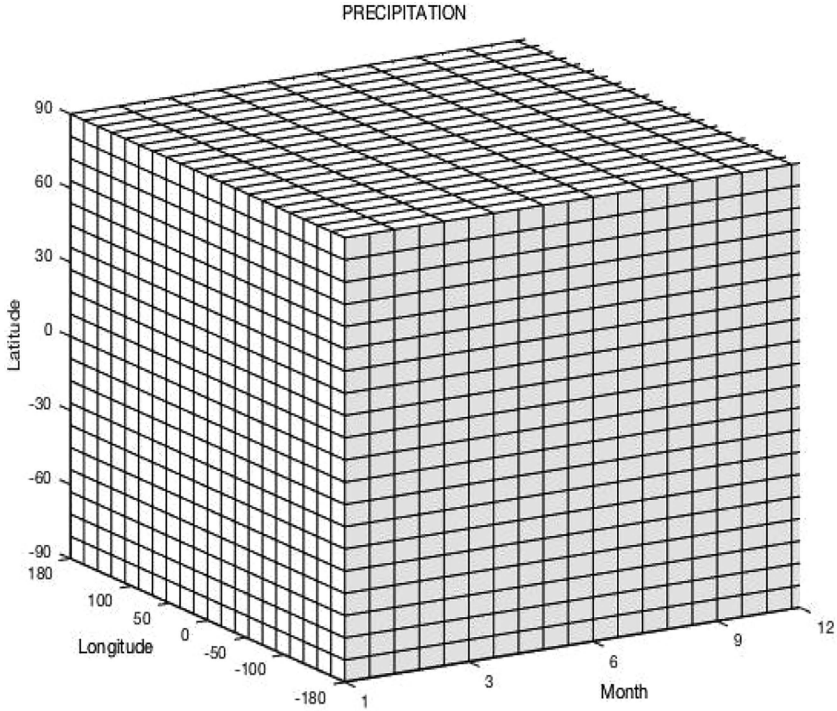


FIGURE 9.4 A Visual Abstraction of a Three-Dimensional Dataset Containing Monthly Precipitation for Each Latitude and Longitude of the Globe

the function in line 24. Line 26 returns the country details to the original file containing monthly air temperatures.

While it is difficult to overestimate the value of such detailed datasets for climate research, their relevance to social psychology may not be immediately clear. However, thanks to a large number of studies, the benefits of accounting for climate-related information, when studying inter- and intra-group phenomena, is now well documented. Using the parsimonious NetCDF data, researchers have showed strong ties between psychological processes and environmental factors. Specifically, a geospatial meta-analysis of experimental and non-experimental studies showed a robust link between temperatures and interpersonal, intergroup, and societal violence (Hsiang et al., 2013).

The characteristics of NetCDF files that make them more efficient than other binary formats can also present challenges absent from the two-dimensional data commonly used in social psychology. Nonetheless, thanks to highly engaged research communities, R, an open-source statistical package, offers packages to facilitate the use of NetCDF files. Examples of such packages include RNetCDF and cshapes (Michna & Woods, 2013; Weidmann & Gleditsch, 2010). The Code section offers several other suggestions for further readings as well as additional packages.

Spatial Polygons

Another way to overcome the challenges presented by finely gridded point data is to convert them into spatial units that use aggregated variables within relevant geographic, administrative, or sociopolitical boundaries. Consider an example of a researcher using states as spatial units. Here, spatial polygons encompass states. One of the benefits of this class of spatial data is its capacity to capture and compare areas of different shapes and sizes, which can be more meaningful, as it allows us to compare between units that differ politically, economically, or socially. Looking at between-state differences in religiosity, or plotting the residuals across states, would not be possible when comparing spatial units of equal size and shape. Doing so seems especially important in studying phenomena of special interest to social psychologists, such as cross-nation differences in the child's cognitive development, or neighborhood-level prejudice as a factor that plays a role in the success or failure of a specific health interventions (e.g., Reid, Dovidio, Ballester, & Johnson, 2014). In these and many similar cases, comparing spatial units of equal size and shape would yield meaningless outcomes.

Spatial polygons often come in the form of ready-to-use shape files accompanied with a set of meta-files that contain information concerning their size, coordinate reference system location, and perimeters. Once imported, the meta-files contain multiple sources of sufficient information to combine geocoded data (e.g., GDP per capita) with the appropriate polygon (e.g., country). Figure 9.3 mapped the residual variance onto spatial polygons—states, using shape files. A variable contained in the Gallup poll data identified states, allowing matching of the corresponding data to the spatial polygon representing the corresponding state and “filled” the state's polygon with a shade whose intensity matched the magnitude of the residual variance.

While less detailed than NetCDF files, spatial polygon data also allow the user to obtain important information for spatial analysis. This information includes distances between (population-weighted or absolute) centers of the polygons, neighbors' classifications, and other details. It also bypasses the potential risks entailed in inadvertently disclosing the location of individual respondents presented by the use of latitude and longitude coordinates with participant-level data.

This issue is closely tied to another argument in favor of incorporating spatiotemporal data into social psychology and understanding their structure: the improvement and rise in popularity of online survey platforms. Recently, survey platforms commonly used in social psychological research (e.g., Qualtrics) have allowed for capture of latitude and longitude coordinates linked to responses. For obvious reasons of protecting participant anonymity, the coordinates usually can be used only for mapping and aggregated spatial analysis. As we alluded to above, geocoded social media posts, some of which identify particular individuals, also can be harvested and modeled in meaningful ways.

Spatial data can be also categorized using geometric classes, many of which rely on NetCDF data, such as points and lines. Each of these can capture a

spatial component via a different scale: discrete, ordinal, or interval. And similar to psychological statistics, the type of the scale is one of the factors that should drive the choice of the appropriate statistical test.

Conclusions

Our hope is that this chapter will serve as a starting point for researchers to develop useful skills and encourage them to venture into the rarely explored spaces that can help better study what we all are so eager to better understand—socially embedded psychological processes.

Placing individuals' responses and behaviors in a broader context of the political, societal, and natural habitat they occupy can aid researchers in better capturing the interplay between individual-level processes and environmental factors. Examples of direct benefits of spatial techniques include (a) the re-evaluation of research outcomes against potential false positives; (b) ensuring that sufficient variance remains explained by predictor variables rather than the overlap between those variables and unaccounted spatially embedded societal, environmental, or temporal factors (or their combination, interaction, etc.); (c) reduction of the risk of erroneous interpretations driven by misconstrued models; and (d) comparison and statistical testing of non-spatial and spatial models using recently popular indices of the goodness of fit such as AIC.

Spatial databases are easily accessible and well-documented statistical methods with a long history of applications and validations in a wide range of domains. It is our hope that this chapter has encouraged some readers to explore the applications of these methods. At the same time, for those who reached for this chapter with a specific project in mind, our hope is that the carefully selected references from various fields, appearing throughout the chapter, will serve as a good starting point.

Code

R code

Packages and data

Packages used in this chapter:
nlme, xtable, lm.beta, maps, ggmap, rgdal, rgeos, maptools,
tmap, dplyr, tidyr, sp, ape, RNetCDF

Please install and activate them to follow the examples using:

```
1 packages.install(c("nlme", " xtable ",..))
2 library(pckg name1)
3 pop<-read.table("http://www2.census.gov/geo/
docs/reference/cenpop2010/CenPop2010_Mean_ST.txt",
header=T, sep=", ")
```

Traditional model

```

4 M1=lm(relimp ~ female + age + income + moderate
+ conservative, data=glp2l, na.action=na.omit)
5 summary(M1)

```

Unconditional means model

```

6 glp.int <- groupedData (relimp ~ 1 | STATE,
data=glp2l)
7 MLM00 <- lme(relimp ~ 1, data=glp.int, random = ~
1 | STATE, na.action=na.omit)
8 summary(MLM00)

```

*Mixed effects models***Level-1 Fixed Effects**

```

9 glp.int01 <- groupedData (relimp ~ 1+ age + female
+ income + moderate + conservative| STATE, data=glp2l)
10 MLM01 <- lme(relimp ~ 1 + age + female + income
+ moderate + conservative, data=glp.int01, random = ~ 1 |
STATE, na.action=na.omit,method="REML", corr=NULL)
11 summary(MLM01)

```

Level-1 and Level-2 Fixed Effects

```

12 glp2l.int11 <- groupedData(relimp ~ 1+ age + female
+ income + moderate + conservative| STATE, data=glp2l,
outer= ~ POPPCT_URBAN)
13 MLM11 <- lme(relimp ~ 1+ age + female + income
+ moderate + conservative + POPPCT_URBAN, data=glp2l.
int11, random = ~ 1 | state, na.action=na.omit)
14 summary(MLM11)

```

Spatial autocorrelation

```

15 popC=pop[pop$STNAME%in%unique(glp2l$state),]
16 locations<- as.matrix(popC[,c(5,4)])
17 DistancesI<- spDists(locations,longlat=TRUE)
18 weights=1/DistancesI
19 weights[is.infinite(weights)]<- 0
20 ResM<- unique(glp2l[,c("STATE","resM")])
21 Moran.I(x=ResM$resM,weight=weights)

```

NetCDF FILE

Air temperature data is provided by the Earth System Research Laboratory of the National Oceanic & Atmospheric Administration, Physical Sciences Division, Boulder, Colorado, USA. And available at their web site at <http://www.esrl.noaa.gov/psd/>

For this specific file please visit: https://www.esrl.noaa.gov/psd/data/gridded/data.UDeI_AirT_Precip.html#detail

```
22 air=read.nc(open.nc("~/air.mon.mean.v401.nc"))
23 points=cbind(air$lon],air$lat)
24 coords2country = function(points) {
  countriesSP<-getMap(resolution='low')
  pointsSP = SpatialPoints(points, proj4string=CRS
    (proj4string(countriesSP)))
  indices = over(pointsSP, countriesSP)
  indices$ADMIN
  indices$ISO3
}
25 OUT=data.frame(coords2country(points))
26 AIR=cbind(air,OUT)
```

STATA code

Traditional model

```
4 reg relimp age female income moderate conservative
```

Unconditional means model

```
6 mixed relimp || state:
```

Mixed effects models

Level-1 Fixed Effects

```
10 mixed relimp age female income moderate conservative
|| state:
```

Level-1 and Level-2 Fixed Effects

```
13 mixed relimp age female income moderate conservative
POPPCT_URBAN || state:
```

Spatial autocorrelation

```
17 spatwmat, name(weights) xcoord(lon) ycoord(lat)
band(0 5000)
21 spatgsa ResM, weights(weights) moran twotailed
```

FURTHER INFORMATION

Additional packages and instructions related to spatiotemporal analyses are available at <https://cran.r-project.org/> or <https://www.r-pkg.org>. The second link contains information about new packages available through the first web-page, but also those deposited by the authors to GitHub (<https://github.com/>) along with manuals, descriptions, and suggestions.

Notes

- 1 1 degree = 60 minutes = 0.01745 radians
- 2 We reiterate that our focus here is on methods to detect socio-spatial dependencies. We employ several shortcuts for the sake of illustration (e.g., using a single item as a dependent variable rather than average scores of multiple questions capturing the same construct), but we discourage researchers from following suit in their own analyses.
- 3 $\varepsilon_{ij} \sim N(0, \sigma^2)$; $u_{0j} \sim N(0, \tau_{00})$.
- 4 Where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]$:
variance of the intercept τ_{00} ; the slope τ_{11} ; covariance τ_{01}, τ_{10}

References

- Begg, R., & Kamruzzaman, J. (2004). A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *Journal of Biomechanics*, 38(3), 401–408. doi:10.1016/j.jbiomech.2004.05.002
- Benenson, I. (1998). Multi-agent simulations of residential dynamics in the city. *Computers, Environment and Urban System*, 22(1), 25–41. doi:10.1016/S0198-9715(98)00017-00019
- Berkman, L. F., Glass, T., Brissette, I., & Seeman, T. E. (2000). From social integration to health: Durkheim in the new millennium. *Social Science & Medicine*, 51, 843–857. doi:10.1016/S0277-9536(00)00065-00064
- Boski, P. (2009). Kulturowe ramy zachowań społecznych. In E. Betlejewska (Ed.), *Podręcznik psychologii międzykulturowej*. Warsaw: Wydawnictwo Naukowe PWN.
- Breznau, N., Lykes, V. A., Kelley, J., & Evans, M. D. R. (2011). A clash of civilizations? preferences for religious political leaders in 86 nations. *Journal for the Scientific Study of Religion*, 50(4), 671–691. doi:10.1111/j.1468-5906.2011.01605.x
- Burke, M., Hsiang, S. M., & Miguel, E. (2015). Climate and conflict. *Annual Review of Economics*, 7, 577–617.
- Chan, W-S. (1999). A comparison of some of pattern identification methods for order determination of mixed ARMA models. *Statistics & Probability Letters*, 42(1), 69–79. doi:10.1016/S0167-7152(98)00195-00193

- Cohen, J., Cohen, P., & West, S. G. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London: Routledge.
- Dasen. (1994). Culture and cognitive development from a Piagetian perspective. In W. J. Lonner & R. S. Malpass (Eds.), *Psychology and culture*. Boston: Allyn and Bacon.
- Department of Defense. (2000). *NIMA technical report TR8350.2 World Geodetic System 1984: Its definition and relationships with local geodetic systems* (3rd ed.). Retrieved from <http://earth-info.nga.mil/GandG/publications/tr8350.2/wgs84fin.pdf>
- Desmet, K., & Rossi-Hansberg, E. (2014). Spatial development. *American Economic Review*, 104(4), 1211–1243. doi:10.1257/aer.104.4.1211
- Diggle, P. J. (2010). Historical introduction. In A. E. Gelfand, P. J. Diggle, P. Guttorp, & F. Montserrat (Eds.), *Handbook of Spatial Statistics* (pp. 3–17). New Jersey: Taylor & Francis.
- EUREF (2016). *Reference frame sub commission for Europe*. Retrieved from www.euref.eu/
- Gallup Organization. (2007). *Lifestyles—economy/religion* [Survey Poll]. USAIPOGNS2007–2040.
- Gallup Organization. (2008). *Lifestyles—economy/religion* [Survey Poll]. USAIPOGNS2008–2046.
- Gallup Organization. (2009a). *Social series—values and beliefs* [Survey Poll]. USAIPOGNS2009–2009.
- Gallup Organization. (2009b). *USA Today* [Survey Poll]. USAIPOUSA2009–2022.
- Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3), 115–145. doi:10.2307/2986645
- Geography Division. (2011). *Centers of population computation: For the United States 1950–2010*. Washington, DC: U.S. Census Bureau.
- Haining, R. (2003). *Spatial data analysis : Theory and practice*. Cambridge: Cambridge University Press.
- Hsiang, S. M., Burke, M., & Miguel, E. (2013). Quantifying the influence of climate on human conflict. *Science*, 341(6151), 671–714. doi:10.1126/science.1235367
- ITRF (2016). *International Terrestrial Reference Frame (ITRF)*. Retrieved from <http://itrf.ign.fr/>
- Johnson, B. T., Cromley, E., & Marrouch, N. (2017). Spatiotemporal meta-analysis: Reviewing health psychology phenomena over space and time. *Health Psychology Review*, 3, 280–291. doi:10.1080/17437199.2017.1343679
- Johnson, B. T., Redding, C., DiClemente, R., Mustanski, B., Dodge, B., Sheeran, P., . . . Fishbein, M. (2010). A network-individual-resource model for HIV prevention. *AIDS and Behavior*, 14(S2), 204–221. doi:10.1007/s10461-10010-19803-z
- Kaufman, M., Cornish, F., Zimmerman, R., & Johnson, B. T. (2014). Health behavior change models for HIV prevention and AIDS care: Practical recommendations for a multi-level approach. *Journal of Acquired Immune Deficiency Syndromes*, 66(Supplement 3), S250–S258.
- Kay, A. C., Whitson, J. A., Gaucher, D., & Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. *Current Directions in Psychological Science*, 18(5), 264–268. doi:10.1111/j.1467-8721.2009.01649.x
- Lee, I. C., Pratto, F., & Johnson, B. T. (2011). Intergroup consensus/disagreement in support of group-based hierarchy. *Psychological Bulletin*, 137(6), 1029–1064. doi:10.1037/a0025410
- Leung, A. K. Y., & Cohen, D. (2011). Within-and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of Personality and Social Psychology*, 100(3), 507–526. doi:10.1037/a0022151
- Maynard, A. E. (2008). What we thought we knew and how we came to know it: Four decades of cross-cultural research from a Piagetian point of view. *Human Development*, 51(1), 56–65. doi:10.1159/000113156

- Michna, P., & Woods, M. (2013). RNetCDF—a package for reading and writing NetCDF Datasets. *The R Journal*, 5, 29–36.
- Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1), 17–23. doi:10.2307/2332142
- Nisbett, R. E., & Cohen, D. (1996). *Culture of honor*. Boulder, CO: Westview Press.
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–430. doi:10.1016/j.tics.2006.07.005
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).
- Pearman, P. B., Guisan, A., Broennimann, O., & Randin, C. F. (2008). Niche dynamics in space and time. *Trends in Ecology & Evolution*, 23(3), 149–158. Retrieved from <http://dx.doi.org/10.1016/j.tree.2007.11.005>
- Pearson, E. S. (1990). *Student: A statistical biography of William Sealy Gosset*. Oxford: Oxford University Press.
- Piaget, J. (1957). *Construction of reality in the child*. London: Routledge.
- Pickett, K. E., & Wilkinson, R. G. (2015). Income inequality and health: A causal review. *Social Science & Medicine*, 128, 316–326. doi:10.1016/j.socscimed.2014.12.031
- Povinelli, R. J., & Feng, X. (2003). A new temporal pattern identification method for characterization and prediction of complex time series events. *IEEE Transactions on Knowledge and Data Engineering*, 15(2), 339–352. doi:10.1109/TKDE.2003.1185838
- Reid, A. E., Dovidio, J. F., Ballester, E., & Johnson, B. T. (2014). HIV prevention interventions to reduce sexual risk for African Americans: The influence of community-level stigma and psychological processes. *Social Science & Medicine*, 103, 118–125. doi:10.1016/j.socscimed.2013.06.028
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. doi:10.2307/2087176
- Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. London: McGraw-Hill.
- Shayle, R. S., George, C., & Charles, E. M. (2006). *Variance components*. New Jersey: Wiley & Sons.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4), 323–355. doi:10.2307/1165280
- Stephenson, A. G. (1999). *Mishap investigation board, phase I Report*. Washington, DC: NASA.
- Twenge, J. M. (2000). The age of anxiety? The birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of Personality and Social Psychology*, 79(6), 1007–1021. doi:10.1037//0022-3514.79.6.1007
- Upton, G. J. G., & Fingleton, B. (1985a). *Spatial data analysis by example: Categorical and directional data* (Vol. 2). Chichester: John Wiley & Sons.
- Upton, G. J. G., & Fingleton, B. (1985b). *Spatial data analysis by example: Point pattern and quantitative data* (Vol. 1). Chichester: John Wiley & Sons.
- Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23), 6454–6459. doi:10.1073/pnas.1521897113
- Weidmann, N. B., & Gleditsch, K. S. (2010). Mapping and measuring country shapes: The cshapes package. *The R Journal*, 2(1), 18–24.

10

SOCIAL MEDIA HARVESTING

*Man-pui Sally Chan, Alex Morales,
Mohsen Farhadloo, Ryan Joseph Palmer,
and Dolores Albarracín*

The use of social media and social networking sites is currently widespread and is only expected to increase in the coming years. In a recent survey (Greenwood, Perrin, & Duggan, 2016), over 70% of U.S. adults aged 18 to 29 and 55% of adults aged 30 to 49 reported having a Facebook profile. Further, a solid 33% of adults aged 50 to 64 reported using Facebook, and 24% of the 86% of Americans who use the Internet interact on Twitter (Greenwood et al., 2016). Social media comprise websites and applications that facilitate the creation, expression, and sharing of information and ideas among users who (a) maintain a personal profile within the system, (b) privately or publicly interact with other users within the social networks, (c) expand their connections by searching for other users or accepting connections suggested by the platforms, and (d) may also leave the social networks and remove their connections (Boyd & Ellison, 2007).

Social media is an enjoyable outlet for people to express themselves and interact with other network members, and the increasing number of users (Mangukiya, 2016; Statista, 2010), along with the growing usage (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015), position such outlets as powerful sources of information to be used in research (e.g., with the goal of identifying discussion topics on a Facebook page; see “Topics as Important Semantic Features”). In recent decades, social media analysis has received considerable attention in various areas of research, from examining the associations between the use of social media and mental health (Jelenchick, Eickhoff, & Moreno, 2013; Lin et al., 2016) to analyzing the effects of social media on interpersonal relationships (Finkel, Eastwick, Karney, Reis, & Sprecher, 2012; Ward, 2016). Furthermore, the number of psychology articles that utilize social media as a tool or treat it as the subject of scrutiny has risen rapidly in the last decade. According to the Psychology Article Database *Psychinfo*, there were over seven times more studies involving social

media since 2010 than in the entire previous decade. The increase of research publications is expected to rise because social media are growing in popularity and becoming ever more influential in our everyday lives. Researchers can now use social media platforms to harvest a wide range of information about a population, such as the demographics of personal profiles (i.e., non-semantic features) as well as likes, favorites, follow, and text posts/messages (i.e., semantic features).

The harnessing of social media data has allowed researchers to uncover numerous aspects about its users at the individual, community, and national levels. In fact, an emerging group of scholars has analyzed social media data to understand a wide range of behaviors and attitudes, including but not limited to consumer decisions (Bennett & Lanning, 2007; De Souza & Ferris, 2015; Farhadloo, Winneg, Chan, Jamieson, & Albarracín, 2018), influenza infections (Signorini, Segre, & Polgreen, 2011), and political orientation/opinions (Schwartz & Ungar, 2015; Wu, Kosinski, & Stillwell, 2015). In the following sections, we provide a detailed overview of some sample platforms (“Social Media Platforms”) and describe different harvesting methods to collect social media data (“Harvesting Social Media Data: Approaches and Sources of Data Collection”) as well as a range of harnessing techniques to analyze non-semantic and semantic features (“Harnessing Social Media Data: Analytical Techniques for Non-Semantic and Semantic Features”). We then provide a discussion of important semantic features, including topics and the use of sentiment analysis and opinion spam detection. In the last section, we present an example to illustrate how social media data can be utilized for predictive and explanatory models. Finally, we end this chapter by describing ongoing challenges and future directions of measuring social media data in psychological research.

Social Media Platforms

At first glance, social media might appear to generate data streams that are far too shallow to advance knowledge in any meaningful way because most platforms impose constraints on how users express themselves. For instance, Twitter has a limit of about 280-character on each post/reply (i.e., tweet), Facebook has a 63,206-character allotment for a status update, and Weibo has an approximately 2,000-character restriction for every message, augmented with additional space given for photos, videos, polls, GIFs, and quotes. Given that, by design, these messages are limited, it might seem reasonable to conclude that there is little to learn from the seemingly shallow communications these sites typically generate. However, this is not what we found in a review of the relevant literature. Table 10.1 presents sample studies that have analyzed data from social media and differ in key functions, including networking, microblogging, messaging, commenting and discussion, media sharing, and news and classified advertisements. In the coming sections, we provide an overview of harvesting methods and analytical techniques in relation to the key functions of the social media used in previous studies.

TABLE 10.1 An Overview of Key Functions and Top Social Media With Sample Studies

Key Functions and Top Social Media	Sample Studies in Social Science	Methods of Data Collection (see “Harvesting Social Media Data” and Table 10.2 for approaches and sources)	Methods of Data Analyses (see “Harvesting Social Media Data” and Table 10.3 for the uses of semantic features)
<i>Networking: Users manage a personal profile, which can be used to connect with people with similar interests and background and to create groups for interactions.</i>			
• <i>LinkedIn</i> — www.linkedin.com	(Zide, Elman, & Shahani-Denning, 2014)	Three hundred user profiles were collected by groups of human resources and sales/marketing professionals and industrial/organizational psychologists	Carried out human coding of profiles and conducted chi-square tests
• <i>Facebook</i> — www.facebook.com	(Kosinski, Stillwell, & Graepel, 2013)	Over 58,000 Facebook–user profiles and a list of their Likes were collected via myPersonality Project	Created a user-Like matrix, reduced the dimensions using singular–value decomposition, and performed linear/logistic regressions
<i>Microblogging: Microblogging focuses on short updates, and users can push updates out to anyone who is subscribed to/following the corresponding account. Followers can pass along updates by reposting.</i>			
• <i>Twitter</i> — https://twitter.com	(Ireland, Schwartz, Chen, Ungar, & Albarracín, 2015)	Over 150 million tweets were obtained via Twitter APIs	Mapped tweets to U.S. counties and conducted text analyses, including the use of the Linguistic Inquiry and Word Count (LIWIC) and natural language processing (NLP) techniques
• <i>Weibo</i> — www.weibo.com	(Yuan, Feng, & Danowski, 2013)	About 18,000 Weibo messages were collected by searching keyword on weibo.com	Identified a semantic network of messages using WORDij 3.0 and performed a cluster analysis with software package NodeXL.
<i>Messaging: Users can send and receive multimedia messages instantly from family, friends, and other publishers. Messages are presented in various formats, including text, audio, photo, video, and emoticons.</i>			
• <i>Snapchat</i> — www.snapchat.com/en-gb/	No study available	–	–
• <i>WhatsApp</i> — https://web.whatsapp.com	(Cheung et al., 2015)	Messages of 40 WhatsApp users who participated in a smoking cessation intervention were obtained	Carried out human coding of messages and performed Mann-Whitney <i>U</i> tests

Commenting and Discussion: Online forums and blog comments allow users to make interactive conversations by posting messages. However, discussion of blog comments usually centers around the topic of the blog post, such as a particular restaurant.

• <i>TripAdvisor</i> — www.tripadvisor.com	(Lawrence & Perrigot, 2015)	Over 6,000 hotel reviews from 134 hotels were collected automatically	Carried out human coding of reviews and performed regression analyses
• <i>Yelp</i> — www.yelp.com	(Gui, Zhou, Xu, He, & Lu, 2017)	More than 1 million reviews and user data were obtained from Yelp 2013 and 2014 Data Challenge Dataset	Conducted sentiment classification on reviews

Media Sharing: These platforms allow users to upload and share various media such as pictures and video. Additional functions include creating profiles, creating/subscribing channels, commenting, etc.

• <i>You Tube</i> — www.youtube.com	(J. Huang, Kornfield, & Emery, 2016)	Over 28,000 videos related to e-cigarette were recorded via a YouTube crawling program, ContextMiner	Performed human coding of content and carried out descriptive analyses
• <i>Instagram</i> — www.instagram.com	(Moreno, Ton, Selkie, & Evans, 2016)	Over 1 million Instagram posts were obtained by searching for nonsuicidal self-injury (NSSI) hashtags www.instagram.com	Performed human coding to assess the NSSI hashtags meaning

News and Classified Advertisements: Users can post various items, such as classified advertisements, news, and links to third-party articles, pictures, or videos. Users are then allowed to interact with the items. For example, the order of display of the items on Reddit is subject to time or to number of votes, which is the core social aspect in these communities. The Reddit community jointly decides which news items get seen by more people.

• <i>Reddit</i> — www.reddit.com	(Zhan, Liu, Li, Letschow, & Zeng, 2017)	Over 27,000 posts were collected by keyword searches and analyses of metadata via Reddit API	Performed topic modeling using natural language processing (NLP) techniques
• <i>Craigslist</i> — www.craigslist.org	No study available	—	—

As revealed in Table 10.1, social media functions related to networking are more appropriate to address research questions about social networks. For example, researchers may use profile information on LinkedIn to explore how users of different professions present themselves on LinkedIn (Zide et al., 2014). Similarly, social media designed for commenting and discussion, such as Yelp, may allow researchers to examine the use of positive versus negative words in reviews of restaurants and shops (Gui et al., 2017). Additionally, some platforms, such as Facebook and Twitter, combine networking, microblogging, and commenting functions, which offers ample opportunity for research. The use of data from these multi-function social media is thus less restrictive than that of data generated from social media with a single function. Previous studies collected Twitter data to examine the relation between the usage of pre-identified vocabularies and health outcomes (Ireland, Chen, Schwartz, Ungar, & Albarracin, 2015; Ireland, et al., 2015) and harvested Facebook Likes data to predict dispositional characteristics (Kosinski et al., 2013; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones et al., 2013). As different harvesting methods yield distinct data characteristics, we next discuss the available harvesting methods.

Harvesting Social Media Data: Approaches and Sources of Data Collection

Social media data are available from a variety of sources, and the data can be collected via different approaches varying in cost, programming techniques required, and data completeness (see Figure 10.1). Figure 10.1 illustrates how six harvesting approaches relate to the data costs, the representativeness of the sample, and the subsequent cleaning and processing procedures. Researchers must carefully select which sources of data best suits their research needs. For example, researchers with limited resources or those who want to pilot test research ideas may use existing free datasets, whereas researchers who have sufficient IT resources and want to monitor the influence of policies in the public for a period, may use the services of monitoring vendors and/or set up application program interfaces (APIs).

Given the availability of multiple sources for each approach, we present the most common and up-to-date sources with their main features in Table 10.2. In the following section, we discuss harvesting social media data from the least expensive approach to the most expensive ones. However, the data access policy of social media platforms can change and thus requires researchers to check the social media regulations at the time of conducting their research.

The Least Expensive Approach: Downloading Free Datasets

The fastest and least expensive approach to accessing social media data is through sample data libraries and directly from the social media sites, such as Yelp, Wikipedia, and YouTube (see Table 10.2). These free datasets have several limitations

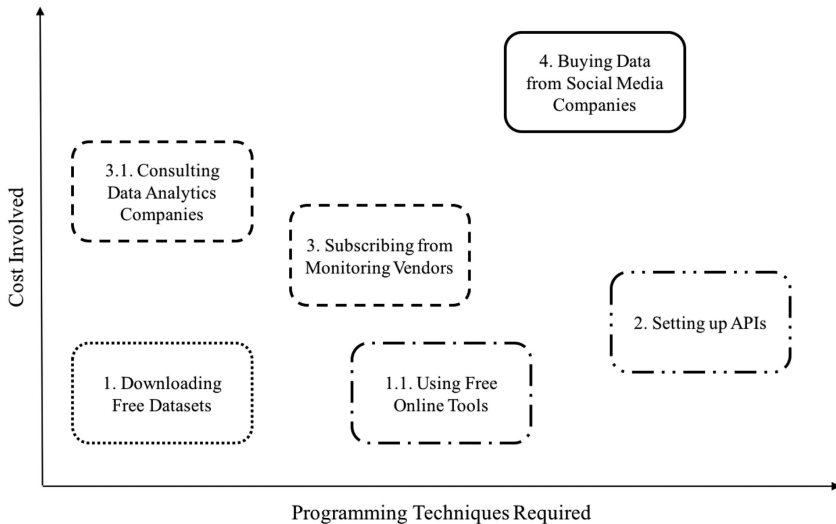


FIGURE 10.1 Various Approaches to Social Media Data Harvesting

Note: By cost, programming techniques and data completeness (solid line refers to the maximum level of completeness, long-dash-dot-dot lines refer to the moderate level of completeness, dash lines refer to the moderate-to-minimum level, and dot line refers to the minimum level of completeness).

because the data collection is already completed, limited user metadata are included, and the researchers have no control over the cleaning/preprocessing processes. Despite these limitations, researchers can use such data sets for validating techniques and pilot testing of their hypotheses.

Apart from these dataset repositories, researchers can also access free social media data (e.g., the Observatory on Social Media (OSoMe), developed by academic initiatives at Indiana University to promote public access to social media data). Moreover, Twitter Search, the search function available on the Twitter website, can also provide a small sample of Twitter data (i.e., retrieving up to seven days of historical data or 1,500 tweets). This method requires researchers to manually copy and paste search results into a database, which is cumbersome for a project examining an extended period.

The Less Expensive Approach: Setting Up Application Program Interfaces (APIs)

Social media platforms publish different APIs, which are sets of protocols and tools to enhance the functionalities of software applications developed by researchers. Researchers are required to register as a developer and obtain consumer and access token credentials to set up the API. Although using the APIs are free, there are tangible costs in setting up/monitoring the API and storing the data.

TABLE 10.2 A Summary of Approaches and Sample Sources of Social Media Data Collections

<i>Approaches</i>	<i>Sample Sources</i>	<i>Sample Social Media</i>	<i>Major Characteristics</i>	<i>Websites</i>
Download from Data Repositories	Stanford Large Network Dataset Collection	Facebook, Google+, Friendster, etc.	<ul style="list-style-type: none">• Datasets have different time periods• Each dataset has different attributes	https://snap.stanford.edu/data/#onlinecoms
	Social Computing Data Libraries Network Repository	Flickr, Twitter, YouTube, etc. Facebook, Twitter, YouTube, etc.	<ul style="list-style-type: none">• Datasets have different time periods• Each dataset has different attributes• Repositories donated by other users• Built-in interactive graph analytic tools for visualizing social networks	http://socialcomputing.asu.edu/pages/datasets http://networkrepository.com/
	myPersonality Project	Facebook	<ul style="list-style-type: none">• Match participants' self-reported questionnaires with Facebook data• Access to participants' Facebook profile and social network data via a Facebook application	https://www.psychometrics.cam.ac.uk/productservices/mypersonality
	Yelp Dataset Challenge	Yelp	<ul style="list-style-type: none">• Release annually by Yelp Co.• Datasets include information for a small number of cities and countries	https://www.yelp.com/dataset_challenge
	Wikipedia Database	Wikipedia	<ul style="list-style-type: none">• Download database directly from Wikipedia	https://en.wikipedia.org/wiki/Wikipedia:Database_download
Use Free Tools	YouTube-8M Dataset	YouTube videos	<ul style="list-style-type: none">• Databases are released on a regular basis• Videos are pre-processed and selected from a list of popular contents• Each video was tagged with labels	https://research.google.com/youtube8m/
	Observatory on Social Media (OSoMe)	Twitter	<ul style="list-style-type: none">• Allow access to about 1% of total public tweets since 2010• Offer a set of web-tools to study how information/ideas spread online	https://osome.iuni.iu.edu/tools/

Set-up APIs	Twitter APIs	Twitter	<ul style="list-style-type: none"> • Use Streaming and REST APIs to collect tweets and metadata (in JSON format) • Able to stream up to 1% of total public tweets • Rate limits of the REST APIs are applied. • Search a user, page, event, group, place, and topic (in JSON format) • Rate limits of the API are imposed on each page/group • Use Real-Time and REST APIs to collect contents and metadata (in JSON format) • Different rate limits are imposed to the modes of applications 	https://dev.twitter.com/overview/api https://dev.twitter.com/rest/public/rate-limiting
	Facebook Graph API	Facebook	<ul style="list-style-type: none"> • Specify API to collect data at various levels e.g., listings, live threats, and forums (subreddit) etc. (in JSON format) • Rate limit of 60 requests per minute is imposed 	https://developers.facebook.com/docs/graph-api/advanced/rate-limiting
	Instagram APIs	Instagram	<ul style="list-style-type: none"> • Use Weibo API to collect messages and metadata (in JSON format) • Rate limit of 150 requests per hour is applied • Use REST API to collect data of users and companies (in XML or JSON formats) • Rate limits are imposed to each application. 	https://www.instagram.com/developer/limits/ https://www.instagram.com/developer/limits/
	Reddit API	Reddit	<ul style="list-style-type: none"> • Use Weibo API to collect messages and metadata (in JSON format) • Rate limit of 150 requests per hour is applied • Use REST API to collect data of users and companies (in XML or JSON formats) • Rate limits are imposed to each application. 	https://www.reddit.com/dev/api/ https://github.com/reddit/reddit/wiki/API#rules
	Weibo API	Weibo	<ul style="list-style-type: none"> • Use Weibo API to collect messages and metadata (in JSON format) • Rate limit of 150 requests per hour is applied • Use REST API to collect data of users and companies (in XML or JSON formats) • Rate limits are imposed to each application. 	http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3_V2/en http://open.weibo.com/wiki/Account/rate_limit_status/en
	LinkedIn API	LinkedIn	<ul style="list-style-type: none"> • Use REST API to collect data of users and companies (in XML or JSON formats) • Rate limits are imposed to each application. 	https://developer.linkedin.com/legal/api-terms-of-use

(Continued)

TABLE 10.2 (Continued)

<i>Approaches</i>	<i>Sample Sources</i>	<i>Sample Social Media</i>	<i>Major Characteristics</i>	<i>Websites</i>
Subscribe Monitoring Services	You Tube API	You Tube	<ul style="list-style-type: none">• Use data v3 API to collect data of video, channel, or playlist (in JSON format)• Rate limit of 1 million units per day is imposed	https://developers.google.com/youtube/v3/getting-started
	Crimson Hexagon	Twitter, Facebook, Instagram, Blogs, Forums, News, Comments, Reviews, and YouTube, etc.	<ul style="list-style-type: none">• Create monitors by selecting sources of data and setting parameters/filters• Access to tweets since 2008 and export data in CSV format• Rate limits of 10,000 tweets per export and 50,000 tweets per day are imposed• Use built-in algorithms for sentiments, topics, and content analyses• Access to historical data and export data in JSON and CSV formats• Rate limit of 500,000 per day is imposed• Use built-in algorithms for identifying important keywords• Use built-in algorithms for sentiments, topics, and content analyses• Rate limits by types of subscription are imposed	https://www.crimsonhexagon.com/
	DataSift	bitly, Blogs, DailyMotion, Instagram, Facebook, Tumblr, and YouTube, etc.		http://datasift.com/
	Watson Analytics for Social Media	Twitter, Facebook pages, Forums, Blogs, Reviews, and Videos, etc.		https://www.ibm.com/us-en/marketplace/social-media-data-analysis/purchase
Buy Data	Gnip, Inc.	Twitter	<ul style="list-style-type: none">• Use APIs to collect real-time and historical tweets• Use APIs to obtain aggregate interest and demographic information for a collection of Twitter users	https://gnip.com/sources/

Furthermore, basic familiarity with programming techniques, as well as server side programming languages, are necessary for the use of APIs. For instance, researchers have to be familiar with Python, a programming language, to use Tweepy, an open-sourced python program, to communicate with the Twitter API python package (see http://docs.tweepy.org/en/v3.5.0/getting_started.html for details). Other intangible costs include the absence of retrospective data (because data are crawled prospectively) and the time required for data cleaning (because of missing fields and inconsistent information).

Apart from the accredited APIs, free web scraping programs available online supply tools to scrape information on designated websites and save into a JSON and XML format. These automated software programs (also referred to as bots) can also utilize fake user accounts to harvest data on social media.

Despite the availability, researchers should be cautious about the legal constraints of such tools. In 2016, LinkedIn filed a lawsuit against 100 unnamed individuals using bots to harvest user profiles from its website (Conger, 2016; LinkedIn, 2016). Web scraping tools are also subject to regulations (Bilton, 2012). For example, in 2015, the airline company Ryanair sued other travel agencies for screen-scraping price information. The Court of Justice of the European Union (ECJ) ruled that websites can set restrictions to limit scraping (Consonni & Anselmi, 2015). As the legitimacy of web scraping tools is bounded by the laws of respective countries, researchers should consult their institutions' legal services before scraping social media data.

The More Expensive Approach: Subscribing Services From Monitoring Vendors

Although the use of computer programs for harvesting involves concerns about technical and legal issues, subscribing services from monitoring vendors can make data retrieval, preparation, and basic analysis potentially easier (see Table 10.2). Monitoring vendors, such as Crimson Hexagon and DataSift, pre-process social media data, such as from Facebook, Weibo, Twitter, and provide information through automatic dashboards, real-time social listening and influencer identification tools, as well as built-in visualization tools (e.g., word cloud, and figures). However, a major drawback of such vendor services is the subscription cost, which may be very prohibitive depending on the retrieval volume and types of data. Furthermore, users can neither customize the algorithms of the built-in analyses nor modify any parameters of the machine learning model for analyses.

The Most Expensive Approach: Buying Data

The most expensive option regarding harvesting social media is buying data directly from the reseller. Gnip is a Twitter data reseller that provides the full raw Twitter data and sells the data to match the researchers' needs by customizing the

programming infrastructure and computational algorithms. Interested researchers can contact the reseller and purchase a dataset that meets specific needs, and there are occasional promotions and grants for academic work. Apart from Twitter data, to our knowledge, there are no other official resellers of raw social media data currently available. This approach certainly gives investigators complete control of how to retrieve, store, and analyze the full sample of Twitter data, but is infrequently used given the extensive cost and resources needed to build such a system.

Harnessing Social Media Data: Analytical Techniques for Non-Semantic and Semantic Features

Apart from the collection of social media data, a major challenge of using social media data for research is the selection of an appropriate analytical technique to measure the variables of interest. Social media data can be included in the analysis as measured variables or used to extract latent variables, depending on the characteristics. Figure 10.2 presents an overview of social media data, including two main features, i.e., non-semantic and semantic, and the corresponding analytical techniques. Non-semantic features include attributes of non-lexical items, such as age, gender, and location, which are usually specified on user profiles. Semantic features include lexical items with different levels of information, ranging from less detailed contents, such as Facebook Likes, to more detailed ones, such as text messages. As social media data vary in characteristics, the analytical techniques vary. For example, user attributes can be entered directly into a regression analysis whereas text messages require content analysis or natural language processing, followed by regression analyses. In the following sections, we reference published studies to illustrate how various analytical techniques can be used to research non-semantic and semantic features.

Non-Semantic Features

The first and most obvious application of social media analysis is to measure the demographic characteristics of populations. Social media data can reveal characteristics of the populations, especially those that are difficult to reach or less likely to participate in a survey. The majority of participants that are studied using traditional research methods are mainly white, female, Western undergraduate students, a.k.a. the “WEIRD” demographic described in Henrich, Heine, and Norenzayan (2010). Given that sampling directly from other regions of the globe and collecting responses from a national representative samples can be extremely expensive and time-consuming (Teitler, Reichman, & Sprachman, 2003), the analysis of social media data is likely to allow researchers to measure global populations on a larger scale, with a lower investment of money and time. The relatively low costs related to social media has enabled their use in everything from

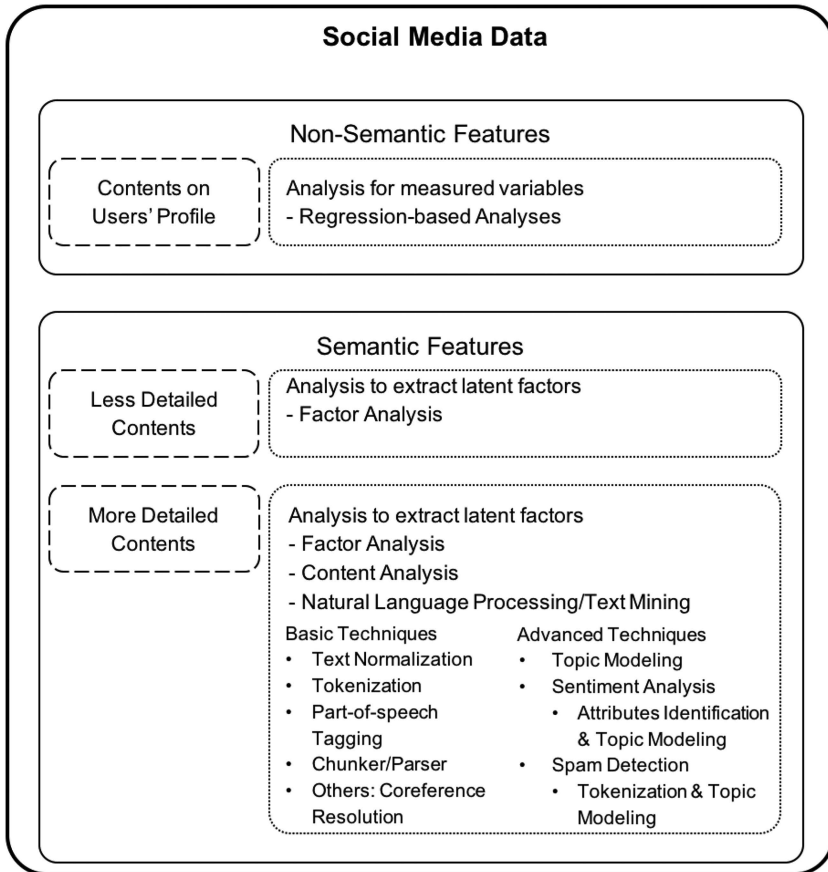


FIGURE 10.2 Data Features and Analytical Techniques of Social Media Analyses

the analysis of the effects of advertisements on consumer behavior, to those of government-led health campaigns on public opinion (Smyser, 2013).

Furthermore, the penetration of social media gives researchers the opportunity to examine broader research questions in which demographics (and other individual differences) can be systematically studied. Researchers can investigate the spread of emerging risk behaviors such as electronic cigarettes (e-cigarette) and vaping (Chu et al., 2015) in particular regions and age groups by analyzing social media non-semantic features such as location information specified on user profiles. Researchers can also examine the effectiveness of tobacco control campaigns in social media because lesbian, gay, and bisexual (LGB) populations are more likely to be smokers and social media users, compared to heterosexual ones

(Kostygina, Tran, & Emery, 2016; Seidenberg et al., 2017; Stevens, Carlson, & Hinman, 2004). Ultimately, researchers can sample diverse users from social media and use their responses to test theories that hypothesize variability on race, gender, education, culture, etc.—the variability unlikely to be found in “WEIRD” college samples.

In addition to the demographic information, researchers can quantify the size of users’ social networks by measuring the number of friends and followers on social media. Lönnqvist and Itkonen (2014) examined the mediating role of social network size on the link between personality traits and life satisfaction. Instead of asking individuals to report how many friends they have, Facebook friend counts can serve as a proxy for their social network size. Likewise, Johnston, Tanner, Lalla, and Kawalski (2013) used Facebook friend counts to gauge the levels of social capital and examined its impact on subjective well-being. Other researchers recorded changes in friendship ties on Facebook and MySpace as a measure of friendship selection, which was then linked to smoking and drinking behavior (G. C. Huang, Soto, Fujimoto, & Valente, 2014).

Although non-semantic features, including users’ demographic information and their social networks, are important for analyses, not everybody is willing to provide complete information on their social media profile. Only 20% of users provide demographic information and meaningful locations in their profile (Cheng, Caverlee, & Lee, 2010). Due to the sparseness of attributes in social media data, researchers have begun to use the available profile information to predict other missing user attributes including age, gender, ethnicity/race, location, language, and other demographic characteristics (S. Chang et al., 2014; Rao et al., 2011; Schwartz, Eichstaedt, Kern, Dziurzynski, Agrawal et al., 2013; Zamal, Liu, & Ruths, 2011). Previous studies have demonstrated satisfactory performance of these predictors and classifiers, even though attribute identification tasks are still resource-intensive due to the use of manual annotation procedures. For example, previous studies relied on users’ first name on their account profiles to infer gender (Burger, Henderson, Kim, & Zarrella, 2011), even though the accuracy of this method is not well validated.

Semantic Features

Less Detailed Contents

Other studies have examined semantic data (e.g., Facebook Likes) to predict personal attributes, personality traits, and psychological outcomes (Kosinski et al., 2013; Wu et al., 2015). Table 10.3 summarizes common semantic data with less detailed contents that are available in the top three social media (see the left side of the table). Favorite/like, follow, and share/retweet are the examples of semantic data that work similarly as web browsing cookies: Clicking Favorite/Like for a message indicates users’ positive evaluations of that post, clicking Share/Retweet

TABLE 10.3 Examples of Less-Detailed and More-Detailed Semantic Features on Facebook, Twitter, and Instagram and Possible Research Questions

<i>Sample Media</i>	<i>Less-Detailed</i>		<i>More-Detailed</i>	
	<i>Features</i>	<i>Possible Research Questions</i>	<i>Features</i>	<i>Possible Research Questions</i>
Facebook	Post reactions	How do post-reactions (i.e., clicking Like, Love, Haha, Wow, Sad, and Angry) towards messages link to voting preferences?	Individual posts	What are the levels of satisfaction with a product?
	Follow/like	How do <i>Likes/Followings</i> (i.e., clicking Like or Follow) of pages/groups relate to dispositional characteristics?	Post conversations	What are the sentiments of a specific event?
	Share	How do <i>Sharing</i> (i.e., clicking Share) of messages on his/her Facebook timeline relate to attitudes, beliefs, and behaviors?		
Twitter	Favorite/like	How do <i>Likes</i> (i.e., clicking Like) of tweets link to related donation campaigns?	Individual tweets	What are the topics discussed in the community?
	Follow	How do <i>Followings</i> (i.e., clicking Follow) of other user accounts relate to mental and physical well-being?	Tweet conversations	What are the important factors in discussion of HIV prevention?
	Retweet	How do <i>Retweets</i> (i.e., forwarding messages) relate to the perceptions of public health crisis?		
Instagram	Like	How do <i>Likes</i> (i.e., clicking Like) of posts link to music/movies preferences?	Photo captions	What are the factors that suggest positive dyadic relationships?
	Follow	How do <i>Followings</i> (i.e., clicking Follow) of user accounts relate to social norms?		

involves forwarding a message posted by other users, and clicking Follow shows users' choice of receiving all updates from that page/group. Even though these semantic features are minimal or condensed, they are useful for examining a wide range of research questions (see Table 10.3). For example, Kalampokis, Tambouris, and Tarabanis (2013) used Facebook Likes data to develop machine learning models to predict personal attributes, going from sexual orientation to intelligence. Wu et al. (2015) further validated the predicted personality scores and revealed that computer-based personality predictions, rather than the estimates made by the participants' Facebook friends, were more highly correlated with participants' self-report scores. Semantic features such as Likes and the related analytical techniques are likely to have a major influence on psychological research in the next decade. Apart from Likes/Follow, researchers can collect users' Share/Retweet as clear expressions of particular events and apply machine learning techniques to predict psychological variables without asking participants to complete self-report questionnaires. The collection of semantic data and the corresponding analyses tend not to be limited to particular social media, with a caveat that Facebook frequently changes APIs for public access to their contents, which creates uncertainties about Facebook as a stable data source.

More Detailed Contents

An expanding body of research has concentrated on the content analysis of semantic features with more detailed contents, such as posts and messages on social media (see Table 10.3 for examples; Curini, Iacus, & Canova, 2015; De Souza & Ferris, 2015). Such messages and posts may include emoticons, which are the use of keyboard characters to represent a facial expression, such as a smile “:-)”, and text content that can be used to directly reveal a user's emotion. Researchers can either use tweets originally codified by Twitter as happy versus unhappy for valence analyses (Curini et al., 2015) or analyze the message content to obtain verbal information. As described in Table 10.3, the analyses of individual messages posted on social media allow marketing campaigners to understand the level of satisfaction with a product (De Souza & Ferris, 2015). Using social media data to measure customer satisfaction resembles the collection of product comments in focus groups, except that the online customers can participate in the product review meeting whenever and wherever they want. Additionally, the general usage of certain words can also reflect an individual's emotions, thoughts, and behaviors. Therefore, an emerging field uses social media data to infer users' behaviors, attitudes, and health status. For example, a study conducted by Asur and Huberman (2010) showed that Twitter data could predict how many tickets would be purchased for the upcoming release of a movie. These findings indicate that the analysis of social media to derive semantic features is likely to provide valuable insights.

Scientists have been studying how to convert raw text and its representations into manageable inputs for computers since the early 1960s. Natural language processing (NLP), which aims to understand human communications using computers, has allowed researchers to extract meaningful representations from text messages (i.e., words, phrases, and sentences) and use them as inputs for machine learning models (see Figure 10.2). The major use of NLP research once concentrated on deriving representations from structured text passages in formal written language, such as news articles, academic journal articles, records, and archives. However, as social media data have become increasingly available, NLP techniques have evolved to analyze the short and unstructured user-generated message contents that characterize posts and messages on social media. The most well-established basic NLP techniques include text normalization, tokenization, part-of-speech tagging, chunkers and parsers, as well as named-entity recognition. Other basic NLP methods that have not yet received much attention in social media analysis include coreference resolution. These new technologies are attempts to respond to the challenges of understanding user-generated content on social media, such as identifying HIV risk among users (Thangarajan, Green, Gupta, Little, & Weibel, 2015) or predicting crime rates using Twitter data (Gerber, 2014).

The first step in applying natural language processing is text normalization, which is an abstraction used to convert raw text into a standardized representation. This step involves some knowledge of the data available and how it will be utilized. For example, Harrison et al. (2014) have collected restaurant reviews in which customers have described various aspects of each restaurant such as location, food quality, atmosphere, and price. If a researcher interested in analyzing price information may find some customers using “\$” to describe the monetary price while others might use the word “dollars,” which requires consolidating different representations into one norm. Researchers can, of course, substitute numerical characters for respective words. Similarly, there are methods for word stemming (e.g., maps the texts *car*, *cars*, *car’s*, and *cars’* to *car*), stop words removal (e.g., removes words like *a*, *an*, and *are*, etc.), and lower-case conversions (e.g., converts *Health* to *health*). However, the adoption of these methods in social media analysis requires consideration of informal language use, idiosyncratic writing styles, and vernacular orthography (e.g., *that* as *dat*; Beckley, 2015). Tweets may signal emphasis with capitalization, which is traditionally used for the starting boundary of a sentence or some named entity. Furthermore, tweets contain punctuations that are used not just to end a sentence, but also as a part of emoticons (Kaufmann, 2010).

The second step in text preprocessing is text tokenization, which reduces raw texts to a number of basic units, typically in the form of words, phrases, sentences, and/or paragraphs (see Figure 10.2). For instance, an *n*-grams tokenizer breaks the text down into a contiguous sequence of *n* items such as words; an *n*-gram of size one refers to as a unigram, and an *n*-gram of size two refers to as a bigram,

etc. The tokenizers also segment sentences into valid partitions. For example, the punctuation period “.” usually indicates the end of a sentence, although applying such a rule to a sentence with a term “U.S.A.” may lead to incorrect segmentation, resulting in meaningless text fragments. In this situation, a text tokenizer would decode the word set correctly into “the United States of America.” This example suggests the need for more analytic tools to tackle the challenges of informal language (Gimpel et al., 2011; Ritter, Clark, Mausam, & Etzioni, 2011). Another basic form of syntactic analysis can be derived from identifying the part-of-speech (POS) components of a sentence (i.e. nouns, verbs, adjectives, etc.). Although many POS taggers and tokenizers are trained using a standard corpus (the Wall Street Journal corpus), Gimpel et al. (2011) developed a Twitter POS tagger and tokenizer tool, which creates an appropriate annotation corpus for the training of the text preprocessing tool. The importance of Gimpel and colleagues (2011) tool lies in the invention of phonetic normalization, which derives a common representation of a word that receives many alternate spellings on Twitter.

The third NLP step is to identify some structure in texts, that is, parsing grammatical components of sentences, such as noun, prepositional, or verb phrases (see Figure 10.2). This goal is achieved by parsers, an umbrella term for fully grammatical parsers and shallow parsers/chunkers. The challenge of identifying structures in texts is that very few structures exist, not to mention the presence of large amounts of noise. Hence, parsers developed for Twitter typically perform less accurately than tools developed for news articles or journals (Kong et al., 2014). Named entity recognition (NER) is another process of identifying and categorizing tokens that refer to people, locations, organizations, etc. NER may be useful when a researcher tries to identify tweets about the World Health Organization (WHO), a case in which the keyword search “WHO” is likely to return noisy results. In that case, tweets can be further processed with NER to identify correct tweets, but currently, this process only works for tweets with sufficient textual content, i.e., the larger the number of characters the better the performance (Ritter et al., 2011).

In addition to the well-established NLP techniques, we present recent NLP techniques that have not yet been widely applied to analyze social media data but might improve analysis in the future (see Figure 10.2). Coreference resolution is a basic NLP technique that involves identifying noun phrases and clustering those that refer to the same named entity (K. Chang, Samdani, & Dan Roth, 2013). Despite the availability of various techniques, their performance at correctly identifying referents depends on the presence of structure or context, both of which are limited in Twitter and other social media. To improve coreference resolution methods, scientists have begun research in cross-document coreference resolution to identify if two mentions refer to the same concept (Upadhyay, Gupta, Christodouloupoulos, & Roth, 2016). Alvarez-Melis and Saveski (2016) have proposed an interesting approach to overcome the limited content issue by keeping track of the conversation on Twitter and aggregating the tweets replying

to the original tweets. Such an approach is likely to gather more tweets that meet the needs of NLP methods and generate more accurate results (Alvarez-Melis & Saveski, 2016).

Given the unique writing style and sentence structure of posts and messages on social media, scientists are actively developing new techniques to address such challenges, leading to steady progress in the advancement of NLP research on social media data. There are many collective efforts and conferences, such as the Workshop on Semantic Evaluation (SemEval), the Text Retrieval Conference (TREC) and the Workshop on Noisy User-generated Text (WNUT), which are dedicated to advance the state-of-the-art (to improve the performance) in text normalization, tokenization, named entity recognition, and other methods for Twitter and other media (Baldwin et al., 2015). In the realm of measuring social media data, NLP techniques can serve both the purpose of language identification and the less attended problem of improving data quality. In the next section, we present other advanced NLP techniques (i.e., topic modeling for text mining, sentiment analyses, and spam detection), which can be incorporated to identify meaningful semantic features and improve data quality for further language identification and analysis.

Topics as Important Semantic Features

Topic modeling is widely used to cluster semantically similar words that frequently co-occur in a collection, and each cluster refers to a topic, which corresponds to a different distribution of words. Among the most popular methods for discovering topic models are Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003; Hofmann, 1999). These topic models adopt hierarchical Bayesian networks that do not require labeled training data and are able to identify topics (i.e., latent semantic features) in an unsupervised fashion. LDA assumes that the documents contain a mixture of topics and that each topic includes a list of words based on their probability distribution. LDA attempts to figure out what topics emerge in a particular set of documents. It is a matrix factorization technique, and the corpus (a collection of documents) can be represented as a document-term matrix. The corpus has N documents D_1, D_2, \dots, D_n and a vocabulary size of M words $W_1, W_2 \dots W_m$. We can apply the LDA model for converting such a document-term matrix into two lower dimensional matrices: M_1 is a document-topics matrix with dimensions (N, K) and M_2 is a topic-terms matrix with dimensions (K, M) , where N is the number of documents, K is the number of topics, and M is the vocabulary size. Although these two matrices provide the distributions of topic word and document topic, such distributions need further improvement by making use of sampling techniques. Therefore, LDA, for example with Gibbs sampling, iterates through each word for each document and tries to adjust the current topic-word assignment with a new assignment (Gilks, Richardson, & Spiegelhalter, 1996).

A steady state, or convergence point, is achieved with satisfactory distributions of the document topics and topic words after multiple iterations. The identified topic model captures topic proportions and assignments as well as the weights of each word in a specific topic in each document (i.e., the measurement unit).

Topic models can help to organize and offer insights about large collections of unstructured text messages. Consider an analysis of Facebook to identify popular topics. In this example, we used the Python package *scikit-learn* (other packages are also available, see <https://pypi.python.org/pypi/lda> for details). Furthermore, the topic modeling analysis can be performed in R and other computer languages based on available resources and familiarity with the programming environments. We first collected data and prepared the documents, that is using Facebook API to collect posts on the Society for Personality and Social Psychology (SPSP) Facebook page from November 10 to December 11 in 2016. As the SPSP page is a public page where subscribers can freely post messages, the sources of messages varied from mainstream news media sites to specific research-oriented outlets. The top five sources of messages include “The Wall Street Journal,” “The New York Times,” “The Atlantic,” “VOX,” and “Washington Post,” all traditional rather than academic media. Second, we used the Python package *scikit-learn* to remove all stop words (e.g., and, the, is, etc.) and then tokenized the corpus into bigrams (i.e., a sequence of two adjacent words). Next, we converted the bigrams into a document-term matrix using the built-in function of the package, created an object for the LDA model, and trained it on a document-term matrix. We set a few parameters as required in the training (see Appendix 10.1 for the sample codes). Finally, from the training corpus, we identified an LDA model that could be used to discover topic distributions of posts on other Facebook pages (i.e., new and unseen documents). Figure 10.3 shows first five topics with top-20 words (due to limited of space) that were identified in this example.

Identification of Sentiments

Social media has become a unique platform for individuals to express their opinions and is a valuable source for researchers to examine attitudes in diverse areas. However, the size and the complexity of the social media data require the development of automatic methods for organizing, analyzing, and extracting attitudes. The main objective of sentiment analysis is to identify attitudes (either positive or negative) in a corpus (i.e., a collection of documents, and each document is a unit of measurement). Sentiment analysis varies in scope, ranging from the document- and sentence-level to the aspect-levels. In the following paragraphs, the discussion focuses on the aspect-level analysis, which first extracts attributes (aspects) of the object and then identifies the sentiments of those attributes (Farhadloo & Roland, 2013; Hu & Liu, 2004; Popescu & Etzioni, 2005; Su et al., 2008).

In recent years, different text mining techniques have flourished to extract attributes (i.e., attitudes) of the object. A group of researchers has proposed automatic



Word Cloud 3



Word Cloud 2



Word Cloud 1



Word Cloud 4



Word Cloud 5

FIGURE 10.3 Top Five Topics Identified Based on Posts of the SPSP Page on Facebook

Note: The size of each word does not represent the relative weight in each topic.

methods, such as an aspect-based summarization model (Blair-goldensohn et al., 2008) to discover attributes, whereas others have used (semi) automatic methods with the same goal. For example, Hu and Liu (2004) used association mining in a combination of pre-identified adjectives with known positive/negative orientations to identify frequent (vs. infrequent) attributes: i.e., how likely are people to talk about those aspects? Other researchers have proposed the use of clustering to extract attributes in a hierarchical manner (Gamon et al., 2005) and the use of nouns to improve the clustering results for attributes identification (Farhadloo & Rolland, 2013).

In the process of identifying sentiments, researchers have mainly used a close-vocabulary approach to reveal the polarity of opinions of text fragments (Andreevskaia & Bergler, 2006; Esuli & Sebastiani, 2006; Hu & Liu, 2004; Subasic & Huettner, 2001; Wiebe, 2000). The close-vocabulary approach involves the use of a list of words (pre-identified terms) as a priori to examine the sentiment, and the presence of such words determines the sentiment polarity. The use of dictionaries words/terms is not limited to supervised learning but is also found in unsupervised learning. Turney (2002) has introduced an unsupervised technique that examines the number of occurrence and co-occurrences between two pre-identified terms and words found via the web search engine. For example, a term that frequently appears with the term “excellent” (a pre-identified positive term) is considered as positive whereas another term that often appears with the term “poor” (a pre-identified negative term) is considered as negative. Whereas a group of researchers identifies the sentiments by measuring the frequencies of specific words/terms (i.e., a regression problem), other researchers consider the sentiment identification as a classification problem (i.e., the presence/absence of features). Different classification techniques have been introduced to identify sentiments (Gamon et al., 2005; Lakkaraju, Bhattacharyya, Bhattacharya, & Merugu, 2011; Moghaddam & Ester, 2012; Pang, Lee, & Vaithyanathan, 2002), and the reliability of these techniques depends on the quality of the features revealed in the process. Hence, recent work has attempted to develop new computing techniques and algorithms, such as a score representation of positivity, negativity, and neutrality as new features (Farhadloo & Rolland, 2013), and a hierarchical deep learning framework (Lakkaraju, Socher, & Manning, 2014).

In addition to the close-vocabulary approach, a topic modeling, which attempts to identify attributes and sentiments simultaneously, is also frequently adopted for the analysis of sentiments. Topic modeling uses probabilistic methods to discover aspects and their associated sentiments at the same time. Topic modeling algorithms can distinguish between attribute-topics and sentiment-topics and determine the probability distribution of each term within a particular topic. One of the main advantages of such topic models as hierarchical Bayesian networks is that they do not require labeled training data and find the topics by analysis of the original collection of documents. As explained in the previous section, Latent Dirichlet Allocation (LDA) assumes the presence of a mixture of topics in each

document (Blei et al., 2003). In the case of sentiment analysis, when individuals talk about an attribute of an object, they are likely to use different terms. Likewise, individuals tend to use various terms to indicate a particular sentiment of that attribute. For instance, “excellent,” “fabulous,” and “extraordinary” are used to suggest a higher level of positive sentiments among individuals. Therefore, each sentiment-level can be considered a topic in topic modeling (see Brody, 2010; Farhadloo, Patterson, & Rolland, 2016 for further details).

Detection of Spam

Detecting spam within social media is a classic problem and is useful in many areas, including consumer, health, political, and social psychology. Although email spam is relatively easy to identify using unigrams or bigrams as input features for machine learning models, spam in social networks can take different forms (e.g., advertising spam, opinion spam, and deceptive opinion spam) and is therefore challenging. In the context of reviews, messages that do not include any opinions, but instead market products/services, are considered as advertising spam or duplicate spam. The detection of this type of spam is relatively easy (Jindal & Liu, 2008). Deceptive opinion spam is defined as “fictitious opinions that have been deliberately written to sound authentic” (Ott, Choi, Cardie, & Hancock, 2011). Some companies hire large numbers of users to post fake, and sometimes malicious, reviews or posts (Wang, Wang, Zhai, & Han, 2011). The detection of this type of spam is more challenging and requires data-driven models to pinpoint anomalous user behaviors (Lim, Nguyen, Jindal, Liu, & Lauw, 2010).

To detect opinion spam in the text, available methods include obtaining basic semantics features (e.g., n-grams) and identifying advanced semantic features (e.g., topics model). Character-level n-grams can be developed to effectively deal with the mistakes, typos, and errors in spelling that are quite common in social media but difficult to detect. Previous research has shown that using these character-level n-grams as features can improve the classification of news articles (Cavnar, Trenkle, & Mi, 1994). Others have demonstrated an 80% accuracy by using unigrams as simple features to identify individuals’ race and ethnicity (Mohammady & Culotta, 2014).

In the area of detecting opinion spam, Ott et al. (2011) found that n-gram based text categorization best identified the opinion spam whereas a combined classifier with both n-grams and psycholinguistic deception features, i.e., terms obtained from the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) performed only slightly better than the former method. Furthermore, n-grams are used within language models for spam detection. Language models are probability distributions over units, where units can be anything from words, phrases, n-grams, or characters. In categorization tasks including text classification, a common method is to develop a language model for each category. For example, messages may be labeled as spam or ham (not spam) by

developing a spam language model and a ham language model. A product review may then be analyzed to determine the probability that it was generated from the spam or the ham language model (Sun, Morales, & Yan, 2013). This approach for using language models assumes that the text in the different categories uses the same words or phrases (“click,” “here,” “online,” “cheap,” etc.) or shares features that can be classified with appropriate models. This line of work has achieved successful spam detection, with a nearly 90% accuracy in spam detection (Ott et al., 2011). Nonetheless, recent research has found that a devious adversary can synthesize faked reviews by using similar data-driven methods (Sun et al., 2013; Tran, Hornbeck, Ha-Thuc, Cremer, & Srinivasan, 2011).

More sophisticated methods such as topic modeling may be more successful in detecting content because no specific words are predetermined in the process. As explained above, a topic model includes a number of topics in which each topic corresponds to a different distribution of words. Therefore, it is widely used to infer latent variables of words that frequently co-occur in a collection. Topic models have been applied to a host of problems, including TopicSpam, a topic modeling approach to identify deceptive opinion spam (Li, Cardie, & Li, 2013). However, topic modeling assumes a bag-of-word (BOW) representation that disregards the word order in the text and requires a sizable corpus to discover meaningful and interpretable topics (for alternate methods, see J. Chang, Gerrish, Wang, & Blei, 2009).

Using Social Media to Obtain Inferences

Predictive and Explanatory Models

An emerging field has used social media data to investigate public health challenges such as influenza infections and sexually transmitted infections, including HIV, chlamydia, and gonorrhea (Chan et al., 2018; Ireland et al., 2015; Young, Rivers, & Lewis, 2014). These studies have either developed a predictive model or an explanatory model. A predictive model is often bottom-up, open vocabulary without predetermined features, whereas an explanatory model is often top-down or closed-vocabulary with pre-established dictionaries of terms/phrases (Pennebaker, Mehl, & Niederhoffer, 2003). However, previous research has demonstrated the use of a closed-vocabulary approach for predicting influenza outbreaks (Santos & Matos, 2014; Signorini et al., 2011) and investigating links with HIV prevalence (Ireland, Schwartz et al., 2015; Young, Rivers, & Lewis, 2014). Potential challenges of closed-vocabulary methods include people’s reluctance to discuss stigmatized conditions (e.g., HIV) or behaviors (e.g., drug use) online. Another limitation is that social media communications are informal and constantly evolving as a function of users’ needs, culture, and idiosyncrasies (Gouws, Metzler, Cai, Hovy, & Rey, 2011).

An open-vocabulary approach can be used for prediction and explanation. Two major available methods differ in the degree to which they use predetermined terms to limit the collection of tweets: (a) a partial method, that is, including only tweets with pre-established dictionaries of terms/words, and (b) a full method, that is, including tweets without filtering by dictionaries. The partial method is likely to obtain more interpretable (explanatory) latent factors, whereas the later one can identify predictive factors relevant to users' needs, culture, and idiosyncrasy. As each method has its strengths and weaknesses, a mixed method is optimal for maximizing the predictability while improving the interpretability of latent factors that are identified on Twitter. For example, we used a Twitter application program interface (API; Garden Hose) to obtain about 10% random sample of all tweets in 2009–2010 and a Twitter streaming API to obtain approximately 1% of its publicly available stream in 2011–2012. We used the time metadata to exclude tweets not originating from U.S. time zones, and combined users' profile location with each tweet's precise coordinates to map tweets to U.S. counties. At the same time, we obtained the available county-level data on HIV prevalence and new diagnoses from the Centers for Disease Control and Prevention and AIDSVu (<http://aidsvu.org/>).

We first carried out an extensive search of research articles, news reports, as well as public health and slang dictionaries to identify a list of relevant terms and phrases of HIV and sexually transmitted infections (STIs). We identified 15 sources from various research teams in psychology and language processing, and together with the public health experts from the Health and Social Media Group at the University of Illinois at Urbana-Champaign, we devised nine categories that are related to HIV/STIs, including (a) HIV including treatment, (b) HIV and STI prevention, (c) drugs and alcohol, (d) other STIs, (e) sex, (f) men who have sex with men, (g) full-service sex work, (h) sexual violence and abuse, and (i) runaway youth. We collected words and phrases for these categories by incorporating prior dictionaries about sex and risky behaviors (Ireland et al., 2015), by using topic-specific glossaries (e.g., Drugs.com, 2013: HIV prevention measures), and by referring to slang databases (e.g., Urban Dictionary). The dictionaries contain 510 words.

Analytical Procedures and Results

Three methods of the open-vocabulary approach were assessed, and the major difference among these methods is the way for which tweets are prepared. For the partial method, we used the pre-established HIV/STIs dictionaries to filter out tweets that did not include one of the terms/words. For the full method, we included all tweets into the analyses. For the mixed method, we used the word embedding techniques to develop a lexicon of HIV and then included the lexicon as a prior in the machine learning model. Altogether, we had three sets of tweets,

and each was converted into a matrix of the token count. We then used the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to identify a model and to automatically discover topics (i.e., latent factors) in a collection of documents. We examined the distributions over words in each document and identified two hundred topics, then using the extremely randomized tree regressor method (Geurts, Ernst, & Wehenkel, 2006) to rank topics that associated with the new HIV diagnoses rates. We learned three topic models and evaluated the performance of each model by obtaining the topics probabilities of the 2012 tweets based on the word distributions and using the topic coefficients to compute the predicted 2012 new HIV diagnoses rate for each county. We then correlated the predicted 2012 HIV rates with the observed ones reported by the CDC to compare the performance.

Table 10.4 presents two model-fit indicators of models for three methods. By definition, a model with more predictive latent factors should yield a higher correlation and a lower mean squared error with the observed outcomes than the other models. As shown in Table 10.4, the proposed mixed method had the highest correlation coefficient and the lowest mean squared error among three methods. The results of different ethnicity representations were consistent with each other, indicating that the mixed method is likely to identify factors that explain the largest amount of variance in HIV prevalence rates. Apart from the numeric indicators, we also compared topics that were identified by different methods. The top three topics were selected from areas with higher and lower ethnic-minority representation. In general, the partial method identified latent factors with more words/terms about sex and drugs compared to other methods and revealed both norms about specific risk behaviors and general risk-taking notions. The partial method is likely more appropriate for detecting the presence of specific risk behaviors whereas the full and mixed methods are likely important for researchers to identify broader norms or perceptions linking to HIV risks in the communities.

TABLE 10.4 Results of Model-Fit Analyses Among Three Methods of the Open-Vocabulary Approach

<i>Models</i>	<i>df</i>	<i>Partial Method^c</i>		<i>Full Method^d</i>		<i>Mixed Method^e</i>	
		<i>r</i>	<i>MSE</i>	<i>r</i>	<i>MSE</i>	<i>r</i>	<i>MSE</i>
Ethnicity representation 1 ^a	2,596	.37***	0.95	.41***	0.83	.47***	0.76
Ethnicity representation 2 ^b	2,768	.29***	0.94	.46***	0.78	.51***	0.72

Note: a = percentages of black population; b = percentages of white population; c = model-fit analyses based on tweets with filtering; d = model-fit analyses based on tweets without filtering; e = model-fit analyses based on including the HIV lexicon as prior; df = degree of freedom; *r* = correlation coefficients; MSE = mean squared errors.

*** < .001.

Ongoing Challenges and Concluding Notes

As a whole, social media data characterize of high spatial resolution (i.e., with an extensive coverage of geographical areas), the location information of individual users and of their contents becomes an important non-semantic feature for researchers to address questions of differences in areas (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Adomavicius & Tuzhilin, 2015; Kalampokis, Tambouris, & Tarabanis, 2013; Mohammady & Culotta, 2014). For example, Mohammady and Culotta (2014) developed a model to predict each Twitter user's ethnicity/race based on the ethnicity/race makeup of tweets that clustered by county. Recent work has also combined geo-mapping techniques with the analysis of social media data to detect terrorism and predict presidential elections (Cody, Reagan, Dodds, & Danforth, 2016; Cohen, Johansson, Kaati, & Mork, 2014). Although location identification is a key, the geo-mapping/geo-tagging of social media data at the user- and message-levels is far from simple (Eisenstein, O'Connor, Smith, & Xing, 2010; Han, Cook, & Baldwin, 2014). Given the growing concern with online privacy and cyberstalking (A. L. Young & Quan-Haase, 2009), less than 2% of social media users enable the GPS functionality (Ireland, Schwartz et al., 2015), and about 26% American teenagers fake their online information, including name, age, or location (Madden et al., 2013).

The limitation and sparseness of location information on social media have become a driving force in geo-mapping research, and different methods have been proposed to identify users locations (Cheng et al., 2010; Eisenstein et al., 2010; Schwartz, Eichstaedt, Kern, Dziurzynski, Agrawal et al., 2013). Schwartz et al. (2013) have proposed a rule-based mapping method, which uses information about the location and coordinates available in the metadata to map each post/message to a county. This method relies on either the coordinates information attached to a tweet/Facebook post (latitude, longitude) or the free-response location information in the users' profile on social media. As reported in recent studies, about 15%–20% of tweets could be mapped to U.S. counties. The percentage depends on the selection/inclusion criteria of tweets (Chan et al., 2018; Eichstaedt et al., 2015; Ireland et al., 2016). Another group of scientists has suggested text-based geo-mapping that uses users' time zones, the number of followers/friends, and/or text messages for location prediction (Cheng et al., 2010; Eisenstein et al., 2010; Roller, Speriosu, Rallapalli, Wing, & Baldridge, 2012). Previous studies have demonstrated that text messages alone with neural network models can predict users' locations, from fine-grained coordinates to regions such as states (Cha, Gwon, & Kung, 2015; Han et al., 2014; Liu & Inkpen, 2015). The state-of-the-art performance is about 42% accuracy for the states prediction (Cha et al., 2015) and 50% for coordinate prediction, with a tolerance of about 161 km (Wing & Baldridge, 2014). The performance of such text-based geo-mapping techniques is subject to several factors, including the choice of activation functions, the number

of neurons per layer, initialization and regularization affects performance on predicting the actual geographical user coordinates, and classifying users per state or region (Morales et al., n.d.). Further work is required until this line of research can be used for “neural geotagging.”

Another challenge with language identification of social media data is code-switching, which is the interchanging of different words in different languages in text messages. Recent work has used neural networks models, a popular classifier which automatically creates higher order representations of the input features for language identification (J. C. Chang & Lin, 2014). The code-switching makes it particularly challenging for tasks such as sentiment analysis, which typically assumes a single language and narrative (Vilares, Alonso, & Gómez-Rodríguez, 2016).

Social media is a unique data source that is worth exploring. Researchers can analyze a wide range of social media data, from demographic information, personal attributes, and location information, to various forms of messages, to determine characteristics of populations, investigate beliefs and attitudes, and ultimately understand behaviors. The widespread use of social media renders social media analysis more generalizable than results produced through conventional self-report methods with convenience samples. Furthermore, individuals and populations that are inherently difficult to reach due to lack of representation in academic settings may be more easily studied through social media analysis. The power and reach of social media analysis makes it a staunch ally to the contemporary researchers in social psychology and its allied sciences.

APPENDIX 10.1

Sample python codes of the topic modeling analysis:

```
### Import packages
from glob import glob
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.decomposition import
LatentDirichletAllocation
from nltk.corpus import stopwords
### Get the social media data file
text_data = glob('facebook_data/*.txt')
### Convert the data into bigrams and remove stopwords
cv = CountVectorizer(input='filename', ngram_range=(2,
2), stop_words=stopwords.words('english'))
### Transform the vocabularies into a matrix
X = cv.fit_transform(text_data)
### Performe the LDA topic modeling
lda = LatentDirichletAllocation(n_topics=15, max_
iter=100, random_state=42)
model = lda.fit_transform(X)
### Create a function to print out the outputs
def print_top_words(model, feature_names, n_top_words):
for topic_idx, topic in enumerate(model.components_):
print("Topic #%d:" % topic_idx, ",
".join([feature_names[i]
```

```

for i in topic.argsort()[::-n_top_words-1:-1]])
print()
### Print the topics with the first 20 words
feature_names = cv.get_feature_names()
print_top_words(lda, feature_names, 20)

```

References

- Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). *Predicting flu trends using Twitter data*. 2011 IEEE conference on computer communications workshops. Shanghai, China: IEEE, pp. 702–707. Retrieved from <http://doi.org/10.1109/INFCOMW.2011.5928903>
- Adomavicius, G., & Tuzhilin, A. (2015). Context-aware recommender systems. *Recommender Systems Handbook* (2nd ed., pp. 191–226). Retrieved from http://doi.org/10.1007/978-1-4899-7637-6_6
- Alvarez-Melis, D., & Saveski, M. (2016). *Topic modeling in Twitter: Aggregating tweets by conversations*. The Tenth International AAAI Conference on Web and Social Media, (IcwsM), pp. 519–522. Cologne, Germany: Association for the Advancement of Artificial Intelligence. Retrieved from [evernote://view/779439927/s24/8594e3b8-85b2-4f4c-8fc8-1f4b55e818cb/](http://evernote://view/779439927/s24/8594e3b8-85b2-4f4c-8fc8-1f4b55e818cb/8594e3b8-85b2-4f4c-8fc8-1f4b55e818cb/)
- Anderson, L. (2013). HIV prevention. Retrieved June 11, 2018, from <https://www.drugs.com/aids-preventative.html>
- Andreevskaia, A., & Bergler, S. (2006). *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. Annual Meeting of The European Chapter of The Association of Computational Linguistics (T. 6, pp. 209–216). Association for Computational Linguistics: Trento, Italy. Retrieved from <http://doi.org/10.1.1.60.8316>
- Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (T. abs/1003.5, pp. 492–499). Toronto, Canada: IEEE. Retrieved from <http://doi.org/10.1109/WI-IAT.2010.63>
- Baldwin, T., Kim, Y.-B., de Marneffe, M. C., Ritter, A., Han, B., & Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *The ACL 2015 Workshop on Noisy User-generated Text*, (pp. 126–135). Beijing, China: Association for Computational Linguistics. Retrieved from <https://noisy-text.github.io/>
- Beckley, R. (2015). *Bekli: A simple approach to Twitter text normalization*. The 53rd Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2015) (p. 82). Beijing, China: Association for Computational Linguistics
- Bennett, J., & Lanning, S. (2007). The Netflix prize. *KDD Cup and Workshop*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6998>
- Bilton, N. (2012). *Disruptions: Innovations snuffed out by Craigslist*. 2017 m. sausio 30 d. Retrieved from https://bits.blogs.nytimes.com/2012/07/29/when-craigslist-blocks-innovations-disruptions/?_r=0
- Blair-goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., McDonald, R., & Reynar, J. (2008). *Building a sentiment summarizer for local service reviews*. Proceedings of the WWW2008 Workshop: NLP in the Information Explosion Era (NLPiX). Beijing, China: Association for Computational Machinery. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.182.4520>

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4–5), 993–1022. Retrieved from <http://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230. Retrieved from <http://doi.org/10.1111/j.1083-6101.2007.00393.x>
- Brody, S. (2010, June). An unsupervised aspect-sentiment model for online reviews. *Computational Linguistics*, 804–812. Retrieved from www.aclweb.org/anthology/N10-1122
- Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301–1309). Edinburg: Association for Computational Linguistics.
- Cavnar, W. B., Trenkle, J. M., & Mi, A. A. (1994). *N-gram-based text categorization*. The 3rd Annual Symposium on Document Analysis and Information Retrieval (pp. 161–175). Las Vegas: Information Science Research Institute. Retrieved from <http://doi.org/10.1.1.53.9367>
- Cha, M., Gwon, Y., & Kung, H. T. (2015). *Twitter geolocation and regional classification via sparse coding*. The 9th International Conference on Web and Social Media (ICWSM) (pp. 1–4). Oxford: Association for the Advancement of Artificial.
- Chan, M. S., Lohmann, S., Morale, A., Zhai, C., Ungar, L. H., Holtgrave, D. R., & Albaracín, D. (2018). An Online Risk Index for the cross-sectional prediction of new HIV, chlamydia, and gonorrhea diagnoses across U.S. counties and across years. *AIDS and Behavior*. <http://doi.org/https://doi.org/10.1007/s10461-018-2046-0>
- Chang, J. C., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296. Retrieved from <http://doi.org/10.1.1.100.1089>
- Chang, J. C., & Lin, C. C. (2014). Recurrent-neural-network for language detection on Twitter code-switching corpus. *CoRR*, 1412.4314.
- Chang, K., Samdani, R., & Dan Roth. (2013). *A constrained latent variable model for coreference resolution*. The 2013 Conference on Empirical Methods on Natural Language Processing (EMNLP). Seattle: Association for Computational Linguistics.
- Chang, S., Chen, Y., Yip, P., Lee, W., Hagihara, A., & Gunnell, D. (2014). Regional changes in charcoal-burning suicide rates in East/Southeast Asia from 1995 to 2011: A time trend analysis. *PLoS Medicine*, 11(4), e1001622. Retrieved from <http://dx.doi.org/10.1371/journal.pmed.1001622>
- Cheng, Z., Caverlee, J., & Lee, K. (2010). *You are where you tweet : A content-based approach to geo-locating Twitter users*. The 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada: Association for Computing Machinery, pp. 759–768. Retrieved from <http://doi.org/10.1145/1871437.1871535>
- Cheung, Y. T. D., Chan, C. H. H., Lai, C-K. J., Chan, W. F. V., Wang, M. P., Li, H. C. W., . . . Lam, T-H. (2015). Using WhatsApp and Facebook online social groups for smoking relapse prevention for recent quitters: A pilot pragmatic cluster randomized controlled trial. *Journal of Medical Internet Research*, 17(10). Retrieved from <http://doi.org/10.2196/jmir.4829>
- Chu, K-H., Unger, J. B., Allem, J-P., Pattarroyo, M., Soto, D., Cruz, T. B., . . . Yang, C. C. (2015). Diffusion of messages from an electronic cigarette brand to potential users through Twitter. *PLoS ONE*, 10(12), e0145387. Retrieved from <http://doi.org/10.1371/journal.pone.0145387>
- Cody, E. M., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2016). Public opinion polling with Twitter. *Physics and Society*. Retrieved from <https://arxiv.org/abs/1608.02024>

- Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, 26(1), 246–256. Retrieved from <http://doi.org/10.1080/09546553.2014.849948>
- Conger, K. (2016). *LinkedIn sues anonymous data scrapers*. 2016 m. gruodžio 20 d. Retrieved from <https://techcrunch.com/2016/08/15/linkedin-sues-scrapers/>
- Consonni, M., & Anselmi, L. (2015). ECJ rules on screen-scraping of Ryanair's database. *E-Commerce Law and Policy*, 17(2). Retrieved from www.orsingher.com/pdf/ECLP-15-02.pdf
- Curini, L., Iacus, S., & Canova, L. (2015). Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. *Social Indicators Research*, 121(2), 525–542. Retrieved from <http://dx.doi.org/10.1007/s11205-014-0646-2>
- De Souza, I. M., & Ferris, S. P. (2015). Social media marketing in luxury retail. *International Journal of Online Marketing*, 5(2), 18–36. Retrieved from <http://doi.org/10.4018/IJOM.2015040102>
- Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). *Social media update 2014*. 2015 m. kovo 31 d. Retrieved from www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2), 159–169. Retrieved from <http://doi.org/10.1177/0956797614557867>
- Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). *A latent variable model for geographic lexical variation*. The 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, pp. 1277–1287.
- Esuli, A., & Sebastiani, F. (2006). *SENTIWORDNET: A publicly available lexical resource for opinion mining*. The 5th Conference on Language Resources and Evaluation. Genoa, Italy: European Language Resources Association, pp. 417–422. Retrieved from <http://doi.org/10.1.1.61.7217>
- Farhadloo, M., Winneg, K., Chan, M. S., Jamieson, K. H., & Albarracín, D. (2018). Associations of topics of discussion on Twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: Probabilistic study in the United States. *JMIR Public Health and Surveillance*, 4(1), e16. Retrieved from <http://doi.org/10.2196/publichealth.8186>
- Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, 90, 1–11. Retrieved from <http://doi.org/10.1016/j.dss.2016.06.010>
- Farhadloo, M., & Rolland, E. (2013). *Multi-class sentiment analysis with clustering and score representation*. 2013 IEEE 13th International Conference on Data Mining Workshops. Dallas: IEEE, pp. 904–912. Retrieved from <http://doi.org/10.1109/ICDMW.2013.63>
- Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest*, 13(1), 3–66. Retrieved from <http://doi.org/10.1177/1529100612436522>
- Gamon, M., Gamon, M., Aue, A., Corston-Oliver, S., Corston-Oliver, S., . . . Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science*, 3646, 121–132. Retrieved from http://doi.org/10.1007/11552253_12
- Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61(1), 115–125. Retrieved from <http://doi.org/10.1016/j.dss.2014.02.003>

- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.7485&rep=rep1&type=pdf>
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. *Technometrics*. Retrieved from <http://doi.org/10.2307/1271145>
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). *Part-of-speech tagging for Twitter: Annotation, features, and experiments*. The 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers. Portland: Association for Computational Linguistics, pp. 42–47. Retrieved from <http://doi.org/10.1.1.206.3224>
- Gouws, S., Metzler, D., Cai, C., Hovy, E., & Rey, M. (2011). *Contextual bearing on linguistic variation in social media*. The Workshop of Language in Social Media (LSM 2011). Oregon: Association for Computational Linguistics, pp. 20–29.
- Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social media update 2016*. 2016 m. lapkrićio 12 d. Retrieved from www.pewinternet.org/2016/11/11/social-media-update-2016/#fn-17239-1
- Gui, L., Zhou, Y., Xu, R., He, Y., & Lu, Q. (2017). Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*. Retrieved from <http://dx.doi.org/10.1016/j.knosys.2017.02.030>
- Han, B., Cook, P., & Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49, 451–500. Retrieved from <http://doi.org/10.1613/jair.4200>
- Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., . . . Balter, S. (2014). *Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013*. Retrieved from <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6320a1.htm>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. Retrieved from <http://doi.org/10.1017/S0140525X0999152X>
- Hofmann, T. (1999). *Probabilistic latent semantic indexing*. The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999). Berkeley: Association for Computational Linguistics, pp. 50–57. Retrieved from <http://doi.org/10.1145/312624.312649>
- Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. The 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 04, T. 4. Seattle: Association for Computing Machinery, p. 168. Retrieved from <http://doi.org/10.1145/1014052.1014073>
- Huang, G. C., Soto, D., Fujimoto, K., & Valente, T. W. (2014). The interplay of friendship networks and social networking sites: Longitudinal analysis of selection and influence effects on adolescent smoking and alcohol use. *American Journal of Public Health*, 104(8), e51–e59. Retrieved from <http://search.proquest.com/docview/1549549180?accountid=9851>
- Huang, J., Kornfield, R., & Emery, S. L. (2016). 100 million views of electronic cigarette YouTube videos and counting: Quantification, content evaluation, and engagement levels of videos. *Journal of Medical Internet Research*, 18(3). Retrieved from <http://dx.doi.org/10.2196/jmir.4265>
- Ireland, M. E., & Iserman, M. (2018). *Lusi lab development dictionaries*. Retrieved June 11, 2018, from <https://www.depts.ttu.edu/psy/lusi/resources.php>

- Ireland, M. E., Chen, Q., Schwartz, H. A., Ungar, L. H., & Albarracín, D. (2015). Action tweets linked to reduced county-level HIV prevalence in the United States: Online messages and structural determinants. *AIDS and Behavior*. Retrieved from <http://doi.org/10.1007/s10461-015-1252-2>
- Ireland, M. E., Chen, Q., Schwartz, H. A., Ungar, L. H., & Albarracín, D. (2016). Action tweets linked to reduced county-level HIV prevalence in the United States: Online messages and structural determinants. *AIDS and Behavior*, 20(6), 1256–1264. Retrieved from <http://doi.org/10.1007/s10461-015-1252-2>
- Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L. H., & Albarracín, D. (2015). Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, 34(Supplement), 1252–1260. Retrieved from <http://doi.org/10.1037/hea0000279>
- Jelenchick, L. A., Eickhoff, J. C., & Moreno, M. A. (2013). “Facebook depression?” Social networking site use and depression in older adolescents. *Journal of Adolescent Health*, 52. Retrieved from www.sciencedirect.com/science/article/pii/S1054139X12002091
- Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *The International Conference on Web Search and Web Data Mining 2008*, pp. 219–230. Retrieved from <http://doi.org/10.1145/1341531.1341560>
- Johnston, K., Tanner, M., Lalla, N., & Kawalski, D. (2013). Social capital: The benefit of Facebook ‘friends’. *Behaviour & Information Technology*, 32(1), 24–36. Retrieved from <http://dx.doi.org/10.1080/0144929X.2010.550063>
- Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, 23(39), 544–559. Retrieved from <http://doi.org/10.1108/IntR-06-2012-0114>
- Kaufmann, M. (2010). *Syntactic normalization of Twitter messages*. International Conference on Natural Language Processing, T. 2, pp. 1–7. Kharagpur, India: Macmillan Publishers. Retrieved from www.cs.uccs.edu/%7B~%7Dkalita/work/reu/REUFinalPapers2010/Kaufmann.pdf
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). *A dependency parser for Tweets*. The Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, pp. 1001–1012. Retrieved from <http://doi.org/10.3115/v1/D14-1108>
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805. Retrieved from <http://doi.org/10.1073/pnas.1218772110>
- Kostygina, G., Tran, H., & Emery, S. (2016). *Follow even if you don't smoke: The amount and themes of cigarillo and marijuana co-use content on Instagram*. The APHA 2016 Annual Meeting & Expo. Denver: American Public Health Association
- Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). *Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments*. The 2011 SIAM International Conference on Data Mining. Arizona: American Statistical Association, pp. 498–509.
- Lakkaraju, H., Socher, R., & Manning, C. D. (2014). *Aspect specific sentiment analysis using hierarchical deep learning*. NIPS 2014 Workshop on Deep Neural Networks and Representation Learning. Montreal, Canada: Neural Information Processing System Foundation, pp. 1–9.
- Lawrence, B., & Perrigot, R. (2015). Influence of organizational form and customer type on online customer satisfaction ratings. *Journal of Small Business Management*, 53(Supplement 1), 58–74. Retrieved from <http://dx.doi.org/10.1111/jsbm.12184>

- Li, J., Cardie, C., & Li, S. (2013). *TopicSpam: A topic-model-based approach for spam detection*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, pp. 217–221.
- Lim, E-P., Nguyen, V-A., Jindal, N., Liu, B., & Lauw, H. W. (2010, April 2016). *Detecting product review spammers using rating behaviors*. The 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada: Association for Computing Machinery, 939–948. Retrieved from <http://doi.org/10.1145/1871437.1871557>
- Lin, L. yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., . . . Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety*, 33(4), 323–331. Retrieved from <http://doi.org/10.1002/da.22466>
- LinkedIn. (2016). *Prohibition of scraping software*. 2016 m. gruodžio 15 d. Retrieved from www.linkedin.com/help/linkedin/answer/56347/prohibition-of-scraping-software?lang=en
- Liu, J., & Inkpen, D. (2015). *Estimating user location in social media with stacked denoising auto-encoders*. Proceedings of NAACL-HLT 2015. Denver: Association for Computational Linguistics, pp. 201–210.
- Lönnqvist, J-E., & Ikonen, J. V. A. (2014). It's all about Extraversion: Why Facebook friend count doesn't count towards well-being. *Journal of Research in Personality*, 53, 64–67. Retrieved from <http://dx.doi.org/10.1016/j.jrp.2014.08.009>
- Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). *Teens, social media, and privacy*. Retrieved from www.lateledipenelope.it/public/52dff2e35b812.pdf
- Mangukiyi, P. (2016, gegužės 26). Social media by the numbers. *The Huffington Post*. Retrieved from www.huffingtonpost.com/piyush-mangukiyi/social-media-by-the-numbe_b_9757926.html
- Moghaddam, S., & Ester, M. (2012). *On the design of IDA models for aspect-based opinion mining*. The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012). Association for Computing Machinery, p. 803. Maui, Hawaii. Retrieved from <http://doi.org/10.1145/2396761.2396863>
- Mohammady, E., & Culotta, A. (2014). *Using county demographics to infer attributes of Twitter users*. The Joint Workshop on Social Dynamics and Personal Attributes in Social Media. Baltimore: Association for Computational Linguistics, pp. 7–16. Retrieved from <http://acl2014.org/acl2014/W14-27/W14-27-2014.pdf%7B#%7Dpage=19%7B%25%7D5Cn>; www.aclweb.org/anthology/W/W14/W14-2702.pdf
- Moreno, M. A., Ton, A., Selkie, E., & Evans, Y. (2016). Secret society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health*, 58(1), 78–84. Retrieved from <http://dx.doi.org/10.1016/j.jadohealth.2015.09.015>
- Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding deceptive opinion spam by any stretch of the imagination*. The 49th Annual Meeting of the Association for Computational Linguistics. Portland: Association for Computational Linguistics, p. 11. Retrieved from <http://arxiv.org/abs/1107.4557>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. The 2002 Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), T. 10. Philadelphia: Association for Computational Linguistics, pp. 79–86. Retrieved from <http://doi.org/10.3115/1118693.1118704>
- Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC*. Austin, TX. Retrieved from liwc.net
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547–577. Retrieved from <http://doi.org/10.1146/annurev.psych.54.101601.145041>

- Popescu, A-M., & Etzioni, O. (2005). *Extracting product features and opinions from reviews*. The conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005), pp. 339–346. Vancouver, British Columbia, Canada: Association for Computational Linguistics. Retrieved from http://doi.org/10.1007/978-1-84628-754-1_2
- Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). *Hierarchical Bayesian models for latent attribute detection in social media*. The Fifth International AAAI Conference on Weblogs and Social Media, T. 11. Barcelona, Spain: Association for the Advancement of Artificial Intelligence, pp. 598–601. Retrieved from www.cs.jhu.edu/%7B~%7Dmpaul/files/2011.icwsm.nigeria.pdf%7B~%7D25%7D5Cn; www.cs.jhu.edu/%7B~%7Ddelip/icwsm.pdf
- Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). *Named entity recognition in Tweets: An experimental study*. The 2011 Conference on Empirical Methods in Natural Language Processing. Edinburg, United Kingdom: Association for Computational Linguistics, pp. 1524–1534. Retrieved from <http://doi.org/10.1075/li.30.1.03nad>
- Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). *Supervised text-based geolocation using language models on an adaptive grid*. The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1500–1510. Jeju Island, Korea: Association for Computational Linguistics.
- Santos, J. C., & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology & Medical Modelling*, 11(Supplement 1), S6. Retrieved from <http://doi.org/10.1186/1742-4682-11-S1-S6>
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., . . . Ungar, L. (2013). *Characterizing geographic variation in well-being using tweets*. The Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13). Cambridge: Association for the Advancement of Artificial Intelligence, pp. 583–591. Retrieved from <http://doi.org/papers3://publication/uuid/43E3E88F-EFDC-4F9C-85AC-C60B4B8C8BCA>
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9), e73791. Retrieved from <http://doi.org/10.1371/journal.pone.0073791>
- Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 78–94. Retrieved from <http://doi.org/10.1177/0002716215569197>
- Seidenberg, A., Jo, C., Ribisl, K., Lee, J., Butchting, F., Kim, Y., & Emery, S. (2017). A national study of social media, television, radio, and internet usage of adults by sexual orientation and smoking status: Implications for campaign design. *International Journal of Environmental Research and Public Health*, 14(4), 450. Retrieved from <http://doi.org/10.3390/ijerph14040450>
- Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, 6(5), e19467. Retrieved from <http://doi.org/10.1371/journal.pone.0019467>
- Smyser, J. D. (2013). *Health communication and social media: A case study of the California Tobacco Control Program's "Toxic Butts" campaign*. ProQuest Dissertations and Theses. Retrieved from <http://sfx.scholarsportal.info/guelph/docview/1494825145?accountid=>

- 11233%7B%25%7D5Cn; http://sfx.scholarsportal.info/guelph?url%7B_%7Dver=Z39.88-2004%7B%7Ddrft%7B_%7Dval%7B_%7Dfmt=info:ofi/fmt:kev:mtx:dissertation%7B%7Dgenre=dissertations+%7B%25%7D26+these
- Statista. (2010). Number of social media users worldwide from 2010 to 2020 (in billions). 2017 m. gegužės 21 d. Retrieved from www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
- Stevens, P., Carlson, L. M., & Hinman, J. M. (2004). An analysis of tobacco industry marketing to lesbian, gay, bisexual, and transgender (LGBT) populations: Strategies for mainstream tobacco control and prevention. *Health Promotion Practice*, 5(3), 129–134. Retrieved from <http://doi.org/10.1177/1524839904264617>
- Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., . . . Su, Z. (2008). *Hidden sentiment association in Chinese web opinion mining*. The 17th International Conference on World Wide Web. Beijing, China: Association for Computing Machinery, pp. 959–968. Retrieved from <http://doi.org/10.1145/1367497.1367627>
- Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4), 483–496. Retrieved from <http://doi.org/10.1109/91.940962>
- Sun, H., Morales, A., & Yan, X. (2013). Synthetic review spamming and defense. *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, p. 1088. New York, NY: Association for Computing Machinery. Retrieved from <http://doi.org/10.1145/2487575.2487688>
- Teitler, J. O., Reichman, N. E., & Sprachman, S. (2003). Costs and benefits of improving response rates for a hard-to-reach population. *Public Opinion Quarterly*, 67(1), 126–138. Retrieved from <http://doi.org/10.1086/346011>
- Thangarajan, N., Green, N., Gupta, A., Little, S., & Weibel, N. (2015). *Analyzing social media to characterize local HIV at-risk populations*. The Conference on Wireless Health (WH 2015), pp. 1–. New York, NY: Association for Computing Machinery. Retrieved from <http://doi.org/10.1145/2811780.2811923>
- Tran, H., Hornbeck, T., Ha-Thuc, V., Cremer, J., & Srinivasan, P. (2011). *Spam detection in online classified advertisements*. The 2011 Joint International Workshop on Information Credibility on the Web and Adversarial Information Retrieval on the Web Workshop on Web Quality, pp. 35–41. Hyderabad, India: Association for Computing Machinery. Retrieved from <http://doi.org/10.1145/1964114.1964122>
- Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. The 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 417–424). New York: Association for Computational Linguistics. Retrieved from <http://doi.org/10.3115/1073083.1073153>
- Upadhyay, S., Gupta, N., Christodoulopoulos, C., & Roth, D. (2016). *Revisiting the evaluation for cross document event coreference*. The 26th International Conference on Computational Linguistics. Osaka, Japan: International Committee on Computational Linguistics.
- Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2016). *En-es-es: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis*. The Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association, pp. 4149–4153.
- Wang, H., Wang, C., Zhai, C., & Han, J. (2011). *Learning online discussion structures by conditional random fields*. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China: Association for Computing Machinery, pp. 435–444. Retrieved from <http://doi.org/10.1145/2009916.2009976>

- Ward, J. (2016). What are you doing on Tinder? Impression management on a matchmaking mobile app. *Information, Communication & Society*, 1–16. Retrieved from <http://doi.org/10.1080/1369118X.2016.1252412>
- Wiebe, J. M. (2000). *Learning subjective adjectives from corpora*. The National Conference on Artificial Intelligence (pp. 735–741). Austin: Association for the Advancement of Artificial Intelligence. Retrieved from <http://doi.org/http://portal.acm.org/citation.cfm?id=721121&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618>
- Wing, B., & Baldridge, J. (2014). *Hierarchical discriminative classification for text-based geolocation*. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 336–348). Doha, Qatar: Association for Computational Linguistics. Retrieved from <http://anthology.aclweb.org/D/D14/D14-1039.pdf>
- Wu, Y., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040. Retrieved from <http://doi.org/10.1073/pnas.1418680112>
- Young, A. L., & Quan-Haase, A. (2009). *Information revelation and internet privacy concerns on social network sites*. The Fourth International Conference on Communities and Technologies (C&T 2009) (p. 265). New York: Association for Computing Machinery. Retrieved from <http://doi.org/10.1145/1556460.1556499>
- Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, 63, 112–115. Retrieved from <http://doi.org/10.1016/j.ypmed.2014.01.024>
- Yuan, E. J., Feng, M., & Danowski, J. A. (2013). “Privacy” in semantic networks on Chinese social media: The case of Sina Weibo. *Journal of Communication*, 63(6), 1011–1031. Retrieved from <http://dx.doi.org/10.1111/jcom.12058>
- Zamal, F. Al, Liu, W., & Ruths, D. (2011). *Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors*. The Sixth International AAAI Conference on Weblogs and Social Media (pp. 387–390). Dublin, Ireland: Association for the Advancement of Artificial Intelligence.
- Zhan, Y., Liu, R., Li, Q., Leischow, S. J., & Zeng, D. D. (2017). Identifying topics for e-cigarette user-generated contents: A case study from multiple social media platforms. *Journal of Medical Internet Research*, 19(1). Retrieved from <http://doi.org/10.2196/jmir.5780>
- Zide, J., Elman, B., & Shahani-Denning, C. (2014). LinkedIn and recruitment: How profiles differ across occupations. *Employee Relations*, 36(5), 583–604. Retrieved from <http://doi.org/10.1108/ER-07-2013-0086>

INDEX

Note: Page numbers in *italics* indicate figures and numbers in **bold** indicate tables.

- Action Interference Paradigm 37
- activation of mental contents 31, 38, 45, 48
- adaptive immunity 90
- adverb qualifiers/anchors for rating scales 21–24, 25
- Affect Misattribution Procedure (AMP) 35–36, 47
- age, as confound 181
- aggregation approach to measurement error 4–5
- altruism 178
- ambiguity, avoiding 14, 15
- analysis of covariance (ANCOVA) 141
- ANS *see* autonomic nervous system
- anticipatory processes and ERPs 107–109
- APE (associative-propositional evaluation) model 38–39, 41–42
- application of activated contents 38, 45, 48
- application program interfaces (APIs), setting up 233, 237
- approach-avoidance tasks 36–37
- archival research: confounds and 181–183; on counterfactual thinking and emotions 186–187; on ethnic stereotyping 184, 185, 186; external validity and 177–178, 179, 180–181; on false consensus 183–184; internal validity and 175–177; on longevity and death 192–196, 195; overview 174, 196–197; on self-concept and identity 188–192, 190, 191; strengths and weaknesses of 174–175; theory and 183
- artificial intelligence 167
- aspect-level analysis 246, 248–249
- associative-propositional evaluation (APE) model 38–39, 41–42
- attitudes: defined 31; in social media data 246, 248–249
- automaticity of experimental effects and implicit measures 45–46
- autonomic nervous system (ANS): anatomy and physiology of 76–77; cardiovascular indices 80–81; electrodermal indices 81–82; indices in social psychological research **78–79**; quantification of activity of 77; research applications 82, **83**
- behavior: archival data as source of information on 196–197; elicitation research in interventions to change 59–60, 68–71; prediction of, with implicit and explicit measures 39–41
- Berra, Yogi 174
- between-person effects 129
- bias: experimenter 175; implicit racial 114–116
- blood samples 91
- broad constructs 5–6, 11, 13
- buying social media data 237–238

- cardiovascular indices 80–81
- career choice and implicit egotism 188, 189–190, 190
- causality: confounds and 181–183; plausible 214; requirements to establish 175–177
- closed-ended quantitative assessments 58, 70
- closed-vocabulary approach and social media data 248, 250
- code-switching 254
- coefficient alpha 6, 8
- coefficients of multilevel models **212**
- cognitive development, theory of 204–205
- cognitive interviewing 137
- cognitive science 167
- communication, analysis of 153–154; *see also* textual analysis
- components of ERPs 105
- composite reliability 8
- computational social science 167
- computer-mediated communication 153
- confirmatory factor analysis 7
- confounds and causality 176–177, 181–183
- construct of interest: defining 5–8; measurement of 4; scope of 11, 13
- content words 154–155
- context effects and implicit measures 45
- contextual information for items 15
- contingent negative variation 107–108
- convenience samples, non-probability based 135
- coordinate reference systems 203
- coreference resolution 244
- correlation: between explicit and implicit measures 37–39; intraclass 135; spatial autocorrelation 217–218, 223, 225
- cortisol 86–87
- counterfactual thinking and emotions 186–187
- covariation and causality 175–176
- cross-temporal meta-analysis 218
- cultural diversity in elicitation research 65–67
- culture, textual analysis research on 165–166
- daily diary research: advantages of 132–133; burden of 134; data analysis in 141–144; ego depletion theory and 129–131; EMA compared to 144; interactive voice response systems and 131–132; as intervention tool 145; limitations of 133–134; methodological considerations of 134–140; overview 128, 131, 146–147; wearables and 145–146
- data: buying from resellers 237–238; from daily diary research 133, 141–144; geospatial 207–214, **210**, **212**; management of 140; from psychobiological assessment 95–96; spatiotemporal 201, 216–222, 220, 253–254; use of, and opt-outs/opt-ins 168; *see also* archival research; Network Common Data Format; social media data
- datasets, combining 203
- deaths: archival studies of 192–196, 195; by season 179
- deception, language markers for 162–163
- definition of broad constructs 5–6
- demographics: of respondents, and systematic error 16; from social media data 138–140; WEIRD (Western, educated, industrialized, rich, and democratic) 182–183, 238
- dependent variables, implicit and explicit measures as 41–42
- descriptive research, elicitation research in 57, 60–65
- detection of spam within social media 249–250
- dimensionality of constructs 5–6
- direct measurement instruments 31
- distortions, motivated 44
- “don’t know” responses 20
- double negations, avoiding 14–15
- downloading free social media datasets 232–233, **234–236**
- “dual processes” for thought and action 102
- duration of daily diary studies 136
- early stage research, elicitation research in 58, 65–67
- EAST (Extrinsic Affective Simon Task) 34–35
- ecological momentary assessment (EMA) 130, 138, 144–145
- effect sizes, as varying over space and time 205–206
- ego depletion theory 129–131
- electrocardiography 80–81
- electrodermal indices 81–82
- electroencephalogram (EEG) 103, 107
- Electronically Activated Recorders 146

- elicitation research: in descriptive, qualitative research 57, 60–65; in early stage research 58; in intervention research 59–60; overview 56–57; rationale for 71–72; in theory development 58–59, 67–68; uses of 71
- EMA (ecological momentary assessment) 130, 138, 144–145
- endocrine system: anatomy and physiology of 83–86; cortisol 86–87; hormones and effects **84–85**; other hormones 87; research applications 88, **89**
- enumerative assays 91, **92**
- ERPs *see* event-related potentials
- error *see* measurement error
- error-related negativity (ERN) 113–116, 115
- ethnicity: as confound 182; stereotyping 184, 185, 186
- event-contingent measures 144
- event-related potentials (ERPs): applying to information processing 107, 108; measurement of 105–106; overview 103, 104–105, 118; reliability of 106–107; schematic representation of 104; social-personality process and 103–104; of textual analysis 168; validity of 106
- experience sampling 144
- experimenter bias 175
- explanatory models and social media data 250–251
- explicit measures: correlations between implicit measures and 37–39; as dependent variables 41–42; overview 29–31; prediction of behavior with 39–41; *see also* self-reports
- external validity 177–178, 179, 180–181
- extreme anchors for rating scales 23–24
- Extrinsic Affective Simon Task (EAST) 34–35
- face perception 109–110
- factor loading 7
- false consensus effect 183–184
- feedback processing and ERPs 116–118
- figural aspects to research 205
- Fitbits 145–146
- focus group elicitation research 61–62, 65–66
- forensic analyses, language markers for 162–164
- functional assays 91, **92**
- function words 154–155, 158
- GAGES confounds 181–183
- Galton, Francis 159
- gender, as confound 181–182
- generalized estimating equation (GEE) models 142
- generational diversity in elicitation research 65–67
- geography, as confound 181
- geo-mapping of social media data 253–254
- geospatial data, applications of 207–214, **210, 212**
- Go/No-Go Association Task 34
- grids of geospatial units 203–204, 204
- Guttman scaling 9, 11, 12
- health: social media data on 250–252; textual analysis research on 164–165; *see also* deaths; immune system
- heart rate 80
- holidays and mortality 195, 195–196
- homogeneity of multi-item scales 6–7
- hormones: classes of 84–85; primary effects of **84–85**; release of 85–86; research applications for **89**
- hypothalamic-pituitary-adrenal (HPA) axis 86
- hypothalamus 85
- IMB (Information-Motivation-Behavioral Skills model) 58–59, 69, 69–70
- immune system: anatomy and physiology of 88–90; assessment of cells and processes of 91–93; functional and enumerative PNI measures **92**; research applications 93, **94**
- Implicit Association Test (IAT) 33–34, 47
- implicit egotism 188–192, 190, 191
- implicit measures: assumptions in research using 42–47; automaticity of experimental effects and 45–46; context effects and 45; correlations between explicit measures and 37–39; defined 30; as dependent variables 41–42; explicit measures compared to 29–31; mental association and 31–32; prediction of behavior with 39–41; as process-pure reflections 46–47; as providing access to old representations 44–45; as providing window to unconscious 43–44; reliability of 47; scoring procedures and numerical values for 42–43; self-reports

- compared to 24; social desirability problems and 44; use of 47–48; *see also* performance-based instruments
- implicit racial bias 114–116
- Implicit Relational Assessment Procedure 37
- incentives for daily diary studies 139
- indirect measurement instruments 31
- individual-level indices and spatial variables 210–211
- Information-Motivation-Behavioral Skills model (IMB) 58–59, 69, 69–70
- information-processing system *see* event-related potentials
- innate immunity 89–90
- intensive repeated measures (IRM) designs 128, 146–147; *see also* daily diary research
- interactive voice response systems 131–132
- internal consistency of multi-item scales 6–7
- internal reliability 106–107
- internal validity: GAGES confounds and 181–183; overview 175–177
- interpretation of results of
 - psychobiological assessment 97
- intervention, daily diary methods as 145
- intervention research, elicitation research in 59–60, 68–71
- interviews: cognitive 137; semi-structured 62–63; structured 64
- intraclass correlation 135
- IRM (intensive repeated measures) designs 128, 146–147; *see also* daily diary research
- item-operating characteristics (IOCs): articulating 13–14; overview 9, 10, 12
- iterative process of scale construction 6

- James, William 75
- jargon, avoiding 15

- known-groups validity approach 106

- lab experiments 127–128
- language: natural language processing 166–167, 243–245; style matching 161; *see also* Linguistic Inquiry and Word Count; textual analysis
- Latent Dirichlet Allocation 245, 248–249, 252
- late positive potential 119n4

- lateralized readiness potential (LRP) 112–113, 114
- latitude 202
- Lexical Hypothesis of Personality 159–160
- Linguistic Inquiry and Word Count (LIWC): opinion spam detection and 249; overview 155–158; research applications of 158–166
- loading 2
- location information on social media 253
- longitude 202
- LRP (lateralized readiness potential) 112–113, 114

- MANOVA/MANCOVA (multivariate analysis of variance/covariance) 141–142
- marriage and implicit egotism 190–192, 191
- measurement-burst design 136
- measurement error: approaches to 4–5; prediction error 116; random error 2–3, 14–16; systematic error 3, 3–4, 16–19
- measurement model 1–2, 2
- measurement outcomes, explicit and implicit 30–31
- measures development for daily diary studies 137
- mental association and implicit measures 31–32
- mental events: during affective priming 108; timing of 102–103
- meta-analysis, spatiotemporal 218–219
- midpoints in rating scales 22
- Mill, John Stuart 175–176
- mind-body phenomenon 75
- Mischel, Walter 129
- mixed-effects models with geocoded fixed-effects variables 211–214, **212**
- mixed modeling 142–144
- MODE (Motivation and Opportunity as DEterminants) model 38, 39
- Modern Racism Scale 48n2
- monitoring vendors, subscribing services from 237
- multilevel modeling (MLM) 142–143
- multilevel models: applications of
 - geospatial data in 207–214, **210**;
 - distribution of residuals across space 214, 215, 216; mixed-effects models with geocoded fixed-effects variables 211–214; overview 205–206; space as

- random-effects variable in 208–209, **210**; spatial variables and individual-level indices in 210–211; standardized coefficients of **212**; traditional approach to 206–207, **207**
- multivariate analysis of variance/covariance (MANOVA/MANCOVA) 141–142
- N170 component of ERPs 109–110
- N179 *110*
- named entity recognition 244
- narrow constructs 5
- naturalism of daily diary methods 132
- natural language processing (NLP) 166–167, 243–245
- negations, avoiding 14–15
- negatively keyed items 18–19
- Network Common Data Format (NetCDF) 219–220, 220, 224
- neural transmission speed 102–103
- non-response in daily diary studies 138–139
- non-semantic features of social media data 238–240, 239
- “no opinion” responses 20
- numeric ratings, combining with adverb anchors 21–22, 23
- observation, sensitivity of humans to 174–175
- Observatory on Social Media 233
- occasions and external validity 178, 179
- OOPS heuristic 177–178
- open-ended qualitative approaches 58
- open-ended questions 60–61
- open-vocabulary approach 251–252, **252**
- operationalizations and external validity 177–178
- opinion spam, detection of 249–250
- P3 component of ERPs 111–112
- parasympathetic nervous system 77
- parsers 244
- participant-informed methods 56–57; *see also* elicitation research
- participant recruitment for daily diary studies 135–136
- partner effects 133
- pattern identification methods 217
- performance-based instruments: Affect Misattribution Procedure 35–36; approach-avoidance tasks 36–37; extrinsic affective priming tasks 34–35; Go/No-Go Association Task 34; Implicit Association Test 33–34; list of 37; overview 29; sequential priming tasks 32–33; *see also* implicit measures
- personalizing items 15
- persuasion, textual analysis research on 161–162
- Piaget, Jean 204
- pituitary gland 85
- plausible causality 214
- PNI (psychoneuroimmunology) 88–89, **92**, 92–93, **94**
- point-pattern analysis 216–217
- populations and external validity 180
- positively keyed items 18–19
- precision of metrics or scales 19–20
- prediction error 116
- prediction of behavior with implicit and explicit measures 39–41
- predictive models and social media data 250–251
- prejudice 31
- Prime Meridian 202
- principles of measurement 1–9, 11
- Probabilistic Latent Semantic Indexing 245
- process-pure reflections, implicit measures as 46–47
- programming requirements for daily diary studies 139–140
- pro-inflammatory cytokines 90
- pronouns, use of 160, 162–163
- psychobiological assessment: data storage, processing, analysis, and interpretation 95–97; future of 97–98; methodologic and analytic considerations 93; overview 75–76; timing and number of measurements 95; tonic versus phasic measures 94–95; *see also* autonomic nervous system; endocrine system; immune system
- psychoneuroimmunology (PNI) 88–89, **92**, 92–93, **94**
- quad-model 46
- quantitative inventory development 65–67
- quantitative measures and spatiotemporal data 216–222
- questionnaire development 65

- questions in self-reports: abbreviations, avoiding 15–16; acceptance, conveying 17; keeping short, simple, and understandable 14; leading 17; “linear wording” approach to asking 13–14; orienting around response categories 20; *see also* writing items for self-reports
- random-effects variable, space as 208–209, **210**
- random error 2–3, 14–16
- rating (Q) 1–2, 2
- real-time advantage of daily diary research 132–133
- regional differences, textual analysis research on 165–166
- regression coefficient 2
- Relational Responding Task 37
- relationship dynamics, textual analysis research on 160–161
- reliability: assessment of 8; defined 3; of ERPs 106–107; of implicit measures 47; of textual analysis 168
- research: assumptions in, using implicit measures 42–47; figural and spatial aspects to 205; methodologies of 127–128; *see also* archival research; daily diary research; elicitation research; textual analysis
- resellers, buying data from 237–238
- residential choice and implicit egotism 188–189
- resistance to change of implicit measures 44–45
- response categories 20
- response preparation and ERPs 112–113
- response processing and ERPs 113–116
- response styles 17
- response times (RTs) 103, 107
- reverse-oriented, counterintuitive scales 18–19
- reward positivity (RewP) 106, 116–118, **117**
- rubric for exclusion from daily diary studies for missing data 140
- salience network 113
- salivary assays 91–92
- saliva samples 86–87
- sampling: convenience samples, non-probability based 135; experience sampling 144
- scale construction, as iterative process 6
- scales, multi-item: characteristics of 6–8; factor structure of 18; item analyses on 9, 11; item-operating characteristics of 13–14; uses of 4; *see also* self-reports
- scales, rating: adverb qualifiers/anchors for rating scales 21–24, 25; numeric ratings, combining with adverb anchors 21–22, 23
- scales, response metric 18–19
- scaling function, making explicit 8–9, **10**, **11**, **12**
- secret-keeping, language markets for 163
- self-concept: defined 31; implicit egotism and 188–192, **190**, **191**
- self-control, capacity for 129–131
- self-reports: appeal of 1; context effects on 45; implicit measures compared to 24; limitations of 29; psychobiological assessment as replacement for or complement to 75; *see also* writing items for self-reports
- semantic features of social media data: less detailed contents **240**, **241**, **242**; more detailed contents **241**, **242**–**245**; overview **239**; topics as **245**–**246**, **247**, **248**–**250**
- semi-structured interviews 62–63
- sentiment analysis **246**, **248**–**249**
- sequential priming tasks 32–33
- signal-contingent measures 144
- situations and external validity 180–181
- skin conductance 81, **82**
- skin potential 81–82
- slang, avoiding 15–16
- SNS (sympathetic nervous system) **76**–**77**
- social cognition, archival studies of 183–187
- social desirability: defined **3**, **3**–**4**; implicit measures and **44**; reducing effects of **16**–**17**
- social media: location information on **253**; platforms **229**, **232**; users of sites **228**
- social media data: analysis of, as research tool **228**–**229**; challenges to use of **253**–**254**; harnessing **238**–**240**, **242**–**246**, **248**–**250**; harvesting **232**–**233**, **233**, **234**–**236**, **237**–**238**; non-semantic features **238**–**240**, **239**; to obtain inferences **250**–**252**; : overview **238**, **239**; sample studies **230**–**231**; semantic

- features 239, 240, **241**, 242–246, 248–250
- social networks, textual analysis research on 162, 165
- social psychology: foundation of discipline of 127; research methodologies of 127–128
- socioeconomic standing, as confound 182
- space: distribution of residuals across 214, 215, 216; as random-effects variable 208–209, **210**
- spam detection 249–250
- spatial aspects to research 205
- spatial autocorrelation 217–218, 223, 225
- spatial clustering measures 216–218
- spatial polygons 221–222
- spatial techniques: benefits of 222; code for 222–225; relevance of 206–207, **207**
- spatiotemporal data: mapping of social media data 253–254; overview 201; quantitative measures and 216–222, 220
- spatiotemporal meta-analysis 218–219
- spatiotemporal variables: individual-level indices and 210–211; overview 202–203, 203, 204; space as random-effects variable 208–209, **210**
- SPN (stimulus-preceding negativity) 108–109
- stability of implicit measures 44–45
- Stapel, Diederik 163–164
- statistical analysis of psychobiological assessment 96
- status, textual analysis research on 159
- stereotypes 31, 184, 185, 186
- stimulus-preceding negativity (SPN) 108–109
- stimulus processing and ERPs 109–112
- strategic control 29
- structural equation modeling 4
- structured interviews 64
- subscribing services from monitoring vendors 237
- sympathetic nervous system (SNS) 76–77
- systematic error 3, 3–4, 16–19
- temporal sequence and causality 176
- test-retest reliability 107
- text normalization 243
- text structure 244
- text tokenization 243–244
- textual analysis: content versus function words 154–155; goals across fields for 166–167; Linguistic Inquiry and Word Count (LIWC) 155–158; overview 153–154; research applications of 158–166; social psychological applications of 167–168; software for **157**; technological enablers of growth of 167
- theory: archival research and 183; of cognitive development 204–205; of ego depletion 129–131; elicitation research in development of 58–59, 67–68; importance of 204–205
- Theory of Reasoned Action 67–68
- Thurstone scaling 9, 10
- time-contingent measures 144
- timing for daily diary studies 137–138
- tonic versus phasic measures 94–95
- topic modeling 245–246, 247, 248, 250
- Ultimatum Game 116–117, 117
- unconditional means model 209, **210**, 214, 223, 224
- unconscious, implicit measures as providing window to 43–44
- unidimensionality, assessment of 7–8
- validity of measures: external validity 177–178, 179, 180–181; internal validity 175–177; systematic error as threat to 4
- variable centering 143–144
- von Helmholtz, Hermann 102–103
- wearable technology 132, 145–146
- web scraping programs 237
- WEIRD (Western, educated, industrialized, rich, and democratic) demographic 182–183, 238
- within-person effects 129
- word counting programs 155–158, **157**
- working memory capacity 40
- writing items for self-reports: articulating item-operating characteristics 13–14; item metrics, designing 19–24; random error, reducing 14–16; scope of constructs, defining 11, 13; systematic error, reducing 16–19
- zero value in rating scales 22