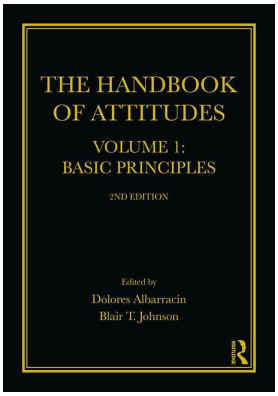

The Handbook of Attitudes
Volume 1: Basic Principles

Dolores Albarracín, Blair T. Johnson

The Measurement of Attitudes

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315178103-2
Jon A. Krosnick, Charles M. Judd, Bernd Wittenbrink
Published online on: 04 Sep 2018

2

THE MEASUREMENT OF ATTITUDES

Jon A. Krosnick, Charles M. Judd, and Bernd Wittenbrink

Attitude measurement is pervasive. Social psychologists routinely measure attitudes when studying their causes (e.g., Fishbein & Ajzen, 1975; Tesser, Whitaker, Martin, & Ward, 1998; Zajonc, 1968); how they change (e.g., Festinger, 1957; Hovland, Janis, & Kelley, 1953; Petty & Cacioppo, 1986); and their impact on cognition and behavior (e.g., Lord, Ross, & Lepper, 1979). Attitude measurement is also frequently done by political scientists, sociologists, economists, and other academics. Commercial market researchers are constantly engaged in measuring attitudes toward real and imagined consumer products and services. Beginning in the 1990s, all agencies of the U.S. federal government have initiated surveys to measure attitudes toward the services they provided. And the news media regularly conduct and report surveys assessing public attitudes toward a wide range of objects. One of the most consequential examples is the routine measurement of Americans' approval of their president.

In order to gauge people's attitudes, researchers have used a wide variety of measurement techniques. These techniques have varied across history, and they vary across professions today. This variation is due to both varying philosophies of optimal measurement and varying availability of resources that limit assessment procedures. When attitude measurement was first formalized, the pioneering scholars presumed that an attitude could only be accurately assessed using a large set of questions that were selected via an elaborate procedure (e.g., Likert, 1932; Thurstone, 1928). But today, attitudes are most often assessed using single questions with relatively simple wordings and structures, and the variability of the approaches is striking, suggesting that there is not necessarily one optimal way to achieve the goal of accurate measurement.

Recently, however, scholars have begun to recognize that the accumulating literature points to clear advantages and disadvantages of various assessment approaches, so there may in fact be ways to optimize measurement by making good choices among the available tools. Furthermore, some challenging puzzles have appeared in the literature on attitude measurement that are stimulating a re-evaluation of widely shared presumptions. This makes the present a particularly exciting time for reconsidering the full range of issues relevant to attitude measurement.

In this chapter, we offer a review of issues and literatures of use to researchers interested in assessing attitudes. We begin by considering the definition of attitudes, because no measurement procedure can be designed until the construct of interest has been specified. We review a range of different definitions that have been adopted throughout the history of social psychology but settle in on one that we believe captures the core essence of the notion of attitudes and that we use to shape our discussions throughout.

Because attitudes, like all psychological constructs, are latent, we cannot observe them directly. So all attitude measurement depends upon those attitudes being revealed in overt responses, either verbal or nonverbal. We therefore turn next to outlining the processes by which we believe attitudes are expressed, so we can harness those processes to accurately gauge the construct. And we outline the criteria for optimal measurement that we use throughout the rest of the chapter: reliability, validity, and generalizability.

Having thus set the stage, we turn to describing and evaluating various techniques for measuring attitudes, beginning with direct self-reports (which overtly ask participants to describe their attitudes). We outline many ways by which a researcher can design direct self-report measures well and less well. Next, we acknowledge the limits of such direct self-reports. A range of alternative assessment techniques, some old and others very new, have been developed to deal with this potential problem, and we review those techniques next.

Defining the Construct

Attitudes have been central to social psychology since its inception. In the first edition of the *Handbook of Social Psychology* (1935), Gordon Allport started his highly influential chapter on the topic with the following observation:

> The concept of attitude is probably the most distinctive and indispensable concept in contemporary social psychology. . . . This useful, one might almost say peaceful concept has been so widely adopted that it has virtually established itself as the keystone in the edifice of American social psychology. In fact several writers (cf. Bogardus, 1931; Thomas & Znaniecki, 1918; Folsom, 1931) *define* social psychology as the scientific study of attitudes. (p. 798; emphasis in original)

Given this centrality, one might expect to find great consistency over years and consensus across scholars in the discipline on a definition of attitudes. But such is certainly not the case. Early on, attitudes were very broadly defined. As Allport (1935) put it, "An attitude is a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related" (p. 798). Given this definition, it is hardly surprising that attitudes were seen as the central construct of social psychology, for they were whatever internal sets or predispositions motivated social behavior (see also Albarracín, Sunderrajan, Lohmann, Chang, & Jiang, this volume).

Since Allport, the definition of attitudes has evolved considerably, focusing much more on approach and avoidance behaviors and defining attitudes as the evaluative predispositions that lead to these. Thus, for instance, Eagly and Chaiken (1993) defined the construct as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (p. 1). Accordingly, an attitude is focused on a particular entity or object, rather than "all objects and situations with which it is related." Additionally, an attitude is a predisposition to like or dislike that entity, presumably with approach or avoidance consequences.

Although the evolution of the definition of attitudes in the discipline has many causes, it is interesting to note that measurement considerations were at least partly responsible. The early definitions, as sets or predispositions that motivated social behavior, were so broad that early measurement attempts were necessarily forced to simplify and place limits on the construct. Indeed Thurstone (1931), among the first to systematically address attitude measurement, noted that:

> an attitude is a complex affair which cannot be wholly described by any single numerical index. For the problem of measurement this statement is analogous to the observation

> that an ordinary table is a complex affair which cannot be wholly described by any single numerical index. So is a man [sic] such a complexity which cannot be wholly represented by a single index. Nevertheless we do not hesitate to say that we measure the table.
>
> (p. 255)

He then more narrowly defined what he proposed to measure: "Attitude is here used to describe potential action toward the object with regard only to the question whether the potential action will be favorable or unfavorable toward the object." The demands of measurement meant that the construct was limited only to evaluative predispositions and that it was narrowed to predispositions towards a single attitude object, in a very similar manner to Eagly and Chaiken's more recent definition.

The need for measurement not only mandated the narrowing of the construct; it also led to the important recognition that manifestations of attitudes, as assessed by any measurement procedure, are not the same as the attitude itself. Measurement permits one to assign values to individuals in a theoretically meaningful manner, such that differences in those values are thought to reflect differences in the underlying construct that is being measured (Dawes & Smith, 1985; Judd & McClelland, 1998). However, measurement is imperfect: The numerical values that are assigned contain both random errors and systematic errors, with the latter reflecting differences in underlying constructs other than the attitude that one intended to measure. All measurement procedures are necessarily errorful in both of these ways. Accordingly, the attitude is a *latent* evaluation of an object, manifested imperfectly both by our measurement procedures and by other observable behaviors that it in part motivates.

To say that an attitude is a latent evaluation of an object is not to say that it necessarily exists as a single entity in the mind of the attitude-holder. It may, of course. And in that case, it seems reasonable to think of an attitude as a single evaluative association with the attitude object, capable of being reported (albeit with error) in any given measurement scenario. But there are alternatives.

Perhaps a person has many stored associations with a particular attitude object, and these stored associations each have evaluative implications. But, for whatever reason, these evaluative implications have never been integrated or crystallized into a single evaluative summary stored in memory. For instance, perhaps when you think about your neighbor, you think about the fact that his yard is messy, that he accumulates rusting cars in his driveway, and that he has a couple of dogs that are nuisances. Each of these attributes that you associate with your neighbor tend to have negative evaluative overtones: You generally don't like messy yards, rusting cars, and nuisance dogs. But, somehow, you have never integrated these evaluative implications into a net evaluation of your neighbor. In this case, when there is no summary evaluation of the object (i.e., the neighbor), can we really speak of an attitude? We believe that we can, although the latent evaluation is doubly latent. Not only is it not observable by someone who wishes to measure it, but it also never exists as a discrete stored association. Rather, it becomes crystallized only under circumstances that demand a summary evaluation, such as when an overall attitude is demanded by a behavioral encounter (e.g., when you are asked "So, do you like your neighbor?").

When a single evaluative association does not exist, attitude reports may vary depending on the particular context in which those attitudes are reported, because different contexts may invoke different integration rules. For instance, if you are asked how much you like your neighbor when he has just acquired a new puppy, then the negative implications of the nuisance dogs might be perceptually overshadowed by the cuteness of the new arrival. And an integrated overall evaluation constructed at that point in time might be slightly less negative as a result. If time were to pass and the salience of the new puppy were to decrease, the overall evaluation of your neighbor might become increasingly negative again.

Because of this context-driven variability in attitude reports, some theorists have suggested that there is in fact no single attitude stored in memory for anyone (for reviews, Gawronski & Brannon,

this volume; Schwarz & Lee, this volume). Instead, these scholars argue that attitudes are constructions, fleeting by their very nature and subject to the direction in which the proverbial wind is blowing at the moment the construction is built. And the construction vanishes shortly thereafter, to be replaced by another construction, built largely independently sometime later. Indeed, some speak of individuals as having multiple attitudes toward an object instead of just one (Schwarz & Strack, 1991; Tourangeau & Rasinski, 1988; Wilson, Lindsey, & Schooler, 2000). However, we see great theoretical and practical value in resisting this extreme formulation and prefer still to hypothesize that a single attitude exists in a person's mind: the net evaluation associated with the object. The observable report of the attitude, representing the integration of evaluative implications at a given point in time, may vary as a function of the specific context in which that integration takes place, but the underlying ingredients from which that report is built (and which constitute the attitude in our formulation) are relatively stable over time.

Because an attitude is a latent construct, either existing in a relatively crystallized form or yet to be integrated into a summary representation, it is important to recognize that the attitude is *not* the numerical summary or the behavioral response that our measurement procedure produces as a product. Nevertheless, the process of attitude measurement is one of attempting to work backwards, going from the response back to the latent construct that is the attitude. To understand this process, it behooves us to better understand the cognitive processes that intervene between the latent attitude and particular responses that are manifested when attitude measurement is attempted. As we will see, understanding these processes, from the latent evaluation to manifest responses, will help us define some of the differences between what we will call "direct" measurement procedures (where we take literally the verbal self-reports of attitudes as indicative of latent attitudes) and "indirect" procedures (where we infer attitudes without asking people directly to report them).

A Processing Framework for Attitude Reports

In this section, we outline a framework for the cognitive processes by which an attitudinal evaluation is generated and by which this evaluation then subsequently shapes response tendencies. The past 20 or so years of attitude research have seen a variety of such processing accounts (e.g., Bassili & Brown, 2005, Chaiken, 1987; Fazio, 1990; Petty & Cacioppo, 1986; Strack & Martin, 1987; Wegener & Petty, 1997; Wilson, Lindsey, & Schooler, 2000). The specific framework that we present here is largely based on these accounts and distinguishes between three stages of the evaluation process: (a) an initial spontaneous activation of memory contents, (b) a deliberation phase, and (c) a response phase.

Automatic Activation Phase

During the initial stage of evaluative processing, an attitude object or its symbolic representation (e.g., a lexical or verbal reference) may elicit evaluations automatically, without intent, effort, or even conscious awareness. Supplementing early demonstrations (e.g., Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Kunst-Wilson & Zajonc, 1980), many studies now document such spontaneous evaluations, which are commonly thought to result from an automatic activation of associated contents in long-term memory (e.g., Bargh, Chaiken, Govender, & Pratto, 1992; De Houwer, Hermans, & Eelen, 1998; Giner-Sorolla, Garcia, & Bargh, 1999; Greenwald, Klinger, & Liu, 1989; Wittenbrink, Judd, & Park, 2001a), although they may also arise from non-declarative processes such as those underlying fluency effects (Bornstein & D'Agostino, 1994; Murphy & Zajonc, 1993; Reber, Winkielman, & Schwarz, 1998) or physiological feedback effects (Laird, 1974; Strack, Martin, & Stepper, 1988).

Memory activation occurs fast, within a few hundred milliseconds after encountering the attitude object (Fazio et al., 1986; Klauer, Roßnagel, & Musch, 1997). This initial activation requires only very limited cognitive resources and does not emanate from an active search for relevant memory

contents. Instead, it is the result of a passive process that runs its course automatically following exposure to the attitude object (Roskos-Ewoldsen & Fazio, 1992; Shiffrin & Schneider, 1977). Due to the passive nature of this initial activation, a person does not have to be aware of the attitude object or of the activation (e.g., Devine, 1989; Greenwald et al., 1989; Wittenbrink, Judd, & Park, 1997)—a fact that can have important consequences for subsequent stages of the evaluation process.

Automatic processes are thought to develop from frequent, repetitive experiences with a given stimulus (Shiffrin & Schneider, 1977). As a result, the particular memory contents that can be triggered automatically by an attitude object depend on the strength of their association with the object. If, as a result of past experiences, an overall evaluation of the attitude object has already been formed and strongly associated with the object, the evaluation itself may be spontaneously activated (e.g., spinach—"yuck!"). At the same time, other associations that have been strongly linked to the object can be activated as well. To the extent that they have evaluative implications, these evaluations may also shape subsequent evaluative responses (e.g., spinach—"bitter taste").

Because automatic activation depends on the accessibility of evaluative information, not all attitudes are equally likely to be activated automatically. Instead, automatic activation should occur especially for strong attitudes, which are more accessible and more consistent in their evaluative implications (see Petty & Krosnick, 1995). Empirical findings generally support the notion that attitude accessibility and consistency moderate automatic activation (Fazio et al., 1986), although in some instances, automatic activation has been observed for evaluatively consistent but inaccessible attitudes (Bargh et al., 1992; De Houwer et al., 1998).

Deliberation Phase

To the extent that a person has the opportunity and is sufficiently motivated, the initial activation phase is followed by a deliberation stage. During this second stage of evaluative processing, a controlled search for relevant information takes place. Both stored evaluations ("I liked the spinach at dinner last week") and other relevant associations ("spinach—it's healthy") might be retrieved from memory. Whether a particular piece of information will be retrieved at this point depends on its temporary accessibility (Salancik & Conway, 1975; Tourangeau et al., 1989), which in turn is influenced by a variety of factors.

First, memory contents vary in their chronic accessibility. Certain beliefs and experiences come to mind more easily than others, and certain memory contents are more closely linked to the attitude object than others. Second, as numerous studies have shown, this chronic accessibility may be moderated by the context in which the attitude object is encountered (for reviews, see Sudman, Bradburn, & Schwarz, 1996; Tesser, 1978; Wilson & Hodges, 1992). For example, the order of questions in a questionnaire may impact the deliberation phase by influencing the temporary accessibility of certain memory contents (e.g., Tourangeau et al., 1989). Likewise, the wording of a question or the particular exemplar of an attitude object that is encountered may highlight specific aspects of the object and thereby raise the temporary accessibility of certain pieces of information (e.g., Bodenhausen, Schwarz, Bless, & Wänke, 1995; Kinder & Sanders, 1990). Moreover, the search strategy that a person uses for retrieval can affect what information comes to mind during deliberation (e.g., Lord, Lepper, & Preston, 1984; Zajonc, 1960).

The deliberation phase requires motivation and opportunity because it involves effortful and willful processes. If these prerequisites are not met, input from the initial automatic activation stage will instead have a direct impact on a person's evaluative response. Motivation to spend time and effort on this process is the first critical determinant of the extent to which an attitude report will be deliberated. Having the opportunity to do so is the second.

There are many reasons why a person may be motivated to carefully reflect on his or her attitude before reporting it. Circumstances in the reporting situation may induce such motivation. That

is, situational cues that highlight the positive consequences of being accurate and/or increase the perceived costliness of making a judgmental error are likely to increase a person's motivation to deliberate. For example, situations where people feel accountable for their evaluations (e.g., because people expect to have to explain their attitudes to others) tend to foster deliberation (e.g., Kruglanski & Freund, 1983; Tetlock, 1983). Likewise, salient cues in a situation that highlight the normative implications of stating one's attitude also lead to more systematic deliberation of evaluations (e.g., Chen, Shechter, & Chaiken, 1996).

Aside from situational cues, motivation to deliberate can also be induced by internal factors. For example, some individuals have a higher overall need for accuracy (e.g., Kruglanski, 1989) or enjoy thinking (Cacioppo & Petty, 1982) and are therefore more motivated to exert mental effort in reaching an evaluation. Others are especially inclined to consider their own opinions and thus are more likely to introspect and deliberate about an issue (e.g., Snyder, 1979).

Assuming that a person is motivated to deliberate about an attitude, the opportunity to do so must also exist. This second prerequisite for deliberation is constrained first by a person's awareness of the attitude object. As long as the object remains outside of conscious awareness, no deliberation can take place. Although this precondition is probably met in very few situations in everyday life, this possibility is important for attitude measurement. Techniques that prevent the attitude object from reaching participants' conscious awareness (e.g., short exposure times) allow the assessment of evaluation effects free of further deliberation (Greenwald et al., 1989; Wittenbrink et al., 1997).

A second constraint on the opportunity to deliberate is the availability of cognitive resources. Many situations in everyday life place significant cognitive demands on people, as when multiple tasks occur simultaneously or when judgments must be made under time pressure (Bargh, 1997; Gilbert, 1989). As a result, a person's capacity for deliberation may often be limited, or, in extreme cases, entirely lacking (e.g., Kruglanski & Freund, 1983; Sanbonmatsu & Fazio, 1990). In these cases, the input from the initial automatic activation stage will be the primary determinant of a person's evaluative response, even though the person may be quite motivated to reflect upon the evaluation in a more controlled fashion.

Response Phase

The evaluations generated either automatically or deliberately then shape overt responses. These influences can be either explicit, with the person aware of the connection between attitude and response, or they can be implicit, with the person remaining unaware of the link (Greenwald & Banaji, 1995). In the case of explicit influence, the response follows from a deliberate consideration of the input generated during the previous two processing stages. For this response to occur, the information has to be integrated, creating the crystallized form of the attitude in working memory, and then it is linked to the available response alternatives.

Of particular interest for understanding attitude measurement is the role that metacognitions play in the integration of inputs to yield a final response (e.g., Metcalfe & Shimamura, 1994). For example, a person may reflect upon his or her subjective experience of the deliberation process itself. Specifically, the ease with which information comes to mind during deliberation may be regarded as diagnostic for one's evaluation. That is, having a difficult time generating reasons for why one might like an object has been found to negatively affect one's evaluation of the object (e.g., Wänke, Bohner, & Jurkowitsch, 1997).

Likewise, metacognitions about the appropriateness of information shaping a particular response may also influence this final step of evaluative processing. That is, people hold naïve theories about how a particular situation might bias their judgments and how to correct for the bias. Thus, if a person's theories suggest that an evaluation is the result of inappropriate information, he or she may attempt to correct the final evaluation accordingly (Martin, 1986; Schwarz & Bless, 1992; Strack,

1992; Wegener & Petty, 1997). For example, in evaluating an ordinary target person, a judge may adjust for the fact that he or she just saw a picture of Adolf Hitler, possibly making the target person seem more appealing and therefore justifying a downward correction in evaluations of him or her (Wegener & Petty, 1995). Correction strategies of this kind are closely related to the control mechanisms that operate during the deliberation stage and that guide the controlled search of information. However, correction during the response stage may simply consist of an adjustment of one's reported evaluation, without any further information search.

Finally, the result of integration has to be mapped onto the available response alternatives. To the extent that the alternatives are clearly prescribed by the situation, as they are in standard self-report measures of attitudes, this step requires that the response be formatted in accordance with the specified options, according to inferences made about the intended meaning of response alternatives (Strack & Martin, 1987).

So far, our description of the response phase has focused on explicit influences of the prior evaluation process on overt responses. These explicit influences require an effortful review of how the available information should be used. In other situations, the evaluation process may influence overt responses implicitly. First, when the attitude object remains outside of awareness, information generated during the evaluation process may impact responses implicitly. When an attitude object triggers an automatic activation, it may influence responses as long as it remains activated. Subliminal priming techniques assess implicit evaluation effects of this kind (e.g., Wittenbrink et al., 1997). Second, the attitude object itself may be noticed, but the evaluation it triggers may remain outside of conscious awareness and influence subsequent responses. Various response latency procedures for attitude measurement assess such implicit evaluation effects (e.g., Fazio et al., 1995). Finally, a third way by which evaluations may implicitly affect responses is through misattribution of the evaluation. That is, a person may deliberately recall or construct an evaluation, and this evaluation may subsequently influence a response, but the person does not recognize the link between evaluation and response. This kind of implicit evaluative influence is illustrated by the impact that answering one question can have on answers to later questions in a questionnaire (e.g., Strack, Martin, & Schwarz, 1988).

Conclusion

The cognitive processes by which evaluations of objects are generated are multifaceted, complex, and variable over time and across situations and individuals in systematic ways. Therefore, there is no reason to believe that a single person will always report the same attitude toward an object when asked about it on multiple occasions in different contexts. Yet, this variability does not mean that the person lacks an attitude or that the attitude concept should be revised to remove notions of stability or consistency. The goal of attitude measurement is to gauge the stable construct underlying responses. Accordingly, the variability in the processes that generate those responses must be understood.

Criteria for Attitude Measurement

The fundamental question in attitude measurement is whether the obtained response appropriately indexes the latent attitude construct. Because that construct itself is not directly observable, any attempt to measure it will necessarily do so only inadequately and incompletely. Consequently, it is important to index that inadequacy—in other words, to index the degree to which our measurement procedures capture the latent construct that we seek to measure.

In the history of attitude measurement, there have been two rather different approaches for addressing the issue of measurement adequacy: (a) the axiomatic or representational approach and

(b) the psychometric approach. The first of these has its origins in some of the earliest work on attitude measurement (e.g., Thurstone, 1927) and has since been developed in mathematically rigorous and even elegant detail (e.g., Luce, Krantz, Supper, & Tversky, 1990). Nevertheless, the second of these approaches currently dominates the field of attitude measurement. There are a variety of reasons for its dominance (see Cliff, 1992; Dawes, 1994), not the least of which is that it was never clear that the representational approach, for all its mathematical rigor, really did a better job than the much more straightforward psychometric approach. Accordingly, in what follows, we focus exclusively on the psychometric approach (for comprehensive treatments of the other tradition, see Dawes & Smith, 1985; Judd & McClelland, 1998).

The fundamental issue in psychometrics is the issue of construct validity (Cronbach, 1984; Messick, 1989): To what extent do the variables we measure adequately represent or capture the psychological construct that is of interest? And the fundamental approach to answering this question is to examine patterns of covariances or correlations between alternative measures. Initially, the focus of such work was on the assessment of the reliability of a measure. Subsequently, issues of convergent and discriminant validity were addressed as a part of the larger issue of construct validity.

Reliability

Initial psychometric formulations assumed that any measured variable had two underlying components: true score and random error (the *i* subscript refers to individuals):

$$X_i = T_i + E_i$$

Errors were assumed to be exclusively random perturbations, so they were assumed to be uncorrelated with true scores (and all other variables). The variance in the measured variable was therefore presumed to equal the sum of the variance in the true scores and the variance of the random errors of measurement:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

From this equation followed the definition of reliability: The proportion of the variance in a measured variable that was true score:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

This provides only a definition of reliability. To estimate it, a researcher must have at least two measures of a construct, sometimes referred to as "parallel forms," sharing the true score to the same extent and having random errors of the same magnitude. It can be shown that the correlation between the two measures equals the reliability of each:

$$r_{X_1X_2} = \rho_{X_1X_1} = \rho_{X_2X_2}$$

In practice, the reliability of a measure could be estimated by correlating two (almost) perfectly equivalent measures of the same construct. Alternative ways of doing this acquired different names: Split-half reliability involved parallel forms based on two randomly selected subsets of a battery of questions; test-retest reliability assumed that measurements at different time points were parallel.

With multiple questions in a battery, all of which are assumed to measure the same underlying construct, the random measurement errors in responses to any one question will cancel each other out when a composite score (sum or average) is computed across all the questions. The degree to

which this is true is given by the Spearman-Brown prophecy formula for the reliability of the sum (or average) of *k* parallel items:

$$\rho_{sum} = \frac{k r_{ij}}{1 + (k - 1) r_{ij}}$$

where $\mathbf{r}_{ij}$ is the correlation between every pair of items (assumed to be constant across all pairs, because of the parallel forms assumption).

The generalization of Spearman-Brown, allowing unequal true score variances across different questions, is coefficient α, the reliability of a sum (or average) of a set of items, all presumed to measure the same construct, albeit with unequal item reliabilities:

$$\alpha = \left(\frac{k}{k - 1} \right) \left(1 - \frac{\sum \sigma_i^2}{\sigma_{sum}^2} \right)$$

where $\sum \sigma_i^2$ is the sum of the variances of the individual items and σ_{sum}^2 is the variance of their sum.

Both of these formulas assume that responses have been coded so that they are all positively correlated. And before items are combined and the reliability of their sum (or average) is estimated, a principal components analysis can be conducted to verify that all questions load highly on the first unrotated component. Most computer programs that compute coefficient alpha will also report item-total correlations, as well as coefficient alpha values omitting each item in turn from the sum. According to this perspective, items that do not load highly on the first principal component or that do not correlate highly with the sum should be omitted, because they may assess other constructs than the one shared by the other items. Doing so will generally increase coefficient alpha computed on the remaining items.

Convergent and Discriminant Validity

The classic psychometric model that we have just reviewed is theoretically inadequate, because it presumes that all nonrandom variation in an attitude measure is due to the construct that we wish to measure, in other words, to the true score. All measures, however, have in them multiple sources of systematic nonrandom variance. Therefore, a more adequate theoretical model for any measure is that it likely taps three classes of phenomena, to varying extents:

(a) the construct of theoretical interest,
(b) other constructs that are not of theoretical interest, and
(c) random errors of measurement.

The broad issue of construct validity concerns the extent to which all three of these contribute to the variance of responses to an item. An item with high construct validity is one in which the construct of interest contributes a great deal to the item's variance, while other constructs and random error contribute very little. How reliable an item is (i.e., the relative absence of random errors of measurement) is accordingly one component of construct validity: It indexes the relative contribution of random errors without differentiating between the two systematic components of item variance. And the reliability of an item therefore sets only an upper limit on the extent to which the item validly measures the construct of interest.

The other two components of construct validity, beyond reliability, concern convergent validity and discriminant validity (Campbell & Fiske, 1959). The former represents the extent to which variance in the items is attributable uniquely to the construct of theoretical interest. The more it does so, the higher the convergent validity. The latter represents the extent to which other constructs, those

that are not of theoretical interest, contribute systematic error variance to an item's overall variance. The more an item contains unwanted systematic error variance due to other constructs, the lower its discriminant validity. In sum, then, the overall construct validity of an item depends on three sources of variation in scores:

1. the more of the variation is attributable to the latent construct of interest, the higher the convergent validity;
2. the less of the variation is attributable to other constructs, i.e., sources of systematic error, the higher the discriminant validity; and
3. the less of the variation is attributable to random error, the higher the reliability.

Campbell and Fiske (1959) were the first to explore ways in which convergent and discriminant validity could be estimated from the patterns of correlations (or covariances) among different measured variables. The tool they used was the multitrait-multimethod matrix, which can be built when a number of different constructs of theoretical interest are measured, each using a number of different assessment procedures. For instance, a researcher might measure attitudes towards three different attitude objects (e.g., three different minority ethnic groups) using each of three different assessment procedures. From these nine items (three attitude objects crossed with three assessment methods), one can construct a 9×9 correlation matrix. As Campbell and Fiske argue, the pattern of these correlations can be used to infer the extent to which there is convergent validity (measures of the same attitude using different methods all correlate highly), there is discriminant validity between the three attitudes (correlations between measures of different attitudes are relatively low), and there is discriminant validity between the measurement methods (correlations between different attitudes measured with the same method are no higher than correlations between different attitudes measured with different methods).

Campbell and Fiske's (1959) approach to the multitrait-multimethod matrix relies upon a fundamental tenant of the psychometric approach to construct validity: To the extent that measures covary, it is because they share systematic variance, either due to the construct(s) of interest or to other constructs that are not of interest (systematic error variance). In general, to argue for discriminant validity, a researcher must show relatively low correlations between items that are thought to measure different constructs, with the caveat of course that those different constructs may themselves be correlated. To argue for convergent validity, a researcher must show large correlations between different items that are all believed to measure the construct of interest. And to rule out other shared systematic sources of error variation as responsible for such high correlations, the different items all thought to measure the construct of interest must be maximally dissimilar in other ways (so that the other constructs they measure are maximally dissimilar). In general, the quest for construct validity mandates what might be called a "multioperationalization" approach: the adequacy of measurement can only be assessed by examining patterns of covariation between alternative measures of the same and different constructs.

Lee Cronbach and colleagues extended notions underlying the multitrait-multimethod matrix to more generalized research designs permitting comprehensive assessments of construct validity (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). One can think about the multitrait-multimethod matrix as a two-factor design, crossing traits (i.e., attitudes) with methods and measuring participants under all levels of both factors. Given this conception, a researcher can conduct an analysis of variance with the resulting data, devoting primary attention to the variance components due to participants, traits, and methods (and their interactions) rather than to the *F* tests typically reported. These variance components and their ratios (which are intraclass correlations, Shrout & Fleiss, 1979) provide information about the construct validity and reliability of the measured variables. For instance, if the different traits (or attitudes) show discriminant validity, then

the variance component due to traits should be large relative to the variance components due to participants and due to the participant by trait interaction (Kenny, 1994).

Cronbach generalized this variance components approach into what became known as "generalizability theory," in which additional factors are added to the analysis of variance design, with factors representing, for instance, occasions, experimenters, locations, etc. In essence, this generalization amounts to an extension of the multitrait-multimethod matrix to incorporate additional factors so that one could examine whether those additional factors systematically affect the variance in responses. From the resulting variance components estimation, a researcher can estimate convergent and discriminant validity for the various factors that were used in the research design. For instance, if multiple attitudes were measured using multiple methods on multiple occasions, one could assess whether different methods yield the same answer (discriminant validity against method variance) and whether different occasions yield the same answer (discriminant validity against time variance—indicating stability of responses).

Although generalizability theory offers a comprehensive approach for examining issues of construct validity, the recommended fully crossed designs are certainly cumbersome. Ideally, researchers would like to estimate the contributions of various factors (i.e., constructs both of interest and those not of interest) to variance in responses with data matrices upon which analysis of variance decompositions are not possible. Doing so is possible in some cases through the use of confirmatory factor analysis (CFA) procedures (see Judd & McClelland, 1998; Kenny & Kashy, 1992; Kline, 1998). In essence, a researcher constructs a theoretical measurement model of the latent constructs thought to be responsible for the variances of and covariances between a set of measured variables. Assuming that the model is "identified" (i.e., there are fewer parameters in the model to estimate than the number of independent bits of information in the observed variance—covariance matrix), then one can estimate the model's parameters, providing direct estimates of convergent validity, discriminant validity, and reliability. The development of such CFA procedures represents a significant recent contribution to the set of tools researchers have available to them for examining issues of construct validity. In fully crossed designs, such as the multitrait-multimethod matrix or the more elaborate designs of generalizability theory, parameter estimates resulting from confirmatory factor analytic estimation provide equivalent information to that which derives from the analysis of variance approach (Judd & McClelland, 1998).

Traditional Direct Self-Report Methods

With this perspective on measurement theory established, we can now turn to the procedures available for measuring attitudes. We begin with a focus on direct self-reports that involve asking participants explicitly to describe their own attitudes. Our discussion starts with a review of the relatively cumbersome measurement techniques proposed by the pioneers of attitude measurement nearly 70 years ago. Although widely appreciated, these techniques are rarely implemented these days, in favor of simpler practices. We therefore review a range of guidelines for optimally building such simpler measures and identify sources of random and systematic measurement error in responses to them.

Classic Self-Report Measurement Methods

The origins of elaborate attitude measurement via direct self-reports lie in the work of Louis Thurstone (1928); Rensis Likert (1932); and Charles Osgood (Osgood, Suci, & Tannenbaum, 1957). Each of these scholars developed a unique technique for measuring attitudes with multiple self-report items that have strong face validity. To put common practices in use today into context, we outline these techniques first.

Thurstone's Equal-Appearing Intervals Method

The title of Thurstone's landmark 1928 publication was "Attitudes Can Be Measured," a phrase that seemed as if it should end with an exclamation point. The method of attitude measurement he proposed involved seven steps of materials preparation (!). The first stage entailed gathering or generating between 100 and 150 statements of favorable or unfavorable evaluations of an object. Next, this set is edited down to a set of 80 to 100 statements that seem to have the most potential to perform effectively in later stages. Then, between 200 and 300 judges place each statement into one of 11 piles, with the piles defined as representing equally spaced points along the evaluative continuum running from extremely negative to extremely positive. Next, each statement is assigned a numeric value from 1 to 11, representing the place at which each participant placed it, and then the mean and variance of the numbers assigned to each statement are calculated. Statements with large variances are interpreted in different ways by different judges, so they are dropped from consideration. Then, two or three statements with means very close to each point along the continuum are selected, thus yielding a final battery with sets of statements that are equally spaced from one another. At this point, the measure is ready for administration. Participants are asked to read all of the selected statements and to indicate those with which they agree. Each participant is assigned an attitude score by averaging the mean scale values of the statements that he or she endorses. Ideally, each participant agrees with just two or three statements, pinpointing his or place along the continuum.

Likert's Method of Summated Ratings

Rensis Likert's (1932) summated rating method is less labor-intensive during the materials preparation phase. First, the researcher prepares about 100 statements that express positions either strongly favorable or unfavorable toward an object. In contrast to Thurstone's method, statements expressing neutrality are not included here. A set of pretest participants are then given a set of five response options (strongly approve, approve, undecided, disapprove, or strongly disapprove) and are asked to choose one response to express their view of each statement. For statements expressing favorable views of the object, responses are coded 1, 2, 3, 4, and 5, respectively. For statements expressing unfavorable views of the object, responses are coded 5, 4, 3, 2, and 1, respectively.

Each pretest participant is then assigned a total score by summing his or her scores on all of the items. Finally, for each item, each person's score is correlated with his or her total score, and items with low item-to-total correlations are dropped. Approximately 20 items with the strongest correlations are retained for use in the final battery. When this final battery is later administered to other samples, participants express their extent of agreement or disagreement with each statement, and total scores are generated accordingly for each participant. This procedure share some of the spirit of Thurstone's but involves a unique feature: assessment of the validity of each item via the item-to-total correlation.

Osgood, Suci, and Tannenbaum's Semantic Differential

The semantic differential is the simplest and easiest to administer of the landmark attitude measurement techniques. Through extensive developmental research, Osgood and his colleagues identified a set of adjective pairs that represent the evaluative dimension, including good-bad, valuable-worthless, wise-foolish, pleasant-unpleasant, and others. Each pair anchors the ends of a seven-point rating scale, and participants select the point on each scale to indicate their evaluation of the object.

Osgood, Suci, and Tannenbaum's (1957, pp. 29, 83) response scale consisted of a long horizontal line, intersected by six short vertical lines dividing the horizontal line into seven sections. At the two ends of each horizontal line were two antonyms, such as "good" and "bad." Participants were

instructed to mark a spot on the horizontal line to evaluate the goodness or badness of an object. In addition, Osgood et al. (1957) provided extensive instructions explaining the meanings of all the points on the rating scale. For example, for a rating scale anchored on the ends by "good" and "bad," participants were told that the endpoint labeled "good" meant "extremely good," the next point over meant "quite good," the next point meant "slightly good," the midpoint meant "neither good nor bad/equally good and bad," the next point meant "slightly bad," and so on. The semantic differential is the foundational technique used most often in research today, but it is typically administered *not* following Osgood et al.'s (1957) procedure. Instead, the horizontal line is presented with no labels on any points except the endpoints, and these endpoints are not labeled extremely ("good" instead of "extremely good" and "bad" instead of "extremely bad"). Typically the scale points are scored 1, 2, 3, 4, 5, 6, and 7, running from the most negative response to the most positive response, and the participant's attitude score is the average of the scores he or she receives on each item in the battery.

Advantages and Disadvantages of These Methods

All three of these foundational methods involve the administration of a large set of questions to measure a single attitude. Therefore, these approaches are time-consuming and demanding for participants. In addition, the Thurstone and Likert procedures entail a great deal of preparatory work up front, prior to the administration of the battery to one's focal sample of participants. However, these methods have at least two key advantages: First, administering many items yields a final score that contains less random measurement error (Allison, 1975). Second, these procedures have the advantage of being built using empirical evidence of convergence of interpretations across people and of correlational validity of the statements.

Unfortunately, the time pressures typical of most data collection efforts these days mean that researchers find it difficult to justify expending the resources necessary to build and then administer full-blown Thurstone, Likert, or Osgood rating batteries to measure a single attitude. Therefore, most researchers measure attitudes using a very small number of questions that have not been selected based on extensive pretesting and development work. This practice means that there is a strong incentive to design these few items to yield maximally reliable and valid assessments. We turn next to the literature on such item design.

Designing Direct Self-Report Attitude Measures Optimally

Designing any question to ask people directly for descriptions of their attitudes requires that researchers make a series of decisions about structure and wording. These decisions were made differently by the three principal founders of attitude measurement, and such heterogeneity continues to this day. This might seem to suggest that there is no optimal measurement approach and that all of the many direct attitude measures are equally reliable and valid.

However, a huge literature has accumulated during the last 100 years throughout the social sciences challenging this conclusion. When taken together, this literature recommends best practices for designing attitude measures, so we turn now to review some of the highpoint of this literature (for a more comprehensive review, see Krosnick & Fabrigar, forthcoming).

We begin by addressing the issue of whether direct attitude measures should be open-ended or closed-ended. Then, we consider a series of design decisions required when building closed-ended questions with rating scales: how many points to put on the rating scales, how to label the scale points, in what order to present the points, and whether or not to offer "don't know" response options.

Open Versus Closed Questions

One of the first decisions a researcher must make when designing an attitude measure is whether to make it an open-ended question (permitting the participant to answer in his or her own words) or a closed-ended question (requiring the participant to select an answer from a set of choices). By a wide margin, closed-ended questions dominate attitude measurement. But open-ended questions can certainly be used to measure attitudes (see, e.g., Holbrook, Krosnick, Visser, Gardner, & Cacioppo, 2001), and the accumulated literature suggests that these may well be worthwhile under some circumstances.

No doubt, a major reason for the widespread use of closed-ended questions is the complexity entailed in the coding of answers to open-ended questions. If a questionnaire is administered to 300 people, nearly 300 different answers will be given to a question asking people what they like and dislike about the President of the United States (for example), if the answers are considered word-for-word. But in order to analyze these answers, a coding scheme must be developed for each open-ended question; multiple people must read and code the answers into categories; the level of agreement between the coders must be ascertained; and the procedure must be refined and repeated if agreement is too low. The time and financial costs of such a procedure no doubt have led many researchers to favor closed-ended questions, which in essence ask participants to code themselves directly into categories that the researcher provides.

Unfortunately, closed-ended questions can have distinct disadvantages. The precise formulation of an attitude rating scale in terms of the number of points on the scale, the extent of verbal labeling of those points, the particular verbal phrases selected to label the points, the order in which the points are presented to participants, and offering "don't know" response options can all be done suboptimally. As a result, reliability and validity can be compromised. Because open-ended questions do not present answer choices to participants, these sources of researcher-induced measurement error do not distort responses in principle. And in practice, past studies show that open-ended questions have higher reliabilities and validities than closed-ended questions (e.g., Haddock & Zanna,1998 Hurd, 1932; Remmers, Marschat, Brown, & Chapman, 1923; Schuman, 2008; see also Smyth, Dillman, Christian, & Mcbride, 2009 on open-ended questions in web surveys).

One might hesitate before using open-ended questions because such questions may themselves be susceptible to unique problems. For example, some scholars feared that open-ended questions might not work well for participants who are not especially articulate, because they might have special difficulty explaining their feelings. However, this fear seems unfounded in most cases (England, 1948; Geer, 1988; Haddock & Zanna, 1998). Second, some scholars feared that participants would be especially likely to answer open-ended questions by mentioning the most salient possible responses, not those that are truly most appropriate. But this, too, seems not to be the case (e.g., Schuman, Ludwig, & Krosnick, 1986). Thus, open-ended questions may be worth the trouble they take to ask and the complexities inherent in analysis of their answers.

Principles of Rating Scale Design

The predominant response format for direct self-report attitude measures these days is the rating scale. When a participant is confronted with such a scale, his or her job is to execute a matching or mapping process. First, the participant must assess his or her own attitude in conceptual terms (e.g., "I like it a lot") and then find the point on the rating scale that most closely matches that attitude (see Ostrom & Gannon, 1996). Given this perspective, a number of general conditions must be met in order for a rating scale to work effectively. First, the points offered should cover the entire measurement continuum, leaving out no regions. Second, these points must appear to be ordinal, progressing from one end of a continuum to the other, and the meanings of adjacent points should

overlap with one another minimally if at all. Third, each participant must have a relatively precise and stable understanding of the meaning of each point on the scale. Fourth, most or all participants must agree in their interpretations of the meanings of each scale point. And a researcher must know what those interpretations are.

If some or all of these conditions are not met, data quality is likely to suffer. For example, if a participant falls in a particular region of an underlying evaluative dimension (e.g., "like somewhat") but no response options are offered in this region (e.g., a scale composed only of "dislike" and "like"), the participant will be unable to rate himself or herself accurately. If a participant interprets the points on a scale one way today and differently next month, then he or she may respond differently at the second time-point, even if his or her underlying attitude has not changed. If two or more points on a scale appear to have the same meaning to a participant, he or she may be puzzled about which one to select, leaving him or her open to making an arbitrary choice. If two participants differ in their interpretations of the points on a scale, they may give different responses even though they may have identical underlying attitudes. And if participants interpret scale point meanings differently than researchers do, the researchers may assign numbers to the scale points for statistical analysis that misrepresent the messages participants attempted to send via their ratings.

Number of Points on Rating Scales

The degree to which the above conditions are met is influenced partly by the number of points on the scale. In the work of academic social scientists, commercial practitioners, and government researchers, the length of rating scales varies tremendously. This variation is evident in the pioneers' attitude measures: Classic Likert (1932) scaling uses 5-point scales; Osgood, Suci, and Tannenbaum's (1957) semantic differential uses 7-point scales; and Thurstone's (1928) equal-appearing interval method uses 11-point scales. Rating scales used to measure public approval of the U.S. President's job performance also vary considerably across commercial survey houses, from 2-point scales to 5-point scales (Morin, 1993; Sussman, 1978). For the last 65 years, the American National Election Study surveys have measured citizens' political attitudes using 2-, 3-, 4-, 5-, 7-, and 101-point scales (Miller, 1982). Robinson, Shaver, and Wrightsman's (1999) catalog of popular rating scales for measuring a range of social psychological constructs and political attitudes describes 37 using 2-point scales, 7 using 3-point scales, 10 using 4-point scales, 27 using 5-point scales, 6 using 6-point scales, 21 using 7-point scales, 2 using 9-point scales, and 1 using a 10-point scale.

Thus, there appears to be no accepted standard for the number of points to be used on rating scales, and common practice varies widely. Nonetheless, the accumulated literature suggests that some rating scale lengths may be preferable to maximize reliability and validity. To review this literature, we begin with a discussion of theoretical issues and then catalogue the findings of relevant empirical studies.

THEORETICAL ISSUES

Translation Ease The length of scales can impact the process by which participants map their attitudes onto the provided response alternatives. The ease of this mapping or translation process varies, partly depending upon the underlying attitude. For instance, if a participant has an extremely positive or negative attitude toward an object, a dichotomous scale (e.g., "like," "dislike") easily permits reporting that attitude. Yet for a participant with a neutral attitude, a dichotomous scale not offering a midpoint would be suboptimal, because it would not offer the point most obviously needed to permit accurate mapping.

A trichotomous scale (e.g., "like," "neutral," "dislike") may be problematic for another person who has a moderately positive or negative attitude, equally far from the scale midpoint and from the extreme end on the underlying continuum. Adding a moderate point on the negative side (e.g., "dislike somewhat") and one on the positive side of the scale (e.g., "like somewhat") seems to be a good way to solve this problem. Thus, individuals who want to report neutral, moderate, or extreme attitudes would all have opportunities for accurate mapping.

The value of adding even more points to a rating scale may depend upon how refined people's mental representations of the construct are. Perhaps a 5-point scale is adequate, but perhaps people routinely make more fine-grained distinctions. For example, most people may be able to differentiate feeling slightly favorable, moderately favorable, and extremely favorable toward objects, in which case a 7-point scale would be more desirable than a 5-point scale.

If people do make such fine distinctions, potential information gain increases as the number of scale points increases, because of greater differentiation in the judgments made (for a review, see Alwin, 1992). This will be true, however, only if two conditions are met. First, participants must make use of the full scale. It is conceivable that when confronted with long scales, participants simply ignore large portions of the scale. Second, no additional information is gained if the number of scale points exceeds the degree to which participants differentiate between levels of an attribute in their minds. If people's psychological representations differentiate into no more than seven categories, for example, then additional scale points gain no more information for a researcher.

The ease of mapping a judgment onto a response scale is likely to be determined in part by how close the judgment is to the conceptual divisions between adjacent points on the scale. For example, when a person with an extremely negative attitude is asked, "Is your opinion of the President very negative, slightly negative, neutral, slightly positive, or very positive?" he or she can easily answer "very negative," because his or her attitude is far from the conceptual division between "very negative" and "slightly negative." However, for a person who is moderately negative, his or her true attitude is close to the conceptual division between "very negative" and "slightly negative," so this person may face a greater challenge in using this 5-point rating scale. The "nearness" of the participant's true judgment to the nearest conceptual division between adjacent scale points is associated with unreliability of responses—participants with greater nearness are more likely to pick one option on one occasion and another option on a different occasion (Kuncel, 1973, 1977).

Clarity of Scale Point Meanings In order for ratings to be reliable, participants must have a clear understanding of the meanings of the points on the rating scale. If the meaning of scale points is ambiguous, then both reliability and validity of measurement may be compromised.

A priori, it seems that dichotomous response option pairs are very clear in meaning. That is, there is likely to be considerable consensus on the meaning of options such as "favor" and "oppose" or "agree" and "disagree." Clarity may be compromised when a dichotomous scale becomes longer, because each point that is added on the rating scale is one more point that must be interpreted. And the more such interpretations a person must make, the more chance there is for inconsistency over time or across participants. That is, it is presumably easier for a participant to decide precisely where the conceptual divisions are between "favoring," "opposing," and being "neutral" on a trichotomous item than in the case of a 7-point scale, where six conceptual divisions must be specified.

For rating scales up to seven points long, it may be easy to specify intended meanings of points with words, as with "like a great deal," "like a moderate amount," "like a little," "neither like nor dislike," "dislike a little," "dislike a moderate amount," and "dislike a great deal." But once the scale point number increases beyond that length, point meanings may become considerably less clear. For example, on 101-point scales measuring attitudes, what exactly do 76, 77, and 78 mean conceptually? Even for 11- or 13-point scales, participants may be hard-pressed to define the meaning of the scale points.

Uniformity of Scale Point Meaning The number of scale points used is inherently confounded with the extent of verbal labeling possible, and this confounding may affect uniformity of interpretations of scale point meanings across people. Every dichotomous and trichotomous scale must, of necessity, include verbal labels on all scale points, thus enhancing their clarity. But when scales have four or more points, it is possible to label only the endpoints with words. In such cases, comparisons with dichotomous or trichotomous scales reflect the impact of both number of scale points and verbal labeling. It may be possible to provide an effective verbal label for each point on a scale containing, say, 11 or fewer scale points, but doing so becomes quite difficult as the number of scale points increases beyond that length.

One could argue that the participant's task is made that much more difficult when presented with numerical rather than verbal labels. To make sense of a numerically labeled rating scale, a participant must first generate a verbal definition for each point and then match these definitions against his or her mental representation of the attitude of interest. Verbal labels might therefore be advantageous, because they may clarify the meanings of the scale points while at the same time reducing participant burden by removing one step from the cognitive processes entailed in answering a rating question.

Satisficing Finally, the optimal number of rating scale points may depend on participants' cognitive skills and motivation to provide accurate reports. Unfortunately, when answering questionnaires, some individuals do not expend the effort necessary to provide optimal answers. Instead, they look for cues in questions pointing to reasonable answer choices that are easy to select with little thought, a behavior termed *questionnaire satisficing* (Krosnick, 1991, 1999). Such satisficing is thought to be more common among individuals with more limited cognitive skills and less motivation to provide accurate answers.

Offering a midpoint on a scale may constitute a satisficing cue to such participants, especially if its meaning is clearly either "neutral/no preference" or "status quo—keep things as they are now." If pressed to explain these answers, satisficing participants would have little difficulty defending such replies. Consequently, offering a midpoint may encourage satisficing by providing a clear cue offering an avenue for doing so.

However, there is a potential cost to eliminating midpoints. Some participants may truly belong at the scale midpoint and may wish to select such an option to communicate their genuine neutrality or endorsement of the status quo. If many people have neutral attitudes to report, eliminating the midpoint will force them to pick a point either on the positive side or on the negative side of the scale, resulting in an inaccurate measurement of their attitudes.

The number of points on a rating scale can also impact satisficing via a different route: task difficulty. High task difficulty is thought to inspire some participants to satisfice instead of optimizing (Krosnick, 1991). The number of scale points offered on a rating scale may be a determinant of task difficulty. Two-point scales simply require a decision of direction (e.g., pro vs. con), whereas longer scales require decisions of both direction and extremity. Very long scales require participants to choose between many options, so these scales may be especially difficult in terms of scale point meaning interpretation and mapping. Yet providing too few scale points may contribute to difficulty by making impossible the expression of moderate positions. Consequently, task difficulty (and satisficing as well) may be at a minimum for moderately long rating scales, resulting in more accurate responses.

EXISTING EVIDENCE ON THE OPTIMAL NUMBER OF SCALE POINTS

During the last 40 years, many research investigations have produced evidence useful for inferring the optimal number of points on rating scales. Some of this work has systematically varied the number of scale points offered while holding constant all other aspects of questions, examining effects

on reliability and validity. Other work has attempted to discern people's natural discrimination tendencies in using rating scales. We review this work next. Some of the studies we review did not explicitly set out to compare reliability or validity of measurement across scale lengths but instead reported data that permit us to make such comparisons post hoc.

Reliability Lissitz and Green (1975) explored the relation of number of scale points to reliability using simulations. These investigators generated sets of true attitudes and random errors for groups of hypothetical participants and then added these components to generate hypothetical responses to attitude questions on different-length scales in two hypothetical "waves" of data. Cross-sectional and test-retest reliability increased from 2- to 3- to 5-point scales but were equivalent thereafter for 7-, 9-, and 14-point scales. Similar results were obtained in simulations by Jenkins and Taber (1977), Martin (1978), and Srinivasan and Basu (1989).

Some studies have found the number of scale points to be unrelated to cross-sectional reliability. Bendig (1954) found that ratings using either 2-, 3-, 5-, 7-, or 9-point scales were equivalently reliable. Similar results have been reported for scales ranging from 2 to 7 points (Komorita & Graham, 1965; Masters, 1974) and for longer scales ranging from 2 to 19 points (Birkett, 1986; Matell & Jacoby, 1971; Jacoby & Matell, 1971). Other studies have yielded differences that are consistent with the notion that scales of intermediate lengths are optimal (Birkett, 1986; Givon & Shapira, 1984; Masters, 1974). For example, Givon and Shapira (1984) found pronounced improvements in item reliability when moving from 2-point scales toward 7-point scales. Reliability continued to increase up to lengths of 11 points, but the increases beyond 7 points were quite minimal for single items. Matell and Jacoby (1971; Jacoby & Matell, 1971) reported lower reliabilities for scales with 19 points as compared to scales with 7 to 8 points.

Another way to assess optimal scale length is to collect data on a scale with many points and recode it into a scale with fewer points. If longer scales contain more random measurement error, then recoding should improve reliability. But if longer scales contain valid information that is lost in the recoding process, then recoding should reduce data quality. Consistent with this hypothesis, Komorita (1963) found that cross-sectional reliability for 6-point scales was .83, but was only .71 when the items were first recoded to be dichotomous. Thus, it appears that more reliable information was contained in the full 6-point ratings than the dichotomies. Matell and Jacoby (1971) reported similar findings, indicating that collapsing scales longer than 3 points threw away reliable information.

Although there is some variation in the patterns yielded by these various studies, they can be viewed as supporting the notion that reliability is higher for scales with many points than for scales with only two or three. Furthermore, one might argue that scales with too many points compromise reliability as well.

Validity Research on the effect of the number of scale points on validity has relied on various gauges of validity, including simulations, concurrent and predictive validity, inter-rater agreement, and susceptibility to question order effects and interviewer effects.

Studies estimating correlations between true attitude scores and observed ratings on scales of different lengths using simulated data have found that validity increases as scales increase from 2 points to longer lengths; however as the scales grow longer, the gains in validity become correspondingly smaller (Green & Rao, 1970; Lehmann & Hulbert, 1972; Lissitz & Green, 1975; Martin, 1973; Martin, 1978; Ramsay, 1973). Besides simulation, several other techniques have been used to assess the validity of scales of different lengths: correlating responses obtained from two different ratings of the same construct (e.g., Matell & Jacoby, 1971; Smith, 1994a; Smith & Peterson, 1985; Watson, 1988; Warr, Barter, & Brownridge, 1983); correlating attitude measures obtained using scales of different lengths with other attitudes (e.g., Schuman & Presser, 1981, pp. 175–176); and using

the ratings obtained using different scale lengths to predict other attitudes (Rosenstone, Hansen, & Kinder, 1986; Smith & Peterson, 1985). Studies have typically found concurrent validity to increase with increasing scale length (Matell & Jacoby, 1971; Rosenstone, Hansen, & Kinder, 1986; Schuman & Presser, 1981; Smith, 1994a, 1994b; Smith & Peterson, 1985; Watson, 1988; Warr, Barter, & Brownridge, 1983).

Participants' answers to attitude measures are often influenced by prior questions that precede a measure in a questionnaire. One such effect is a *contrast effect*, which can occur when a given stimulus is evaluated partly in comparison with stimuli presented previously. Another source of invalidity in ratings is interviewers' opinions in face-to-face or telephone surveys. Presumably due partly to how interviewers ask questions, participants sometimes express opinions that are distorted toward those of the individuals who interview them (see Groves, 1989). These sources of systematic measurement error are apparently related to scale length in ways that suggest more and less optimal lengths.

Several studies suggest that longer scales are less susceptible to question order effects (Wedell & Parducci, 1988; Wedell, Parducci, & Lane, 1990; Wedell, Parducci, & Geiselman, 1987). However, one study indicates that scales that are especially long might be more susceptible to context effects than those of moderate length (Schwarz & Wyer, 1985). Stember and Hyman (1949/1950) found that answers to dichotomous questions were influenced by interviewer opinion, but this influence disappeared among individuals who were also offered a middle alternative, yielding a trichotomous question. There is again some variation in the patterns yielded by these studies, but they can be viewed as supporting the notion that validity is higher for scales with a moderate number of points than for scales with fewer and that validity is compromised by especially long scales.

Discerning Natural Scale Differentiation In a study by Champney and Marshall (1939), judges provided ratings on various scales by placing "x"s on 9-centimeter-long lines. Five, six, or seven points along the lines were labeled with sentences to establish the meanings of the parts of the scale. The continuous measurement procedure allowed Champney and Marshall (1939) to divide the lines into as many equally sized categories as they wished and then assess the cross-sectional reliability of the various divisions for two items that were both designed to measure sociability. Cross-sectional reliability increased dramatically from a 2-point scale (r = .56) to a 9-point scale (r = .70), and a further significant increase appeared when moving to 18 scale points (r = .74). Reliabilities, however, were essentially the same for 22 (r = .75), 30 (r = .76), 45 points (r = .77), and 90 points (r = .76). The judges returned 3 weeks later to re-rate the objects on a total of 12 scales, which allowed the computation of test-retest reliability of ratings, and results were consistent with the cross-sectional results.

McKelvie (1978) had participants rate various objects by marking points on lines with no discrete category divisions. Participants also indicated their "confidence interval" around each judgment. By dividing the total line length by the average magnitude of the confidence interval, McKelvie (1978) could estimate the number of scale points participants were naturally employing, which turned out to be 5.

Another study along these lines examined the number of scale points that participants used on scales of increasing length. Matell and Jacoby (1972) had participants provide a series of ratings on scales of lengths ranging from 2 points to 19 points. Nearly all participants used both points on the dichotomous items, and most participants used all three points on the trichotomous items. For longer scales, participants used about half the points offered, regardless of length. That is, the more scale points that were offered up to 19, the more points participants used, up to about 9.

Rundquist and Sletto (1936) had participants complete a set of ratings either by marking points on lines or by using 5- or 7-point category scales. When the line marks were coded according to a 7-point division, the distribution of ratings was identical to that obtained from the 7-point scale. But when the line marks were coded according to a 5-point division, the distribution was significantly

different from the 5-point scale, with fewer extreme and midpoint ratings being made for the latter than the former. This finding, again, supports the use of 7-point scales.

Middle Alternatives and Satisficing The validity of the satisficing perspective regarding middle alternatives can be gauged by determining whether attraction to them is greatest under the conditions that are thought to foster satisficing, two of which are low cognitive skills and low attitude strength (see Krosnick, 1991). However, Kalton, Roberts, and Holt (1980); Schuman and Presser (1981); O'Muircheartaigh, Krosnick, and Helic (1999); and Narayan and Krosnick (1996) concluded that attraction to middle alternatives was unrelated to participants' education (a proxy measure for cognitive skills). Krosnick and Schuman (1988) and Bishop (1990) found more attraction among those for whom the issue was less important and whose attitudes were less intense, and O'Muircheartaigh et al. (1999) found that attraction to middle alternatives was greater among people with less interest in the topic. But Stember and Hyman (1949/1950) found attraction to middle alternatives on a specific policy issue was unrelated to general interest in foreign policy, and O'Muircheartaigh et al. (1999) found no relation of attraction to middle alternatives with volume of knowledge about the object. Thus, at best, the available evidence on this point is mixed with regard to predictors of attraction to middle alternatives.

More importantly, O'Muircheartaigh and colleagues (1999) found that adding midpoints to rating scales improved the reliability and validity of ratings. Structural equation modeling of error structures revealed that omitting the middle alternative led participants to randomly select one of the moderate scale points closest to where a midpoint would appear. This suggests that offering midpoints is desirable.[1]

Labeling of Rating Scale Points

Once the length of a rating scale has been specified, a researcher must decide how to label the points on the scale. The semantic differential has verbal labels on the endpoints, no verbal labels on any intermediate points, and no numbers on any point. Some scales put numbers on all scale points, with verbal labels either on the endpoints, the endpoints plus some intermediate points, or on all the points. Thus, there are many different ways to do labeling, especially because various different numbering systems can be used (e.g., 1 to 7, 0 to 6), and many different words could be chosen to label a point on a scale.

Various studies suggest that the reliability of attitude rating scales is higher when all scale points are labeled with words than when only some are (e.g., Krosnick & Berent, 1993; Weng, 2004; Weijters, Cabooter, & Schillewaert, 2010). Furthermore, participants are more satisfied when more rating scale points are verbally labeled (e.g., Dickinson & Zellinger, 1980). When selecting verbal labels, researchers can maximize reliability and validity by selecting ones with meanings that divide up the continuum into approximately equal units (e.g., Klockars & Yamagishi, 1988; for a summary, see Krosnick & Fabrigar, forthcoming). For example, "very good, good, and poor" is a combination that should be avoided, because the terms do not divide the evaluative continuum equally.

Adding numbers to words on scale points seems to be unwise for various reasons. First, twice the amount of labeling means twice the amount of cognitive work that the respondent must do in order to interpret the meanings of the scale point. And in order to interpret the meanings of numbers on scale points, respondents naturally interpret them by putting words on them. Since the words are provided already, no additional value is gained by providing numbers as well. Furthermore, if the numbers are integers (e.g., 1 through 7) and therefore equally spaced, but the verbal labels chosen are not exactly equally spaced, then the words and numbers are sending contradictory signals to respondents, who must do extra cognitive work to decide what meanings the researcher actually intended to attach to the scale points. Finally, even arbitrarily chosen numbers can send unintended

signals to respondents about the intended meanings of scale points, which can conflict with verbal labels and thereby create more interpretative challenges (e.g., Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991).

Many closed-ended attitude measures are modeled after Likert's technique, offering statements to participants and asking them to indicate whether they agree or disagree with each or to indicate their level of agreement or disagreement on a scale. Other attitude measures offer assertions and ask participants to report the extent to which the assertions are true or false, and some attitude measures ask people "yes/no" questions (e.g., "Do you favor limiting imports of foreign steel?").

These sorts of item formats are very appealing from a practical standpoint, because such items are easy to write. If one wants to identify people who have positive attitudes toward bananas, for example, one simply needs to write a statement expressing an attitude (e.g., "I like bananas") and ask people whether they agree or disagree with it or whether it is true or false. Also, these formats can be used to measure a wide range of different constructs efficiently. Instead of having to change the response options from one question to the next as one moves from measuring liking to perceived goodness or badness, the same set of response options can be used. The popularity of agree/disagree, true/false, and yes/no questions is therefore no surprise.

Despite this popularity, there has been a great deal of concern expressed over the years that these question formats may be seriously problematic. The concern expressed is that some participants may sometimes say "agree," "true," or "yes" regardless of the question being asked of them. So, for example, a person might agree with a statement that the U.S. should forbid speeches against democracy and might also agree with a statement that the U.S. should allow such speeches. This behavior, labeled "acquiescence," can be defined as endorsement of an assertion made in a question, regardless of the content of the assertion. In theory, this behavior could result from a desire to be polite rather than confrontational in interpersonal interactions (Leech, 1983); from a desire of individuals of lower social status to defer to individuals of higher social status (Lenski & Leggett, 1960); or from an inclination to satisfice rather than optimize when answering questionnaires (Krosnick, 1991).

The evidence documenting acquiescence is now voluminous and consistently compelling, based upon a range of different demonstration methods (for a review, see Krosnick & Fabrigar, forthcoming). For example, consider first just agree/disagree questions. When people are given such answer choices, are not told any questions, and are asked to guess what answers an experimenter is imagining, people guess "agree" much more often than "disagree" (e.g., Berg & Rapaport, 1954). In other studies, pairs of statements were constructed stating mutually exclusive views (e.g., "I enjoy socializing" vs. "I don't enjoy socializing"), and people were asked to agree or disagree with both. Although answers to such pairs should be strongly negatively correlated, 41 studies yielded an average correlation of only −.22 (citation). This correlation may be far from—1.0 partly because of random measurement error, but it may also be because of acquiescence. Consistent with this claim, combining across ten studies, an average of 52% of people agreed with an assertion, whereas an average of only 42% of people disagreed with the opposite assertion. Thus, people are apparently inclined toward agreeing rather than disagreeing, manifesting what might be considered an acquiescence effect of 10 percentage points. Another set of eight studies compared answers to agree/disagree questions with answers to forced-choice questions where the order of the views expressed by the response alternatives was the same as in the agree/disagree questions. On average 14% more people agreed with an assertion than expressed the same view in the corresponding forced-choice question. Averaging across seven studies, 22% of people on average agreed with both a statement and its reversal, whereas only 10% of people disagreed with both. Thus, all of these methods suggest an average acquiescence effect of about 10%.

Other evidence indicates that the tendency to acquiesce is a general inclination of some individuals across questions. For example, the average cross-sectional reliability of the tendency to agree with assertions is .65 across twenty-nine studies. Furthermore, the over-time consistency of the tendency

to acquiesce is about .75 over 1 month, .67 over 4 months, and .35 over 4 years (e.g., Couch & Keniston, 1960; Hoffman, 1960; Newcomb, 1943).

These same sorts of results (regarding correlations between opposite assertions, endorsement rates of items, their reversals, forced-choice versions, and so on) have been produced in studies of true/false questions and of yes/no questions, suggesting that acquiescence is present in responses to these items as well. And there is other such evidence regarding these response alternatives. For example, people are much more likely to answer yes/no factual questions correctly when the correct answer is "yes" than when it is "no" (e.g., Larkins & Shaver, 1967; Rothenberg, 1969), presumably because people are biased toward saying "yes." Similarly, factual reports are more likely to disagree with informants' answers when a yes/no question is answered "yes" than when it is answered "no," again presumably because of a bias toward "yes" answers (Sigelman & Budd, 1986). When people say they are guessing the answer to a true/false question, 71% of answers are "true," and only 29% are "false."

Acquiescence is most common among participants of lower social status (e.g., Gove & Geerken, 1977; Lenski & Leggett, 1960); with less formal education (e.g., Ayidiya & McClendon, 1990; Narayan & Krosnick, 1996); of lower intelligence (e.g., Forehand, 1962; Hanley, 1959; Krosnick, Narayan, & Smith, 1996); of lower cognitive energy (Jackson, 1959), who don't like to think (Messick & Frederiksen, 1958); and of lower bias toward conveying a socially desirable image of themselves (e.g., Goldsmith, 1987; Shaffer, 1963). Also, acquiescence is most common when a question is difficult to answer (Gage et al., 1957; Hanley, 1962; Trott & Jackson, 1967); after participants have become fatigued by answering many prior questions (e.g., Clancy & Wachsler, 1971); and during telephone interviews as opposed to face-to-face interviews (e.g., Calsyn, Roades, & Calsyn, 1992; Holbrook, Green, & Krosnick, 2003). Although some of these results are consistent with the notion that acquiescence results from politeness or deferral to people of higher social status, all of the results are consistent with the satisficing explanation.

If this interpretation is correct, then acquiescence might be reduced by assuring (through pretesting) that questions are easy for participants to comprehend and answer and by taking steps to maximize participant motivation to answer carefully and thoughtfully. However, no evidence is yet available testing whether acquiescence can be reduced in these ways. Therefore, a better approach to eliminate acquiescence is avoiding the use of agree/disagree, true/false, and yes/no questions altogether. This is especially sensible because answers to these sorts of questions are less valid and less reliable than answers to the same questions expressed in a format that offers all competing points of view and asks participants to choose among them (e.g., Eurich, 1931; Isard, 1956; Watson & Crawford, 1930). A number of studies demonstrate how acquiescence can distort the results of substantive investigations (e.g., Jackman, 1973; Saris, Revilla, Krosnick, & Shaeffer, 2010; Winkler, Kanouse, & Ware, 1982), and in a particularly powerful historical example, acquiescence undermined the scientific value of *The Authoritarian Personality's* extensive investigation of fascism and anti-Semitism (Adorno, Frankel-Brunswick, Levinson, & Sanford, 1950).

One alternative approach to controlling for acquiescence is derived from the presumption that certain people have acquiescent personalities and are likely to do all of the acquiescing. According to this view, a researcher needs to identify those people and statistically adjust their answers to correct for this tendency (e.g., Couch & Keniston, 1960). To this end, many batteries of items have been developed to measure a person's tendency to acquiesce, and people who offer lots of "agree," "true," or "yes" answers across a large set of items can then be spotlighted as likely acquiescers. However, the evidence on moderation that we reviewed above suggests that acquiescence is not simply the result of having an acquiescent personality; rather, it is mainly influenced by circumstantial factors. Because this "correction" approach does not take that into account, the corrections performed are not likely to fully and precisely account for acquiescence.

It might seem that acquiescence can be controlled by measuring a construct with a large set of agree/disagree or true/false items, half of them making assertions opposite to the other half (called

"item reversals"; see Paulhus, 1991). This approach is designed to place acquiescers in the middle of the final dimension but will do so only if the assertions made in the reversals are equally extreme as the statements in the original items. Furthermore, it is difficult to write large sets of item reversals without using the word "not" or other such negations, and evaluating assertions that include negations is cognitively burdensome and error-laden for participants, thus adding measurement error and increasing participant fatigue (e.g., Eifermann, 1961; Wason, 1961). Even if one is able to construct appropriately reversed items, acquiescers presumably end up at a point on the measurement dimension where most probably do not belong on substantive grounds. That is, if these individuals were induced not to acquiesce but to answer the items thoughtfully, their final scores would presumably be more valid than placing them at or near the midpoint of the dimension.

Most important, answering an agree/disagree, true/false, or yes/no question always requires a participant to answer a comparable rating question with construct-specific response options in his or her mind first. For example, if a person is asked to agree or disagree with the assertion "I do not like bananas," he or she must first decide how much bananas are liked (perhaps concluding "I love bananas") and then translate that conclusion into the appropriate selection in order to answer the question one was asked ("disagree" to the original item). Researchers who use such questions presume that the arraying of participants along the agree/disagree dimension corresponds monotonically to the arraying of those individuals along the underlying substantive dimension of interest. That is, the more a person agrees with the assertion "I do not like bananas," the more negative his or her true attitude toward bananas is.

Yet consider the following scenario. Our hypothetical banana-lover encounters the following item: "I sort of like bananas." He or she may respond "disagree" because "sort of like" does not express the extremity of his liking. Thus, people who disagree with this question would include those who genuinely dislike bananas, as well as those whose positive regard vastly exceeds the word "sort of like," which clearly violates the monotonic equivalence of the response dimension and the underlying attitude construct of interest.

As this example makes clear, it would be simpler to ask participants directly how much they like or dislike objects. Every agree/disagree, true/false, or yes/no question implicitly requires the participant to make a rating of an object along a continuous dimension in his or her mind, so asking about that dimension directly is bound to be less burdensome. Not surprisingly, then, the reliability and validity of rating scale questions that array the full attitude dimension explicitly (e.g., from "extremely bad" to "extremely good," or from "dislike a great deal" to "like a great deal") are higher than those of agree/disagree, true/false, and yes/no questions that focus on only a single point of view (e.g., Ebel, 1982; Mirowsky & Ross, 1991; Ruch & DeGraff, 1926; Saris & Krosnick, 2000; Wesman, 1946). Consequently, it seems best to avoid agree/disagree, true/false, and yes/no formats altogether and instead ask questions using rating scales that explicitly display the evaluative dimension.

The Order of Response Alternatives

Many studies have shown that the order in which response alternatives are presented to participants can affect their selection among the alternatives, but until recently, it has not been clear when such effects occur, what their direction will be, and why they occur. Some past studies identified primacy effects (in which response choices presented early were most likely to be selected); other studies found recency effects (in which response choices presented last were more likely to be selected), and still other studies found no order effects at all. Fortunately, this apparently disorderly set of evidence can be explained by the theory of questionnaire satisficing (Krosnick, 1991).

Because the vast majority of attitude measurement involves the use of rating scales that ask participants to choose a descriptor from among a set that represents some sort of dimension or

continuum (e.g., from "dislike a great deal" to "like a great deal"), our greatest interest is with such scales. But to understand the satisficing explanation of response order effects, it is helpful to begin with an explanation of how response choice order effects occur in response to categorical questions, which ask people to make a choice among a set of objects that do not represent a continuum (e.g., "Which do you like more, peas or carrots?").

Response order effects in categorical questions appear to be attributable to "weak satisficing," which entails executing all the steps of optimal answering (interpreting a question, retrieving information from memory, integrating the information into a judgment, and reporting the judgment), but in a superficial, biased, and shortcut fashion (see Krosnick, 1991; Krosnick & Alwin, 1987). When confronted with categorical questions, optimal answering would entail carefully assessing the appropriateness of each of the offered response alternatives before selecting one. In contrast, a weak satisficer could simply choose the first response alternative he or she considers that appears to constitute a reasonable answer. Exactly which alternative is most likely to be chosen depends in part upon whether the response choices are presented visually or orally to participants.

When response alternatives are presented visually, either on a show-card in a face-to-face interview or in a self-administered questionnaire, weak satisficing is likely to bias participants toward selecting choices displayed early in a list. Participants are likely to begin at the top of the list and consider each response alternative individually, and their thoughts are likely to be biased in a confirmatory direction (Koriat, Lichtenstein, & Fischhoff, 1980; Klayman & Ha, 1987; Yzerbyt & Leyens, 1991). Given that researchers typically include in questions response choices that are reasonable answers, this confirmation-biased thinking is likely to generate at least a reason or two in favor of selecting almost any alternative a participant thinks about.

After considering one or two response alternatives, the potential for fatigue becomes significant, because participants' minds become cluttered with thoughts about initial alternatives. Also, fatigue may result from proactive interference, whereby thoughts about the initial alternatives interfere with and confuse thinking about later, competing alternatives (Miller & Campbell, 1959). Weak satisficers can cope by thinking only superficially about later response alternatives; the confirmatory bias would thereby give the earlier items an advantage. Alternatively, weak satisficers can simply terminate their evaluation process altogether once they come upon a response alternative that seems to be a reasonable answer to the question. And again, because most answers are likely to seem reasonable, these participants are likely to end up choosing alternatives near the beginning of a list. Thus, weak satisficing seems likely to produce primacy effects under conditions of visual presentation.

When response alternatives are presented orally, as in face-to-face or telephone interviews, the effects of weak satisficing are more difficult to anticipate, because response order effects reflect not only evaluations of each option, but also the limits of memory. When response alternatives are read aloud, participants are not given the opportunity to process the first alternative extensively. Presentation of the second alternative terminates processing of the first one, usually relatively quickly. Therefore, participants are able to devote the most processing time to the final items read; these items remain in short-term memory after interviewers pause to let participants answer.

It is conceivable that some participants listen to a short list of response alternatives without evaluating any of them. Once the list is completed, these individuals may recall the first alternative, think about it, and then progress through the list forward from there. Given that fatigue should instigate weak satisficing relatively quickly, a primacy effect would be expected. However, because this process requires more effort than simply considering the final items in the list first, weak satisficers are unlikely to do this very often. Therefore, considering only the allocation of processing, we would anticipate both primacy and recency effects, though the latter should be more common than the former.

These effects of deeper processing are likely to be reinforced by the effects of memory. Items presented early in a list are most likely to enter long-term memory (e.g., Atkinson & Shiffrin, 1968),

and items presented at the end of a list are most likely to be in short-term memory immediately after the list is heard (e.g., Atkinson & Shiffrin, 1968). Furthermore, items presented late are unusually likely to be recalled (Baddeley & Hitch, 1977). So items presented at the beginning and end of a list are more likely to be recalled after the question is read, particularly if the list is long. Therefore, given that a response alternative must be remembered in order for a participant to select it, both early and late items should be more available for selection, especially among weak satisficers. Typically, short-term memory dominates long-term memory immediately after acquiring a list of information (Baddeley & Hitch, 1977), so memory factors should promote recency effects more than primacy effects. Thus, in response to orally presented questions, mostly recency effects would be expected, though some primacy effects might occur as well.

Schwarz and Hippler (1991; Schwarz, Hippler, & Noelle-Neumann, 1992) pointed out two additional factors that may govern response order effects: the plausibility of the response alternatives presented and perceptual contrast effects. If deep processing is accorded to a response alternative that seems highly implausible, even participants with a confirmatory bias in reasoning may fail to generate any reasons to select it. Thus, deeper processing of some alternatives may make them especially *un*likely to be selected.

Although the results of past studies of response order effects in categorical questions seem to offer a confusing pattern of results when considered as a group, order appears when the studies are separated into those involving visual and oral presentation. Whenever a visual presentation has been used, primacy effects have been found (Ayidiya & McClendon, 1990; Becker, 1954; Bishop et al., 1988; Campbell & Mohr, 1950; Israel & Taylor, 1990; Krosnick & Alwin, 1987; Schwarz, Hippler, & Noelle-Neumann, 1992). And in studies involving oral presentation, nearly all response order effects have been shown to be recency effects (McClendon, 1986; Berg & Rapaport, 1954; Bishop, 1987; Bishop et al., 1988; Cronbach, 1950; Krosnick, 1992; Krosnick & Schuman, 1988; Mathews, 1927; McClendon, 1991; Rubin, 1940; Schuman & Presser, 1981; Schwarz, Hippler, & Noelle-Neumann, 1992; Visser, Krosnick, Marquette, & Curtin, 1999).

If the response order effects demonstrated in these studies are due to weak satisficing, then these effects should be stronger under conditions where satisficing is most likely. And indeed, these effects were stronger when participants had relatively limited cognitive skills (Krosnick, 1990; Krosnick & Alwin, 1987; Krosnick, Narayan, & Smith, 1996; McClendon, 1986; McClendon, 1991; Narayan & Krosnick, 1996). Mathews (1927) also found stronger primacy effects as questions became more and more difficult and as participants became more fatigued. And although McClendon (1986) found no relation between the number of words in a question and the magnitude of response order effects, Payne (1949/1950) found more response order effects in questions involving more words and words that were more difficult to comprehend. Also, Schwarz et al. (1992) showed that a strong recency effect was eliminated when prior questions on the same topic were asked, which presumably made participants' knowledge of the topic more accessible and thereby made optimizing easier for them.

Much of the logic articulated above regarding categorical questions seems applicable to rating scales, but in a different way than for categorical questions. Many people's attitudes are probably not perceived as precise points on an underlying evaluative dimension but rather are seen as ranges or "latitudes of acceptance" (Sherif & Hovland, 1961; Sherif, Sherif, & Nebergall, 1965). If a satisficing participant considers the options on a rating scale sequentially, then he or she may select the first one that falls in his or her latitude of acceptance, yielding a primacy effect under both visual and oral presentation.

Nearly all of the studies of response order effects in rating scales involved visual presentation, and when order effects appeared, they were nearly uniformly primacy effects (Carp, 1974; Chan, 1991; Holmes, 1974; Johnson, 1981; Payne, 1971; Quinn & Belson, 1969). Furthermore, the two oral presentation studies of rating scales found primacy effects as well (Kalton et al., 1978; Mingay & Greenwell, 1989). Consistent with the satisficing notion, Mingay and Greenwell (1989) found that

their primacy effect was stronger for people with more limited cognitive skills. However, these investigators found no relation of the magnitude of the primacy effect to the speed at which interviewers read questions to participants, despite the fact that a fast pace presumably increased task difficulty. Also, response order effects were found to be no stronger when questions were placed later in a questionnaire (Carp, 1974). Thus, the moderators of rating scale response order effects may be different from the moderators of such effects in categorical questions, though more research is clearly needed to full address this question.

How should researchers handle these response choice order effects when designing attitude measures? One possibility would be to ignore them, in the hope that they are relatively rare and, when they do occur, rarely displace variables' distributions by large degrees. Unfortunately, this approach seems overly optimistic. Even if a researcher is interested primarily in associations between variables (rather than univariate distributions), tests of the form-resistant correlation hypothesis suggest that the conclusions of correlational analysis can be significantly altered by response order effects (see Krosnick & Fabrigar, forthcoming). It therefore seems wiser to take some steps to address these effects in the design phase of a research project.

One seemingly effective way to do so is to counterbalance the order in which response choices are presented to participants. Counterbalancing is relatively simple to accomplish with dichotomous questions; half of a set of participants can be given one order, and the other half can be given the reverse order. When the number of response choices increases, the counterbalancing task can become more complex. However, it would make no sense to completely randomize the order in which rating scale points are presented, because that would eliminate the sensible progressive ordering of them from positive to negative, negative to positive, most to least, least to most, or whatever. Therefore, for rating scales, only two orders would presumably be used, regardless of how many points are on the scale.

Unfortunately, counterbalancing order across participants creates a new problem: variance in responses due to systematic measurement error. Once response alternative orders have been varied across participants, their answers will probably differ from one another partly because different people received different orders. One might view this new variance as *random* error variance, the effect of which would be to attenuate observed relations among variables and leave marginal distributions of variables unaltered. However, given the theoretical explanations for response order effects proposed above, this error seems unlikely to be random.

We therefore suggest considering an alternative approach to solving this problem. In addition to counterbalancing presentation order, it seems potentially valuable to take steps to prevent the effects from ever occurring in the first place. The most effective method for doing so presumably depends on the cognitive mechanism producing the effect. If primacy effects in rating scale questions are due to satisficing, then steps that reduce satisficing should reduce the effects. For example, with regard to motivation, questionnaires can be kept short, and accountability can be induced by occasionally asking participants to justify their answers. And with regard to task difficulty, the wording of questions and answer choices can be made as simple as possible.

No-Opinion Filters and Attitude Strength

When we ask participants to report their attitudes, we presume that their answers reflect information or opinions that they previously had stored in memory. And if a person does not have a pre-existing opinion about the object of interest, the question itself presumably prompts him or her to draw on relevant beliefs or attitudes in order to concoct a reasonable, albeit new, belief or evaluation (see, e.g., Zaller & Feldman, 1992). Consequently, whether based upon a pre-existing judgment or a newly formulated one, responses presumably reflect the individual's belief about or orientation toward the object.

What happens when people are asked about an object regarding which they have no knowledge and no opinion? Ideally they will say that they have no opinion or aren't familiar with the object or don't know how they feel about it (we refer to all such responses as no-opinion or NO responses). But when participants are asked a question in such a way as to suggest that they ought to have opinions of the object, they may wish not to appear foolishly uninformed and may therefore give arbitrary answers (Converse, 1964). In order to reduce the likelihood of such behavior, some questionnaire design experts have recommended that no-opinion options routinely be included in questions (e.g., Bogart, 1972; Converse & Presser, 1986; Payne, 1949/1950; Vaillancourt, 1973). In essence, such options tell participants that it is acceptable to say they have no attitude toward an object.

Do no-opinion filters work? Do they successfully encourage people without meaningful opinions to admit it? That is, is the overall quality of data obtained by a filtered question better than the overall quality of data obtained by an unfiltered question? Might filters go too far and discourage people who have meaningful opinions from expressing them? These important issues can be explored by drawing upon a large body of existing research, and this work suggests clearly that no-opinion filters are a bad idea.

Support for this conclusion comes from a series of studies that explored whether the substantive responses provided by people who would have said "don't know" if that had been offered to them are in fact meaningless. In one nonexperimental study, Gilljam and Granberg (1993) asked participants three questions tapping attitudes toward building nuclear power plants. The first of these questions offered a NO option, and 15% of participants selected it. The other two questions, asked later in the interview, did not offer NO options, and only 3% and 4% of participants, respectively, failed to offer substantive responses to them. Thus, the majority of participants who initially said NO offered opinions on the later two questions. However, these later responses mostly reflected meaningful opinions, because the two attitude reports correlated moderately with one another and predicted participants' later voting behavior.

Other studies examined the predictive validity and reliability of attitude reports and reached similar conclusions. Bishop, Oldendick, Tuchfarber, and Bennett (1979) found slightly stronger associations of attitudes with other criterion items when NO options were offered than when they were not, but Schuman and Presser (1981) rarely found such differences. And Alwin and Krosnick (1991), McClendon and Alwin (1993), Krosnick and Berent (1990), Krosnick et al. (2002), and Poe et al. (1988) found no greater reliability of self-reports when NO filters were included in questions than when they were not.

Krosnick et al. (2002) found that offering NO options did not enhance the degree to which people's answers were responsive to question manipulations that should have affected them. Specifically, participants in their study were told about a program that would prevent future oil spills and were asked whether they would be willing to pay a specified amount for it in additional taxes. Different participants were told different prices, on the presumption that fewer people would be willing to pay for the program as the price escalated. In fact, this is what happened. If pressing NO responses into substantive ones creates meaningless answers, then sensitivity to the price of the program would be less among people pressed to offer substantive opinions than among people offered a NO option. But in fact, sensitivity to price was the same in both groups. Finally, Visser, Krosnick, Marquette, and Curtin (2000) found that pre-election polls predict election outcomes more accurately when participants who initially say they don't know are pressed to identify the candidate toward whom they lean.

Taken together, the literature on how filters affect data quality suggests that NO filters do not remove only people without meaningful opinions. Thus, we see here reason to hesitate regarding the use of such filters. In order to make sense of this surprising evidence, it is useful to turn to studies by cognitive psychologists of the process by which people decide that they do not know something.

Norman (1973) proposed a two-step model that seems to account for observed data quite well. If asked a question such as "Do you favor or opposed U.S. government aid to Nicaragua?", a participant's first step would be to search long-term memory for any information relevant to the objects mentioned: U.S. foreign aid and Nicaragua. If no information about either is recalled, the individual can quickly respond by saying he or she has no opinion. But if some information is located about either object, the person must then retrieve that information and decide whether it can be used to formulate a reasonable opinion. If not, he or she presumably replies "don't know," but the required search time make this a relative slow response. Glucksberg and McCloskey (1981) reported a series of studies demonstrating that "don't know" responses can indeed occur either quickly or slowly, the difference resulting from whether or not any relevant information can be retrieved in memory.

This distinction between first-stage and second-stage NO responses suggests different reasons for them. According to the proponents of NO filters, the reason presumed to be most common is that the participant lacks the necessary information and/or experience with which to form an attitude. Such circumstances would presumably yield quick, first-stage NO responses. In contrast, second-stage NO responses could occur, for example, because of ambivalence. That is, some participants may know a great deal about an object and/or have strong feelings toward it, but their thoughts and/or feelings may be highly contradictory, making it difficult to select a single response.

It also seems possible that NO responses can result at what might be considered a third stage, the point at which participants attempt to translate their retrieved judgments onto the response choices offered by a question. For example, a participant may know approximately where he or she falls on an attitude scale (e.g., around 6 or 7 on a 1–7 scale), but because of ambiguity in the meaning of the scale points or of his or her internal attitudinal cues, he or she may be unsure of exactly which point to choose, yielding a NO response. A participant who has some information about an object, has a neutral overall orientation toward it, and is asked a question without a neutral response option might say NO because the answer he or she would like to give has not been conferred legitimacy. Or a participant may be concerned that he or she does not know enough about the object to defend an opinion toward it, so that opinion may be withheld rather than reported.

And finally, it seems possible that some NO responses occur at a pre-first stage, before participants have even begun to attempt to retrieve relevant information. For example, if a participant does not understand the question being asked and is unwilling to answer until its meaning is clarified, he or she might respond "I don't know" (see, e.g., Fonda, 1951).

There is in fact evidence that some NO responses occur for all of these reasons, but when people are asked directly why they give NO responses, people rarely attribute such responses to complete lack of information, are rarely due to lacking an opinion, and most often occur for the other reasons outlined above (Coombs & Coombs, 1976; Faulkenberry & Mason, 1978; Klopfer & Madden, 1980; Schaeffer & Bradburn, 1989).

Another explanation for the fact that NO filters do not consistently improve data quality is satisficing (Krosnick, 1991). According to this perspective, people have many latent attitudes that they are not immediately aware of holding. Because the bases of those opinions reside in memory, people can retrieve those bases and integrate them to yield an overall attitude, but doing so requires significant cognitive effort (a response behavior referred to as "optimizing"). When people are disposed not to do this work and instead prefer to shortcut the effort they devote in generating answers, they will attempt to satisfice by looking for cues in a question that point to an answer that will appear to be acceptable and sensible but that requires little effort to select. A NO option constitutes just such a cue and may therefore encourage satisficing, whereas omission of the NO option would instead inspire participants to do the cognitive work necessary to retrieve relevant information from memory.

This perspective suggests that NO options should be especially likely to attract participants under the conditions thought to foster satisficing: low ability to optimize, low motivation to do so, or high

task difficulty. And consistent with this reasoning, NO filters attract participants with more limited cognitive skills, as well as participants with relatively little knowledge and exposure to information about the attitude object (for a review, see Krosnick, 1999). In addition, NO responses are especially common among people for whom an object is low in personal importance, of little interest, and arouses little affective involvement, and this may be because of lowered motivation to optimize under these conditions. Furthermore, people are especially likely to say NO when they feel they lack the ability to formulate informed opinions (i.e., subjective competence), and when they feel there is little value in formulating such opinions (i.e., demand for opinionation). These associations may arise at the time of attitude measurement: Low motivation inhibits a person from drawing on knowledge available in memory to formulate and carefully report a substantive opinion of an object.

NO responses are also more likely when questions appear later in a questionnaire, at which point participant motivation to optimize is presumably waning (Culpepper, Smith, & Krosnick, 1992; Krosnick et al., 2002; Dickinson & Kirzner, 1985; Ferber, 1966; Ying, 1989). Also, NO responses become increasingly common as questions become more difficult to understand (Converse, 1976; Klare, 1950). Additionally Houston and Nevin (1977) found experimentally that describing a research study as being conducted by a prestigious sponsor for a purpose consistent with its identity (a university seeking to advance knowledge) decreased NO responses, presumably via enhanced participant motivation to optimize.

Hippler and Schwarz (1989) proposed another reason why NO filters discourage reporting of real attitudes: Strongly worded NO filters might suggest to participants that a great deal of knowledge is required to answer an attitude question and thereby intimidate people who feel they might not be able to adequately justify their opinions. Consistent with this reasoning, Hippler and Schwarz found that participants inferred from the presence and strength of a NO filter that follow-up questioning would be more extensive, would require more knowledge, and would be more difficult. If participants were motivated to avoid extensive questioning or were concerned that they couldn't defend whatever opinions they might offer, then they might be biased towards a NO response.

Another reason why people might prefer to select NO options rather than offering meaningful opinions is the desire not to present a socially undesirable or unflattering image of themselves. Consistent with this claim, many studies found that people who offered NO responses frequently would have provided socially undesirable responses (Cronbach, 1950, p. 15; Fonda, 1951; Johanson, Gips, & Rich, 1993; Kahn & Hadley, 1949; Rosenberg, Izard, & Hollander, 1955).

Taken together, these studies suggest that NO responses often result not from genuine lack of attitudes but rather from ambivalence, question ambiguity, satisficing, intimidation, and self-protection. In each of these cases, there is something meaningful to be learned from pressing participants to report their opinions, but NO response options discourage people from doing so. As a result, data quality does not improve when such options are explicitly included in questions.

A better way to accomplish the goal of differentiating "real" opinions from "non-attitudes" is to measure the strength of an attitude using one or more follow-up questions. Krosnick and Petty (1995) proposed that strong attitudes can be defined as those that are resistant to change, are stable over time, and have powerful impact on cognition and action. Many empirical investigations have confirmed that attitudes vary in strength, and the participants' presumed task when confronting a "don't know" response option is to decide whether his or her attitude is sufficiently weak to be best described by selecting that option. But because the appropriate cut point along the strength dimension seems exceedingly hard to specify and unlikely to be specified uniformly by participants, it seems preferable to ask people to describe where their attitudes fall along the strength continuum.

Many different attitude attributes are correlated with attitude strength, and these attributes are all somewhat independent of each other (see, e.g., Krosnick, Boninger, Chuang, Berent, & Carnot, 1993). For example, people can be asked how important the object is to them personally or how much they have thought about it or how certain they are of their opinion or how knowledgeable

they feel about it (for details on measuring these and many other dimensions, see Wegener et al., 1995). Measuring each of these dimensions can help to differentiate attitudes that are crystallized and consequential from those that are not.

Summary

All of the above studies and many others suggest optimal and less optimal ways to produce reliable and valid measurements of attitudes via direct self-reports (see Krosnick & Fabrigar, forthcoming). Each of the sources of error outlined above (e.g., the number of points on a rating scale, the verbal labeling, and order of response choices) may have a relatively small effect, but when a set of compromises are conglomerated, the net measurement error induced may be quite considerable. If researchers wish to make accurate assessments of people's attitudes and to have the greatest chance of finding statistically significant correlations between variables and statistically significant effects of manipulations on attitudes, then following the guidelines outlined above to minimize measurement error seems well-advised.

Alternatives to Direct Self-Reports

Given that direct self-reports will only be valid if participants are willing to describe themselves accurately, it is understandable that researchers have wondered whether motivational forces might sometimes lead participants to abandon this goal and to misrepresent themselves, creating a different sort of measurement error. A great deal of research has addressed this issue, and we turn to that work next.

The Notion of Social Desirability Response Bias

The idea that research participants might lie to researchers is not an implausible proposition, to be sure. For example, DePaulo, Kashy, Kirkendol, Wyer, and Epstein (1996) had people complete daily diaries in which they recorded any lies that they told during a 7-day period. On average, people reported telling one lie per day, with some people telling many more, and 91% of the lies involved misrepresenting oneself in some way. This evidence is in line with theoretical accounts from sociology (Goffman, 1959) and psychology (Schlenker & Weingold, 1989) asserting that an inherent element of social interaction is constructing an image of oneself in the eyes of others in pursuit of relevant goals. The fact that being viewed more favorably by others is more likely to bring rewards and minimize punishments may motivate people to construct favorable self-images, sometimes via deceit. If such behavior is common in daily life, why wouldn't people lie when answering questionnaires as well?

There are in fact a number of reasons to believe that the motivation to lie when answering questionnaires might be minimal. First, when filling out an anonymous questionnaire, no rewards or punishments can possibly be at stake. And second, in most surveys and laboratory experiments, the participants' relationships with a researcher are so short-lived and superficial that very little of consequence is at stake as well. Certainly, a small frown of disapproval from a total stranger can cause a bit of discomfort, but little more than that. And the cognitive task of figuring out which response to each question one is asked will garner the most respect from a researcher is likely to be demanding enough to be worth doing only when the stakes are significant. So perhaps there isn't so much danger here after all.

Unfortunately, however, there is another potential source of systematic distortion in responses to even self-administered anonymous questionnaires: self-deception. Not only do people want to maintain favorable images of themselves in the eyes of others, but they also want to have such images

in their own eyes as well. According to many psychological analyses, the pursuit of self-esteem is a basic human motive (see, e.g., Sedikides & Strube, 1997), and it is driven partly by such inevitable realities as the prospect of death (e.g., Greenberg, Solomon, & Pyszczynski, 1997). So people may be motivated to convince themselves that they are respectable, good people, and doing so may at times entail misconstrual of facts (see Paulhus, 1984, 1986, 1991). If people fool themselves in this way, then of course such misconstrual will find its way into questionnaire responses, even when participants want to accurately report their attitudes to an interviewer and/or researcher. Obviously, it is tricky business to fool oneself, because part of the mind would need to know that it's fooling another part. But such self-deception can be so automatic that people may not be aware of it at all.

Documenting the Extent of Self-Presentational Social Desirability Response Bias

The evidence documenting systematic and intentional misrepresentation in questionnaire responses is now quite voluminous and very convincing, partly because the same conclusion has been supported by studies using many different methods. One such method is the "bogus pipeline technique," which involves telling participants that the researcher can otherwise determine the correct answer to a question they will be asked, so they might as well answer it accurately (see, e.g., Roese & Jamieson, 1993). Under these conditions, people are more willing to report substance use than they would be if asked directly (Evans, Hansen, & Mittlemark, 1977; Murray & Perry, 1987). Likewise, White participants are more willing to ascribe undesirable personality characteristics to African Americans (Sigall & Page, 1971; Pavlos, 1972, 1973) and are more willing to report disliking African Americans (e.g., Allen, 1975) under bogus pipeline conditions. Women are less likely to report supporting the women's movement under bogus pipeline conditions than under normal reporting conditions (Hough & Allen, 1975). And people are more likely to admit having been given secret information under bogus pipeline conditions (Quigley-Fernandez & Tedeschi, 1978).

Another approach to documenting such distortion is to compare responses given when people believe their answers will have significant consequences for them to responses given when no such consequences exist. For example, in one study, participants who believed that they had already been admitted to an apprenticeship program admitted to having less respectable personality characteristics than did comparable participants who believed they were being evaluated for possible admission to the program (Michaelis & Eysenck, 1971).

Yet another approach to this problem involves the "randomized response technique" (Warner, 1965). Here, participants answer one of various different questions, depending upon what a randomizing device instructs. The researcher does not know exactly which question each person is answering, so participants can presumably feel freer to be honest. In one such study, Himmelfarb and Lickteig (1982) had participants secretly toss three coins before answering a yes/no question. Participants were instructed to say "yes" if all three coins came up heads, "no" if all three coins came up tails, and to answer the yes/no question truthfully if any combination of heads and tails came up. People answering in this fashion admitted to falsifying their income tax reports and enjoying soft-core pornography more than did participants who were asked these questions directly.

Still another approach to assessing the impact of social desirability is by studying interviewer effects. The presumption here is that the observable characteristics of an interviewer may suggest to a participant which answers are considered most respectable. So if answers vary in a way that corresponds with interviewer characteristics, it suggests that participants tailored their answers accordingly. For example, various studies have found that African Americans report more favorable attitudes toward Whites when their interviewer is White than when the interviewer is African American (Anderson, Silver, & Abramson, 1988; Campbell, 1981; Schuman & Converse, 1971). Likewise, White participants express more favorable attitudes toward African Americans to African

American interviewers than to White interviewers (Campbell, 1981; Cotter, Cohen, & Coulter, 1982; Finkel, Guterbock, & Borg, 1991). These effects have occurred both in face-to-face interviews and in telephone interviews as well (Cotter et al., 1982; Finkel et al., 1991). And in another study, people expressed more positive attitudes toward firefighters when they thought their interviewer was a firefighter than when they did not hold this belief (Atkin & Chaffee, 1972/1973).

Another approach to this issue involves comparisons of different modes of data collection. In general, pressure to appear socially desirable is presumably greatest when a participant is being interviewed by another person, either face-to-face or over the telephone. This pressure is presumably lessened when participants are completing written questionnaires. Consistent with this reasoning, Catholics in one study were more likely to report favoring legalized abortion and birth control when completing a self-administered questionnaire than when being interviewed by telephone or face-to-face (Wiseman, 1972). And people report being happier with their lives in interviews than on self-administered questionnaires (Cheng, 1988).

Anonymity of self-administered questionnaires further reduces social pressure, so it, too, offers an empirical handle for addressing this issue. In one study, Gordon (1987) asked participants about dental hygiene on questionnaires; half the participants (selected randomly) were asked to write their names on the questionnaires, whereas the other half were not. Dental checkups, brushing, and flossing were all reported to have been done more often when people wrote their names on the questionnaires than when they did not. Thus, socially desirable responses were apparently more common under conditions of high identifiability. Similarly, people reported having more desirable personality characteristics when they wrote their names, addresses and telephone numbers on questionnaires than when they did not (Paulhus, 1984).

Taken together, these studies all suggest that some people sometimes distort their answers to questionnaire items in order to present themselves as having more socially desirable or respectable characteristics or behavioral histories. These studies also validate a series of methods that can be used to detect social desirability bias in responses. That is, if a researcher is worried that answers to a particular question might be distorted by intentional misrepresentation, an experiment can be conducted employing a technique such as randomized response to see whether different results are obtained.

It is important to note that only relatively small distortions in results have been documented in all of the social desirability studies reviewed above. But the social desirability-driven distortions documented above represent only those involving other-deception. Therefore, there may be significant amounts of self-deception going on as well, and when combined with other-deception, social desirability-driven error may be substantial.

Indirect Measurement Techniques

To overcome the problems with intentional and unintentional distortion of direct attitude reports, much research has explored the use of measurement techniques that keep self-presentational concerns from entering a person's deliberation of his or her evaluation in the first place. Such techniques have a long history in attitude research, but modern technology has opened up new means for indirect attitude assessments. We discuss three kinds of indirect measures in this section: unobtrusive behavioral observation, contemporary implicit measures of automatic evaluation, and physiological measures.

Unobtrusive Behavioral Observation

Originally, measures designed to limit self-presentational concerns relied primarily on unobtrusive assessments of overt behaviors. One such strategy has been to simply capture the physical traces

of behavior like, for example, the empty liquor bottles in public garbage collections as an indicator of attitudes toward alcohol consumption (see Webb, Campbell, Schwartz, & Sechrest, 1966). Another option is to disguise what is being measured and/or conceal the measurement itself. Milgram's classic "lost-letter technique" uses this strategy and involves the placement of ostensibly lost letters in public places (Milgram, Mann, & Harter, 1965). The address on the envelopes is manipulated (and in some cases the sender information: Benson, Karabenick, & Lerner, 1976). Based on the assumption that individuals with more positive attitudes toward the addressee will be more likely to pick up the envelope and put it in a mailbox, the rate and speed of return for these letters is recorded as an indicator of attitudes toward the addressee (e.g., "Friends of the Nazi party" in Milgram et al., 1965).

Other examples of unobtrusive observation techniques focus on responses that are more closely linked to the assessed attitude but are rather incidental behaviors that people are unlikely to suspect are monitored by researchers. For instance, in the classic seating task developed by Weitz (1972), participants are asked to take a seat in a waiting room where an outgroup target person is already waiting. The critical measure is how closely the participant sits to the target when given a choice of seats that vary in physical proximity. Presumably, the more negative a person's attitude toward the outgroup, the farther away he or she will choose to sit from the target.

Yet another strategy for unobtrusive observation is to disguise what attitude is actually being studied. For example, studies on intergroup attitudes have considered helping behavior in interpersonal contexts as a measure of racial attitudes. These studies have assessed how a person responds when given the opportunity to aid another individual who is either an ingroup or outgroup member (e.g., Gaertner & Dovidio, 1977; see Penner, Dovidio, Manning, Albrecht, & van Ryn, Volume 2). Likewise, studies by Donnerstein and colleagues used the same approach for assessing the flip side of pro-social behavior. They provided participants with a legitimate opportunity to aggress toward another individual in the context of a learning experiment, varying the individual's group membership (e.g., Donnerstein & Donnerstein, 1975). Although the participants in these helping and aggression studies were in all likelihood cognizant of the fact that their behavior was being recorded, they may nevertheless have been unaware that their attitudes toward a particular social group were the focus of the measurement effort.

Of course, the expressed goal of these kinds of measurement techniques is to reduce the impact of normative concerns on a person's responses and thereby eliminate strategic misrepresentation. The effectiveness of such a measurement strategy is often assumed to be based on the fact that normative concerns will not come to mind during the assessment and are not used for the targeted response. Therefore, the assessment context is designed to curtail the presence of cues that could trigger deliberation about the social acceptability of one's attitude, so responses are ostensibly unmonitored. However, there may be another reason why unobtrusive measures can be effective in limiting self-presentational bias. They may simply assess responses under conditions in which people fail to recognize the impact of their attitudes and thus ignore not only normative implications but all aspects of those attitudes. This possibility is most apparent in the case of techniques designed to disguise the purpose of the assessment. Such strategies may not simply render the normative implications of an attitude less salient for people, but they may also make it more difficult for people to recognize the attitude in question as a potential determinant of their behavior. Thus, when deliberating whether or not to assist another person in need of help, or when choosing a chair in the waiting room, participants may remain unaware of the implicit influences that the target's race has on their decision. Even unobtrusive observation techniques that draw attention to the critical attitude, like the lost-letter technique, may have a similar effect on evaluative processing, as they assess behaviors under circumstances in which the motivation to deliberate is likely to be rather limited. In the absence of much controlled deliberation of one's attitude, its impact on responses may easily go unnoticed. In short, aside from controlling the salience and relevance of normative

considerations during assessment, self-presentational bias in attitude measurement can be limited by assessing implicit evaluative influences on behavior.

Measures of nonverbal communication make up a final set of traditional unobtrusive observation techniques intended to capture implicit evaluations even in circumstances in which people are motivated to monitor the appropriateness of their behavior. In the past, various nonverbal behaviors, including body posture, eye contact, and fidgeting have been used to assess intergroup attitudes (e.g., McConnell & Leibold, 2001; Word, Zanna, & Cooper, 1974). The general idea behind the use of such measures is that nonverbal channels of communication are more difficult to control than are most aspects of verbal behavior (Dovidio, Kawakami, & Gaertner, 2002). Nonverbal channels therefore allow researchers to assess implicit evaluative influences on interpersonal behavior even when people are deliberately trying to control such influences. For example, in an interracial interaction, people may be more successful at keeping negative racial attitudes from influencing their verbal statements than suppressing their impact on nonverbal expressions. Thus, measures of nonverbal behavior would reveal evaluative biases that could be hidden in other, more deliberate, channels of communication.

Of course, none of these measures offer precise control over the exact nature of the evaluative processing that takes place during the assessment. Nor do the measures necessarily guarantee that the attitude in question will be a particularly prominent influence on the assessed response. After all, behavior is generally influenced by a multitude of factors, a person's attitude being just one among many. As a result, measures based on behavioral observation may be particularly noisy. These are just some of the reasons why unobtrusive behavioral observation measures are not especially popular today.

Modern implicit assessment techniques are intended to overcome these problems. Instead of capturing complex behaviors, they aim to assess the activation of an evaluation independent of processes that take place during the deliberation and response phases of evaluative processing (Wittenbrink, 2007; see Gawronski & Brannon, this volume). In other words, implicit measures are meant to capture the automatic operation of an attitude (De Houwer & Moors, 2007). We discuss them in the following sections.

Contemporary Implicit Measures

Among the new kinds of implicit measures that have received the most attention are those based on response latencies and response errors when a certain judgment has to made under time pressure. Based on the mechanisms by which they are thought to operate, implicit measures may be categorized into (a) response interference measures and (b) priming measures.

RESPONSE INTERFERENCE MEASURES

Measures that rely on response interference set up a task where an attitude stimulus may be responded to in multiple ways. The procedure then captures the speed with which individual responses can be executed. A classic example of such a setup is the Stroop Color-Word Test. Here, participants are asked to quickly identify the color of target words. For the critical items, these targets themselves represent a color, sometimes compatible (e.g., the word *green* appearing in green color), other times incompatible (e.g., *green* in red color). In general, responses take longer and are less accurate when the meaning of the word conflicts with the font color. The effect is thought to reflect the influence of two independent response tendencies elicited by the semantic meaning of the stimulus and by its font color. On incompatible trials, these response tendencies are in conflict: the word meaning invoking one response, the font color a different response, thereby interfering with a quick and accurate response to that stimulus. On compatible trials both meaning and font color yield the same

response tendency, allowing for accurate and efficient responding (see MacLeod, 1991). A number of implicit attitude measures operate in this fashion, including the Implicit Association Test.[2]

Implicit Association Test (IAT) Developed by Greenwald, Banaji, and their colleagues (Greenwald et al., 1998), the IAT is easily the most widely used implicit measure of attitudes. It has been applied to virtually all areas of attitude research. Via the online portal http://projectimplicit.net, the IAT has been disseminated to the general public where visitors are offered feedback on their attitudes and invited to participate in online studies. Given its popularity, it has received extensive, at times controversial, coverage in the news media (cf., Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; Singal, 2017). In this task, participants classify as quickly as possible two sets of target items along two dimensions of judgment. For example, one set of items may consist of targets of polarized valence (e.g., poison, love), for which participants perform an evaluative discrimination task using two response keys (e.g., pleasant/unpleasant). A second set of target items may include exemplars of two contrasting categories of attitude objects (e.g., flowers: *tulip*, *rose* versus insects: *spider*, *ant*). The task for this second set of items is to quickly classify them according to their category membership.

During a set of trials, both judgment tasks are combined, and the targets of the two sets of valence and attitude items appear in random order. Both judgment tasks are performed using the same two response keys. Two separate assessment blocks then vary the mapping of categories on the response keys, so that each attitude object is paired once with the positive response key and once with the negative key (e.g., *flower/pleasant* and *insect/unpleasant* versus *flower/unpleasant* and *insect/pleasant*). The critical measure, the IAT Effect, assesses which of these two blocks produces more fluent, faster responses as an indicator of relative evaluative preference. For example, relatively faster responses when *flower* is paired with *pleasant* and *insect* is paired with *unpleasant* indicate an evaluative preference for flowers over insects (for a detailed review of experimental procedure and data analysis, see Greenwald, Nosek, & Banaji, 2003).

Analogous to the Stroop Color-Word Test, operation of the IAT has been explained in terms of response interference (De Houwer, 2003a; Gawronski, Deutsch, LeBel, & Peters, 2008). That is, during the combined IAT trials, attitude targets may elicit both a response tendency that is driven by the targets' evaluative associations (pleasant vs. unpleasant), and another response tendency that is triggered by the targets' category membership (flower vs. insect). In the evaluatively congruent block where category and associated evaluation are assigned to the same response key, both tendencies yield correct responses (e.g., *rose*: pleasant, *rose*: flower). However, in the evaluatively incongruent block, only category-based response tendencies result in correct responses. In contrast, evaluation-based response tendencies elicit the opposite response, thereby interfering with a fluent and accurate execution. In other words, similar to the Stroop Color-Word Test, the main mechanism driving IAT effects is response interference.[3]

IAT Variants Several modifications of the original IAT procedure have been proposed. Although none of these modified procedures have seen widespread adoption, they do target important procedural or interpretative limitations of the IAT. For example, the IAT assesses attitudes not in absolute terms but only in relation to a second contrasting category.[4] In many cases, the contrasting category is not an obviously mutually exclusive category and instead is selected from among many plausible alternatives (e.g., spinach vs. broccoli, corn, peas, beans, asparagus, salmon, hamburger, french fries). The choice may influence what features of the target category become salient (Tversky, 1977). For example, an IAT is likely to yield different results for the attitude toward *spinach* when it is paired with *carrot* than when it appears in contrast to *french fries*. Moreover, even for naturally dichotomous categories (e.g., male, female) or for objects that imply an obvious contrast category (e.g., Republicans vs. Democrats), the relativity of the attitude estimate obtained by the IAT renders its interpretation potentially ambiguous. For example, a score close to zero on a race IAT with *Black people*

and *White people* as contrasting categories may result either from very positive evaluations of both target groups or equally negative attitudes for the two groups. Likewise, a relatively positive IAT score may reflect very positive evaluations associated with White people as much as it may indicate very negative attitudes toward Black people. Obviously, the respective alternative interpretations paint very different portraits of the underlying group attitudes, but an IAT score cannot differentiate among them. This complicates a proper interpretation of IAT results (see Blanton & Jaccard, 2006) but also makes it difficult to use the IAT in situations where separate attitude measures for each of the two target attitudes are required. For example, in research on intergroup attitudes, it is often of interest to differentiate ingroup favoritism (positive evaluations of an ingroup) from outgroup derogation (negative attitudes toward an outgroup; see Brewer, 2001).

Several measures have been developed to address this particular limitation of the IAT. The Single Target IAT (ST-IAT, Karpinski & Steinman, 2006; Wigboldus, Holland, & van Knippenberg, 2004) consists essentially of an IAT with a single attitude category. That category is paired with *pleasant* on the same response key during one block of trials and with *unpleasant* on the next block. In each case, the opposite response key carries only an evaluative label. Thus, the ST-IAT may be used to assess evaluative preferences for a single attitude object, without the presence of a contrasting category. Moreover, a study by Bluemke and Friese (2008) suggests that it may be feasible to obtain multiple independent attitude estimates for several attitude objects in a series of ST-IATs, without the individual estimates being influenced by the serial positioning of the attitude object. The absence of a contrasting category simplifies the task, compared to a standard IAT procedure. As a result, ST-IAT responses may be more susceptible to strategic interference. For example, Stieger, Göritz, Hergovich, and Voracek (2011) show that instructions to participants on how to fake responses related to a socially sensitive attitude proved more effective on an ST-IAT than on a standard IAT.

Another modification of the IAT meant to provide absolute estimates of a single attitude is the Extrinsic Affective Simon Task—EAST. Proposed by De Houwer (2003b), the EAST works by adding color to an IAT with lexical stimuli. As in the IAT, participants classify two separate sets of stimuli, one related to an attribute dimension (e.g., good/bad) and the other made up of object exemplars (e.g., tulip). The object exemplars are presented in one of two font colors, and participants are instructed to press the *good* key whenever a word appears in, say, green, and to press the *bad* key for words in blue. Attribute stimuli are presented in white and have to be classified based on their valence. Because the font color of object stimuli can be varied across trials, each object stimulus can be paired once with the *good* and once with the *bad* response key. Faster responses on trials when the object target is paired with the *good* key indicate a more positive attitude toward the target.

The EAST is based on the assumption that the meaning of words representing the attitude objects is processed despite the fact that only their font color is task relevant. However, this may not always be the case as lexical processing can be impacted by task demands (Risko, Stolz, & Besner, 2005). In fact, the EAST has been found to show less internal reliability and criterion validity than the IAT (De Houwer & De Brucyker, 2007a). To address this issue, De Houwer and De Bruycker (2007b) have proposed a variant, the Identification-EAST or ID-EAST. Here, the word stimuli are presented on some trials in capital letters, on other trials in lowercase letters. Attribute words (*GOOD*, *good*) have to be classified according to the dimension in question (e.g., evaluation); attitude words (*TULIP*, *tulip*) are classified by their font case. Thus, attribute and attitude word sets are no longer differentiated by the task-relevant feature. Instead, lexical processing is require to determine how to respond to a target. Participants first have to identify the meaning of the word.

A third variant of the IAT that can be administered with only a single target attitude is the Go/No-go Association Task (GNAT—Nosek & Banaji, 2001). As in the IAT, presentation of exemplars of this target attitude alternates in random order with stimuli that vary on a particular dimension (pleasant/unpleasant). Unlike the IAT, however, participants have to give a response only

when a stimulus fits one of two categories. That is, participants may be shown names of flowers, positive words, and negative words (in some versions of the task, unrelated distractors as well). On some trials, participants press a key whenever the name of a flower or a positive word appears. On other trials, participants respond to flowers and negative words. Relatively faster responses to the first set of trials indicate a positive attitude toward flowers.

A standard implementation of the IAT typically requires at least 180 trials. The Brief IAT (BIAT, Sriram & Greenwald, 2009) considerably shortens the procedure to a total of 80 trials. The procedure uses only combined trials in which all four types of attitude and valence items are presented (e.g., flowers and insects, pleasant and unpleasant items). However, participants are instructed to focus on just two of the four categories, presenting a full list of items from these categories prior to the response trials. Two blocks then vary which item categories are made focal. For example, in the case of a flower/insect BIAT, participants would first see a list of all flower items and all pleasant items, instructing them to respond to any of these items by pressing the right response key, and by pressing the left key for anything else that might appear on the screen. Next, a list of all insect and all pleasant items would make those categories focal for the second block of trials. To improve measurement reliability, the two blocks are typically repeated once. A recent study by Bar-Anan and Nosek (2014) suggests the BIAT's reliability to be on par with that of a standard IAT (but see Rothermund & Wentura, 2010).

Evaluative Priming Task (EPT) Given that our coverage of implicit measures distinguishes between response interference and priming measures, it may be surprising to find the most common priming procedure used for attitude measurement, the EPT (Fazio, Sanbonmatsu, Powell, & Kardes, 1986), listed among the response interference measures. The procedure was inspired by, and shares its basic task structure with, priming paradigms used in cognitive research on semantic access in lexical processing (e.g., Meyer & Schvaneveldt, 1971). However, recent evidence suggests that its operation has more in common with response interference procedures (e.g., De Houwer, Hermans, Rothermund, & Wentura, 2002; Gawronski et al., 2008; Klauer, Roßnagel, & Musch, 1997; Klinger, Burton, & Pitts, 2000).

In the EPT, participants are shown in rapid succession an attitude prime (e.g., spinach), followed by a target word (e.g., pleasant). Participants indicate as quickly as possible whether the meaning of the target implies either *good* or *bad* by pressing the respective response key. Of interest is whether, across several trials with different targets, the attitude prime facilitates responses to positively valenced targets and/or responses to negatively valenced targets. To limit priming effects to automatic influences and to preclude effects that could result from deliberate processing of the attitude prime, prime and target appear in rapid succession (usually less than 300 milliseconds). Nevertheless, in the EPT 7, the attitude primes are clearly visible for participants. The procedure therefore requires some kind of cover story that instructs participants to respond to the target items, while at the same time justifying the presentation of the primes. For example, the primes may be introduced as being part of a secondary memory task meant to make the actual target response task more difficult (for a review of the measure, see Fazio, 2001).

Despite its name and its overt similarities to semantic priming, the EPT more likely operates by setting up a Stroop-like potential response conflict. The judgment task performed in the EPT is to classify targets according to their valence—for example, requiring a decision between a *good* and a *bad* response. To the extent that the attitude primes themselves are valenced, they may prepare a response that is either consistent or inconsistent with the response tendency triggered by the target word. As a result, on trials where prime and target are of different valence, these response tendencies are in conflict, thereby interfering with a quick and accurate response execution. However, on trials where attitude prime and target word are of the same valence, both prepare the same response tendency, facilitating fast and accurate responding.

PRIMING MEASURES

Like the EPT, priming measures used for attitude measurement that operate through spreading activation rather than response interference are based on the classic paradigm first introduced by Meyer and Schvaneveldt (1971). In their original procedure, participants are shown letter strings (e.g., *BUTTER*, *MARB*) and are asked to decide whether or not the target string forms a word. In addition, the letter string is paired with a prime, another word that in the common implementation of this paradigm precedes the target. The classic finding, replicated in numerous experiments, is that participants are faster in making such lexical decisions when prime and target string are semantically associated, when for example the string *BUTTER* is preceded by the prime *BREAD* (for a review, see Neely, 1991). A common explanation for the effect holds that the prime automatically activates other semantically related concepts in long-term memory, which subsequently reduces the time that is required for the activation of related targets to reach recognition threshold (Neely, 1977; Posner & Snyder, 1975).

Lexical Decision Task (LDT) Wittenbrink, Judd, and Park (1997) adapted Meyer and Schvaneveldt's paradigm for attitude measurement by presenting attitude objects as primes, followed by word and nonword target items. Using a lexical decision task, participants have to decide as quickly as possible whether the target strings form a word. Additional features of the procedure are meant to limit deliberate processing of the attitude primes. For example, very short stimulus presentations, combined with stimulus masking, can be used to conceal the primes from conscious awareness (e.g., Wittenbrink et al., 1997).

The LDT is effectively a semantic priming procedure where attitude objects serve as primes. Thus, differences in response latencies on trials where a word target is preceded by an attitude prime, compared to a control condition, are used as a measure of association between attitude and target. Because the lexical classification task employed in the LDT does not explicitly focus on target valence—or, for that matter, any other particular attribute of the target's meaning—the LDT can accommodate targets that vary on multiple dimensions across trials. Thereby, the LDT does not just assess an overall evaluative response but instead can measure differentiated and specific associations with the attitude object (e.g., spinach—healthy, yucky, fresh, bitter, etc.). For example, the LDT by Wittenbrink et al. (1997) was set up to assess racial attitudes. In this procedure, African American and White group primes are paired with trait attributes contained in the cultural stereotype for either of the two groups (athletic, intelligent). In addition, half of the items for each stereotype are positive in valence, and half are negative. The facilitation observed for the various combinations of primes and types of target items then offers separate estimates for (a) the degree to which a group prime yields automatic stereotype activation, (b) the extent to which this automatic stereotype activation is evaluatively biased (i.e., whether primarily negative or positive traits are activated), and (c) the capacity for a group prime to trigger an overall evaluation (i.e., to facilitate any item of particular valence, independent of the stereotype).

Except for the decision task, both LDT and EPT are largely identical procedures.[5] Nevertheless, this seemingly minor change in task focus impacts the underlying operation in important ways. The target items of interest to the LDT measure are always words, while the nonword targets are only filler items used to implement the lexical classification task. Because the attitude primes are also always words, the LDT does not involve any potential response interference on the trials of interest. That is, the attitude primes may initiate a response tendency for a *word* response. Yet, this response tendency is irrelevant for the emergence of differences in response latency to positive and negative (or related and unrelated) targets as it would apply to all targets, independent of their valence or relatedness. Any such differences will emerge, however, as a result of semantic priming.

Affective Misattribution Procedure (AMP) The AMP was proposed by Payne, Cheng, Govorun, and Stewart (2005), based on a study by Murphy and Zajonc (1993) in which affect primes

influenced the evaluation of otherwise unfamiliar Chinese ideographs. In the AMP attitude primes are used instead and are followed by briefly displayed and masked neutral targets, commonly the Chinese ideographs from Murphy and Zajonc. Participants classify these targets as either *pleasant* or *unpleasant*. The evaluation of targets following a particular attitude prime serve as the critical measure for the procedure.

An interesting feature of the AMP is its use of overt attitude primes, absent any instructions to disguise their function. In fact, there is some evidence that priming effects in the AMP persist even when participants are explicitly warned not to be influenced by the primes (Payne et al., 2005). However, more recent data suggest the evaluative effects obtained in the AMP are partially due to participants responding directly to the primes instead of the targets (Bar-Anan & Nosek, 2012; Rohr, Degner, & Wentura, 2015). Concealed prime presentations with very short display times and masked stimuli may circumvent those deliberate influences (see Rohr et al., 2015).

As reflected in the name, Payne and colleagues assume the AMP to operate via the affect misattribution mechanism, proposed by Murphy and Zajonc (1993). Accordingly, the attitude primes generate an early affective response that participants cannot correctly account for and thus influences the target evaluations. Essentially, participants confuse their evaluation of the targets with their affective response to the primes.

AMP Variants Target judgments other than the standard evaluative classification have been used with the AMP. For example, Blaison et al. (2012) and Rohr et al. (2015) successfully used multiple emotion categories (happy, sad, anger, fear) for the target classifications. Likewise, the Stereotype Misperception Task (SMT) by Krieglmeyer and Sherman (2012) displays face stimuli as primes and targets and asks participants to classify the targets based on valenced trait attributes. The feasibility of such modifications increases the AMP's practical use. However, it also suggests that the priming effects observed in the AMP are not solely due to misattributed affective responses, but also reflect the effects of semantic classification of the attitude primes (see Gawronski & Ye, 2014; Rohr et al., 2015).

Physiological Measures

Physiological attitude measures seek to capture the physiological correlates of evaluative responses. Because people generally have no control over physiological responses, researchers early on considered the assessment of these kinds of responses to be a way of overcoming intentional misrepresentation in direct attitude self-reports. Physiological measures operate implicitly because, in most cases, people have no introspective access to their response and its connection with a specific evaluation.

Early attempts to use physiological responses for attitude measurement focused on noninvasive measures of autonomic responses such as galvanic skin conductance and pupillary responses. Rankin and Campbell (1955) were among the first to use galvanic skin response (GSR), a measure of the ability of skin to conduct electricity, in attitude research. In their experiment, White participants showed an elevated GSR during interactions with an African American experimenter compared to a condition with a White experimenter. Subsequent research, however, indicated that GSR is primarily sensitive to arousal and cannot differentiate whether this arousal is triggered by a positively evaluated stimulus or a negatively evaluated stimulus or by a novel stimulus (Cacioppo & Sandman, 1981).

The use of pupillary responses for attitude measurement has not fared much better. In principle, this measure, first proposed by Hess (1965), was thought to differentiate between positive evaluations, which are believed to yield a dilation of the pupil, and negative evaluations, which are supposed to trigger pupil constriction. However, like the GSR, pupillary responses are influenced by the novelty of a stimulus (Petty & Cacioppo, 1983). In addition, empirical evidence testing whether negatively evaluated stimuli trigger pupil constriction is mixed at best (see Himmelfarb, 1993).

A more effective measurement approach assesses subtle muscle activity in specific areas of the face, commonly over the brow (frowning) and the cheek (smiling). For example, Cacioppo, Petty, Losch, and Kim (1986) found that EMG activity in these areas showed distinct patterns following exposure to either positive or negative stimuli. Observing judges failed to detect any overt expressions of positive or negative emotions, thus documenting the subtlety of the responses (see also Fridlund, Schwartz, & Fowler, 1984).

Facial EMG measures are generally based on multiple recordings of activity over a short period of time, during which participants think about the stimulus. The measure is therefore not well suited for the assessment of automatic evaluative responses free of deliberation. In addition, this measure is open to misrepresentation. People can fake or intentionally distort their facial expressions (Cacioppo et al., 1986). However, extra precautions to disguise the purpose of the assessment—for example, the placement of additional dummy electrodes in places other than the face, may make facial EMG an effective measure of socially sensitive attitudes (McHugo & Lanzetta, 1983; Vanman, Paul, Ito, & Miller, 1997).

Another attitude measure based on facial EMG activity assesses the modulation of eyeblink reflexes during exposure to an object. For this procedure, a startle probe (e.g., a short blast of acoustic noise or a visual flash) is used to elicit a reflexive eyeblink while participants watch images of an object. Startle eyeblink reflexes are modulated as a function of affective valence of the target stimulus. Exposure to positively evaluated stimuli is associated with eyeblink inhibition, whereas negatively evaluated stimuli elicit amplification of the reflex (Lang, Bradley, & Cuthbert, 1990). Some evidence suggests that affective modulation of the eyeblink reflex occurs only for highly arousing stimuli, which would limit its use to the assessment of attitudes involving strong evaluations (Cuthbert, Bradley, & Lang, 1996). Moreover, affective modulation is observable only after considerable exposure to the target stimulus. Early startle eyeblink responses, within 800 ms of stimulus onset, remain insensitive to the valence of the target stimulus (Bradley, Cuthbert, & Lang, 1993). Thus, although this measure captures responses that remain outside of participants' voluntary control, the nature of the responses can be determined by both automatic reactions to the target and by controlled deliberation of it.

A final set of physiological attitude measures is based on the assessment of brain activity. Most recently, these measure have begun to employ newly emerging brain imaging techniques, like positron emission tomography (PET) and functional magnetic resonance imagery (fMRI). These brain imaging techniques determine neural activity based on changes in blood flow in the brain and can be used to identify the brain regions that operate in the processing of a given stimulus.

Initial steps have been taken to link evaluative processing to activity in specific areas of the brain. For, example, activity in the amygdala, a neural structure that is part of the limbic system and that is located in the anterior part of the temporal lobes, is linked to the processing of negatively evaluated stimuli (e.g., Adolphs, Tranel, & Damsio, 1998; LeDoux, 1996). Based on these findings, a recent study by Phelps et al. (2000) explored the role of amygdala activity in more complex social attitudes. Using fMRI, this study recorded amygdala activity for White participants while they were shown images of African American and White faces and found it to be correlated with two other implicit racial attitude measures, an IAT and a startle eyeblink measure. Similarly, Hart et al. (2000) found increased amygdala activity in response to outgroup faces for both African American and White participants. This effect was observed, however, only on later trials, which the authors interpreted as evidence that participants more quickly habituated to ingroup faces.

Another technique for the use of brain activity in attitude measurement is a procedure based on event-related brain potentials (ERP) proposed by Cacioppo and his colleagues (Cacioppo, Crites, Berntson, & Coles, 1993; Cacioppo, Crites, Gardner, & Berntson, 1994). For an ERP, neural electric activity is recorded via electrodes placed on the scalp, and changes in this activity following a critical event (e.g., the presentation of an attitude object) are recorded. The procedure is based on a

particular component of the ERP waveform, known as the P300: a relative increase in neural activity that occurs relatively late in the ERP, approximately 300 ms after event onset.

This component is sensitive to the meaning of an event for the overall task that is performed during an ERP. For example, when participants are asked to classify stimuli according to a certain dimension (high tones vs. low tones), oddball stimuli that are inconsistent with prior stimuli (e.g., a low tone that follows a series of high tones) evoke a larger P300 in a specific location of the scalp (e.g., Fabiani, Gratton, Karis, & Donchin, 1987). The Cacioppo et al. measure, termed Late Positive Potential (LPP), employs such an oddball paradigm with an evaluative classification task, whereby a target stimulus is embedded into a sequence of stimuli of known valence. Ideally, attitude assessments would be derived from this measure by comparing trials in which the target is embedded in a sequence of positive stimuli with trials where it is paired with negative stimuli. However, reliable ERP waveforms can only be obtained across several presentations of the same stimulus sequence. In order to limit the repetitiveness of the procedure, LLP measures typically use only one valence context (Crites, Cacioppo, Gardner, & Berntson, 1995). The LLP amplitude, averaged across several presentations of the target stimulus, can be used as an indicator of the degree of evaluative mismatch between target and context stimuli.

The LLP measure offers precise control over the timing of evaluative processing. It is also unaffected by attempts to deliberately falsify evaluations during the classification task (Crites et al., 1995). Thus, it appears to be an effective measure of automatic evaluative responses free of controlled deliberation.

Other Implicit Measures

A variety of other implicit assessment techniques do not fit squarely into the above categories. For example, the latency and intensity of approach and avoidance motor movements have been used as indicators of evaluations. In a study by Solarz (1960), participants responded to positive and negative words (e.g., smart, stupid) by operating a lever in one of two ways: by pulling it toward them, an arm movement consistent with approach behavior, or by pushing it away from them, an arm movement associated with avoiding an object. Half of the participants were instructed to pull the lever for words that they liked and to push the lever if they saw a word they didn't like. The other participants were told to do the opposite. Participants responded significantly faster when the word valence was consistent with the evaluation implied by the motor movement: They pulled the lever more quickly in response to a positive word and pushed it more quickly in response to a negative one. Chen and Bargh (1999) replicated Solarz' findings and showed that the effect persisted even when participants were not explicitly instructed to evaluate the target stimuli. Moreover, several recent studies have used the strength of arm extension and flexion as indicators of the motivation to approach or avoid a valenced stimulus (see Förster, Higgins, & Idson, 1998).

Paper-and-pencil measures also offer simple means of implicit measurement. For example, a relatively easy way to assess attitude accessibility is by means of a word-fragment completion task. Participants complete letter strings to form complete words (e.g., POL_E—POLITE). Construct accessibility influences participants' choices of how to complete a given word fragment (Bassili & Smith, 1986; Tulving, Schacter, & Stark, 1982). If a letter string can be completed with either attitude-related or unrelated words, the task can be used as a quick indicator of attitude accessibility. Likewise, if the possible completions include both positive and negative alternatives, it may be used to assess attitude valence as well (e.g., B_D—BAD vs. BUD, see Dovidio et al., 1997).

A slightly more complicated implicit paper-and-pencil measure has been used in research on intergroup attitudes. Proposed by von Hippel, Sekaquaptewa, and Vargas (1997), this measure is based on evidence that people tend to describe behavior in more abstract terms when the behavior is consistent with expectations (Maass, Salvi, Arcuri, & Semin, 1989). Participants are presented with

several ostensible news clippings that describe stereotypic and counterstereotypic events involving either ingroup or outgroup targets. The events systematically vary in terms of the valence of the described behavior. Participants then rate a set of possible headlines for how well they capture the described event. The headlines vary in the level of linguistic abstraction (e.g., *Johnson performs 360-degree slam-dunk* vs. *Johnson is athletic*). Of interest is the degree to which participants show a bias in favor of abstract headlines when they describe negative events as opposed to positive behaviors for the outgroup target.

Limitations of Implicit Measures

Since their introduction, implicit measures of attitudes have seen both overly enthusiastic reactions (due to the promise of unprecedented opportunities for attitude measurement) as well as overly harsh criticisms (due to real and perceived shortcomings of the measures). Some of the criticisms are conceptual in nature; others focus on the practical effectiveness of the measures and their psychometric properties.

CONCEPTUAL ISSUES

Extrapersonal Knowledge We begin with the most fundamental criticism of implicit measures, that in fact they do not measure a person's attitudes at all, but simply capture culturally pervasive knowledge. This argument has been raised in specific form regarding particular types of implicit measures (i.e., the IAT, Fiedler, Messner, & Bluemke, 2006; Karpinski & Hilton, 2001; Olson & Fazio, 2004), as well as in general terms against any measure meant to tap automatic evaluations (Arkes & Tetlock, 2004; Kihlstrom, 2004; Mitchell & Tetlock, 2006).

Perhaps not surprisingly, the criticism has been presented most vehemently in contexts where implicit measures are applied to socially sensitive attitudes, such as racial attitudes and prejudice (e.g., Arkes & Tetlock, 2004). The concern is that implicit measures of racial prejudice cannot distinguish between the mere knowledge of cultural stereotypes that associate African Americans with, for example, violent behavior and acceptance of such stereotypic and negative views. Certainly, knowledge does not necessarily imply acceptance. Thus, respondents may be labeled as prejudiced, not because of the beliefs they hold, but because they are familiar with cultural views.

No doubt the distinction between accepted standpoint and merely known positions is important. Yet, the notion that cultural knowledge has no role in evaluative responses is simply at odds with contemporary conceptualizations of attitudes and their operation (see Gawronski & Bodenhausen, 2006; Nosek & Hansen, 2008; Wittenbrink, 2004).

As we described at the outset of this chapter, attitudes are commonly defined as evaluative predispositions. Sometimes, these predispositions may be based on a single source. More commonly, however, attitudes stem from multiple sources of input. With regard to attitudes toward a familiar social group, a person is likely to hold many stored associations, of which cultural stereotypes, known group members, or personal experiences with individual group members may be some. The cultural knowledge criticism implies that of all the input sources, only those will impact a person's evaluative predisposition that are deemed valid and are explicitly endorsed. There are two basic problems with this position. First, as an extensive literature on context effects has shown, explicit endorsement varies considerably across situations (see Sudman, Bradburn, & Schwarz, 1996). Acceptance per se is therefore a poor and ambiguous criterion for determining what kinds of influences a proper attitude measure should actually assess. Second, it is now well-established that evaluation may occur spontaneously, without intent, and without control over or even awareness of the response (Bargh, Chaiken, Govender, & Pratto, 1992; Fazio et al.,

1986; Giner-Sorolla, Garcia, & Bargh, 1999; Greenwald, Klinger, & Liu, 1989; Kunst-Wilson & Zajonc, 1980; Wittenbrink et al., 1997). Of course, without awareness, without control to edit one's response, and in the absence of intent to even evaluate in the first place, evaluations do occur without the opportunity to reflect on one's true attitude. Such evaluative responses are precisely what implicit measures aim to capture. Whether the responses stem from cultural knowledge, or from accepted sentiments, or both is an empirical question. However, there is every reason to assume that in principle both cultural knowledge and personal sentiments have a role in attitudinal responses.

Automaticity Ample empirical evidence exists by now showing that the implicit measures we have reviewed in this chapter do not meet strict automaticity criteria. For example, responses are not obligatory, fixed reactions to the attitude objects included in the measurement. In part, this may reflect the fact that automatic attitudinal responses themselves vary by context (see Blair, 2002; Wittenbrink, Judd, & Park, 2001a) and that implicit measures therefore simply capture this variation in the latent construct. Yet, various studies have also demonstrated influences of the measurement procedure itself. We already mentioned earlier that the EPT and the LDT operate differently despite the fact that these measures are procedurally identical, except for their choice of judgment task. In fact, a procedure's judgment task and the implied processing goals critically influence the measurement outcome (Klauer & Musch, 2002; Wittenbrink, Judd, & Park, 2001b). For example, Wittenbrink et al. (2001b) observed different patterns of activation as a result of manipulating the task instructions in a priming measure of racial attitudes. In the context of a lexical decision task, group primes showed facilitation for trait attributes associated with the respective group stereotype. Moreover, outgroup primes yielded disproportionately strong facilitation for negative stereotypic attributes compared to ingroup primes. However, when the same priming procedure was administered with an evaluative decision task, the stereotypicality of the target items did not matter for the observed priming effects. Now, outgroup primes produced overall stronger facilitation for any negatively valenced attribute.

Aside from processing goals imposed by the task, implicit measures are subject to strategic influences more generally. For example, the exemplars chosen to represent a particular attitude object may often be classified in multiple ways (e.g., *rose* and *butterfly* as *flower* vs. *insect*, or simply as *beautiful things*). The categorization that will ultimately impact responses is influenced by strategic considerations made by the participant (Govan & Williams, 2004; Mitchell, Nosek, & Banaji, 2003; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009) or implied by the task instructions (Livingston & Brewer, 2002; Olsen & Fazio, 2003).

Of course, given the use of implicit measures as less reactive indirect attitude measures, the one automaticity feature that is likely to matter the most to researchers is the controllability of responses. In principle, automatic response are uncontrollable. However, given that implicit attitude measures are open to strategic influences, they may also be subject to intentional control. Several studies have addressed the impact of strategic attempts at misrepresentation for the IAT, the EPT, and the AMP.

Faking Participants can control IAT effects when they have prior IAT experience (Fiedler & Bluemke, 2005; Steffens, 2004) or when they receive explicit advice on how to do so effectively—for example, by slowing down responses on compatible trials and speeding up responses on incompatible trials (Fiedler & Bluemke, 2005; but see De Houwer, Beckers, & Moors, 2007, or Wallaert, Ward, & Mann, 2010 for examples of effective control even without these preconditions).

With regard to the controllability of responses in the EPT, the available data are somewhat less clear. Several studies have concluded that the task cannot be faked, especially when prime and target are presented in rapid succession (stimulus-onset asynchrony of less than 300 ms) and a response deadline (600 ms) is enforced (Bar-Anan, 2010; Degner, 2009). However, Teige-Mocigemba and

Klauer (2013) were able to obtain successfully faked responses on an EPT of political attitudes and outgroup attitudes with increased motivation to misrepresent and more effective instructions on how to exactly mislead the EPT measure. Of course, concealment of the attitude primes from conscious awareness ought to limit any opportunities for strategic misrepresentation.

Payne and colleagues already presented data on the controllability of the AMP's misattribution effect with the original introduction of the measure. Explicit instructions to participants not to let the primes influence their target evaluations did not undermine the measure's effectiveness (Payne et al., 2005). However, more recent studies suggest the AMP to be easily faked, without any specific advice on how to actually influence the measure. The possible flaw here is the obvious transparency of the procedure. In as much as participants recognize the purpose of the prime presentation, responses can be readily edited by ignoring the targets altogether and by responding directly to the prime in the socially desired fashion (Teige-Mocigemba et al., 2016). Like other priming measures, the AMP offers the opportunity to use concealed attitude primes to limit such misrepresentation (Rohr et al., 2015).

MEASUREMENT EFFECTIVENESS: RELIABILITY AND VALIDITY

Several studies have investigated the measurement effectiveness of implicit measures (e.g., Banse, 1999, 2001; Bar-Anan & Nosek, 2014; Bosson, Swann, & Pennebaker, 2000; Cameron, Brown-Iannuzzi, & Payne, 2012; Cunningham, Preacher, & Banaji, 2001; Hofmann et al., 2005; Oswald et al., 2013). A consistent finding is that EPT and LDT show relatively limited reliability, whereas both IAT (and its variants) and AMP fare better on measures of internal consistency (e.g., IAT: $\alpha = .88$ and the AMP: $\alpha = .69$; Bar-Anan & Nosek, 2014). However, even for these measures test-retest correlations remain modest (e.g., IAT: $r = .45$ and the AMP: $r = .50$; Bar-Anan & Nosek, 2014), suggesting implicit measures carry a sizeable measurement error. Another explanation for the modest test-retest reliability is that the to-be-measured construct itself is not stable across measurements. This is an interesting alternative as situational variation in attitudes has been a long-standing challenge in attitude measurement (Ajzen & Fishbein, 1977; Schwarz, 2007; Wicker, 1969). We already noted that automatic components of evaluation, which implicit measures aim to capture, turn out not to be as obligatory and fixed as they had initially been conceptualized (cf. Fazio et al., 1986). Instead they have proven to be more malleable than anticipated (Blair, 2002). In principle, the relatively lower reliability of implicit measures, compared to explicit self-report measures, could mean that automatic components of evaluation are even more fickle than deliberate forms of evaluation. Either way, their moderate reliability places upper limits on the extent to which implicit measures can predict attitude-relevant behavior or, more generally, perform well on measures of predictive validity (see a parallel discussion in Ajzen, Fishbein, Lohmann, & Albarracín, this volume).

Investigations into the predictive validity of implicit measures have produced some markedly disappointing results. For example, in one of the first studies to explore the issue, Bosson and colleagues found IAT and EPT measures of self-esteem virtually unrelated to self-report and observer-based criterion measures (Bosson et al., 2000). The reported bivariate correlations ranged from $r = .25$ to $r = -.19$, with almost half of the coefficients showing inverse relationships and only 3 out of 18 correlations reaching statistical significance. More recently, Oswald and colleagues concluded that, based on a meta-analysis of studies investigating the predictive validity of racial attitude IATs, "the IAT provides little insight into who will discriminate against whom" (Oswald et al., 2013, p. 188). Their reported IAT/criterion correlation was $r = .148$.

Such findings seriously question the usefulness of implicit measures for the prediction of attitudinal responses, and may actually place doubt on their validity altogether. However, individual validity studies are often underpowered. For example, the data by Bosson et al. (2000) were based on a total 44 participants. The diagnosticity of meta-analyses on the other hand depends on the researchers'

decisions regarding the data aggregation. Casting the net widely increases power but also runs the risk of including criteria that the measure is not meant to predict. The race IAT meta-analysis by Oswald et al. (2013), for example, included 309 separate IAT-criterion correlations (from a total of 86 independent samples). Yet, 33 of these correlations involved criteria capturing responses to White Americans or a White target person. In principle, the inclusion of White attitude criteria may be justified by the relative nature of the IAT measure, which as we noted earlier is always based on two contrasting categories (here: Black and White people). But to the extent that for the predominantly White participants a race IAT captures in effect attitudes toward African Americans, the inclusion of behaviors and judgments regarding white targets is bound to reduce its observed predictive validity.[6] In fact, Greenwald, Banaji, and Nosek (2015) note that different decisions regarding data inclusion can readily explain much of the discrepancies between the pessimistic summary offered by the Oswald et al. (2013) meta-analysis and the results of an earlier meta-analytic assessment of race IATs by Greenwald, Poehlman, Uhlmann, and Banaji (2009).

This study by Greenwald et al. (2009) still remains the most comprehensive meta-analysis of the IAT's predictive validity to date. Aggregating the results of 184 independent samples from studies across a variety of attitude domains, it reports an overall average correlation of $r = .274$ between criterion measures and their respective IATs. Explicit self-report measures showed a somewhat higher correlation overall ($r = .361$), yet fared worse for socially sensitive attitudes. In particular, for racial attitudes, the average IAT-criterion correlation was $r = .236$, compared to an $r = .118$ for the self-report-criterion correlation. The effect sizes reported by Greenwald and colleagues for the IAT are also in line with the results of a recent meta-analysis for priming measures by Cameron et al. (2012). Based on a total of 167 independent samples, AMP, LDT, and EPT showed correlations with criterion behaviors of $r = .35$, $r = .29$, and $r = .25$, respectively. Likewise, a large-scale online study with more than 24,000 volunteer participants on racial attitudes, political attitudes, and self-esteem by Bar-Anan and Nosek (2014) yielded correlations of similar magnitude between self-reported attitudes and several implicit measures (ranging from a maximum of $r = .38$ for the BIAT to a minimum of $r = .24$ for the EPT).

Beyond correlational evidence on the predictive validity of implicit measures, a few experimental and quasi-experimental studies have addressed the measures' construct validity. For example, Olson and Fazio (2001) showed that an IAT successfully captured experimentally induced novel attitudes. Likewise, implicit measures have been found to be sensitive to construct-related interventions. Lowery, Hardin, and Sinclair (2001) observed attenuated prejudice on an IAT when the task was administered by a Black, rather than White, experimenter. Wittenbrink et al. (2001a, Study 1) showed moderation of IAT-measured prejudice by preceding movie clips that portrayed African Americans in either stereotypically positive or negative fashion. Other studies have demonstrated that implicit measures successfully detect a priori known group differences (e.g, Bar-Anan & Nosek, 2014; Fazio et al., 1995; Nosek et al., 2007; Payne et al., 2005).

A recurrent finding of studies investigating the validity of implicit measures has been that these measures have at times shown only weak relationships with one another (e.g., Bosson et al., 2000; Olson & Fazio, 2003). Bar-Anan and Nosek (2014) have speculated that low intercorrelations are specific to particular attitude domains. While that may be the case, the finding suggests that the different implicit measures carry considerable method-specific variance. Proposals to data-analytically decompose the measures into separate estimates of underlying processes might help to shed further light on this issue (e.g., Conrey et al., 2005; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007). Such a strategy might also prove useful to further improve the predictive validity of the attitude estimates obtained by implicit measures.

On balance, this available evidence suggests that implicit measures show adequate construct validity and that they offer moderate predictive validity. Overall, their ability to predict attitudinal responses is equivalent to what traditional self-report measures can accomplish (e.g., Kraus, 1995;

Talaska, Fiske, & Chaiken, 2008). By comparison, implicit measures appear to be more effective predictors of attitudes that are socially sensitive. But even in those cases, the portion of criterion variance explained remains limited. The value of implicit measures therefore lies foremost in the prediction of aggregate effects across many individuals or actions, rather than in offering predictions for individual cases. For example, evidence of racial prejudice on an implicit measure among the general public might help to explain societal patterns of discrimination. But a prejudiced IAT score for an individual police officer won't reliably predict the officer's actions on the next traffic stop.

Conclusion

Attitude researchers have many techniques available to them for assessing the constructs they study, and these various techniques all offer useful handles for empirical study. The future of attitude measurement research will no doubt be very interesting, as the relations among implicit measures become better understood and as their relations to direct self-reports of attitudes become better understood as well. In the meantime, we see value in the classic approach to measurement: Any study of a construct is more likely to be informative if multiple measures of that construct are used instead of just one. Only then can issues of construct validity be successfully addressed.

Although implicit measures of attitude offer great promise, in terms of their ability to assess attitudes freed of participants' self-presentational concerns, at present their claims to validity rest largely on intuitive appeals. It seems crucial that researchers in attitude measurement establish that such measures in fact predict socially significant criterion behaviors.

Additionally, as we claimed in the beginning to this chapter, attitudes are not simple productions that emerge intact, ripe for measurement. Rather they manifest themselves in many different shapes, as a result of complex cognitive processes. Our measures need to be sensitive to the ways in which they may be produced. In some situations, assessments of automatically formed evaluations may be most important in predicting behaviors. In others, more deliberative, and potentially critically monitored, evaluative responses may be what we want to measure. Just because a participant is unaware that his or her attitude is being assessed, that does not mean that the attitude in question has been measured with greater construct validity.

Without doubt both traditional self-report and more indirect attitude measures will continue to be used. The goal is not to come up with a single "best" attitude measure but rather to measure attitudes in all their complexity and all their manifestations.

Notes

1 Almost all studies reviewed above involved experimental designs varying the number of rating scale points, holding constant all other aspects of questions. Some additional studies have explored the impact of number of scale points using a different approach: meta-analysis. These studies have taken large sets of questions asked in pre-existing surveys, estimated their reliability and/or validity, and meta-analyzed the results to see whether data quality varies with scale point number (e.g., Alwin, 1992, 1997; Alwin & Krosnick, 1991; Andrews, 1984, 1990; Scherpenzeel, 1995). However, these meta-analyses sometimes mixed together measures of subjective judgments with measurements of objective constructs such as numeric behavior frequencies (e.g., number of days) and routinely involved strong confounds between number of scale points and other item characteristics, only some of which were measured and controlled for statistically. Consequently, it is not surprising that these studies yielded inconsistent findings. For example, Andrews (1984) found that validity and reliability were worst for 3-point scales, better for 2-point and 4-point scales, and even better as scale length increased from 5 points to 19 points. In contrast, Alwin and Krosnick (1991) found that 3-point scales had the *lowest* reliability; found no difference in the reliabilities of 2-, 4-, 5-, and 7-point scales; and found 9-point scales to have maximum reliability (though these latter scales actually offered 101 response alternatives to participants). And Scherpenzeel (1995) found the highest reliability for 4-/5-point scales, lower reliability for 10 points, and even lower for 100 points. We therefore view these studies as less informative than experiments manipulating rating scale length.

2 The Stroop Color-Word Test itself has been used as an implicit measure of attitudes. Pratto and John (1991) reasoned that negative words would show more interference on the color-naming task than would positive words. Results from several studies are consistent with this argument, showing increased response latencies for negative words, whereas positive or neutral words did not affect the color-naming task. However, valence effects are potentially confounded with effects of accessibility in this type of measure as highly accessible attitudes have generally been found to direct attention, not just when they are negative (Smith, Fazio, & Cejka, 1996).
3 Aside from this association-based response interference, other cognitive mechanisms like task switching have been suggested to contribute to IAT effects (Conrey et al., 2005; Klauer & Mierke, 2005, Klauer et al., 2007).
4 In principle, it is possible to calculate a separate IAT score for trials that involve one of the two attitude categories and another score for trials of the second attitude category. However, this computational strategy has been found to be unreliable in obtaining absolute rather than relative measures of evaluation (Nosek, Greenwald, & Banaji, 2005).
5 Different from the LDT, the EPT requires additional instructions to justify the presence of the attitude primes. And because the target dimension of interest is focal to the judgment task, the EPT can only accommodate targets that vary on a single attribute dimension.
6 The average correlation for the 33 race IAT and White attitude criteria in Oswald et al. (2013) is close to zero ($r = -.020$ with a 95% confidence interval of +/− .06; see Greenwald, Banaji, & Nosek, 2015, footnote 3).

References

Adolphs, R., Tranel, D., & Damsio, A. R. (1998). The human amygdala in social judgment. *Nature*, *393*(6684), 470–474.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper and Row.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*, 888–918.

Allen, B. P. (1975). Social distance and admiration reactions of "unprejudiced" Whites. *Journal of Personality*, *43*, 709–726.

Allison, P. D. (1975). A simple proof of the Spearman-Brown formula for continuous length tests. *Psychometrika, 40*, 135–136.

Allport, G. W. (1935). Attitudes. In C. Murchinson (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22*, 83–118.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research, 25*, 318–340.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research, 20*, 139–181.

Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988). The effects of race of the interviewer on measures of electoral participation by Blacks in SRC national election studies. *Public Opinion Quarterly*, *52*, 53–83.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly, 48*, 409–442.

Andrews, F. M. (1990). Some observations on meta-analysis of MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies*. Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry*, *15*, 257–279.

Atkin, C. K., & Chaffee, S. H. (1972–1973). Instrumental response strategies in opinion interview. *Public Opinion Quarterly*, *36*, 69–79.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The psychology of learning and motivation: Advances in research and theory (Vol. 2, pp. 89–195). New York: Academic Press.

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229–247.

Baddeley, A. D., & Hitch, G. J. (1977). Recency reexamined. In S. Dornic (Ed.), *Attention and performance*. Hillsdale, NJ: Erlbaum.

Banse, R. (1999). Automatic evaluation of self and significant others: Affective priming in close relationships. *Journal of Social and Personal Relationships*, *16*, 805–824.

Banse, R. (2001). Affective priming with liked and disliked persons: Prime visibility determines congruency and incongruency effects. *Cognition & Emotion*, *15*, 501–520.

Bar-Anan, Y. (2010). Strategic modification of the evaluative priming effect does not reduce its sensitivity to uncontrolled evaluations. *Journal of Experimental Social Psychology*, *46*, 1101–1104.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality & Social Psychology Bulletin*, *38*, 1194–1208.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven implicit measures of social cognition. *Behavior Research Methods*, *46*, 668–688.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Lawrence Erlbaum Associates.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality & Social Psychology*, *62*(6), 893–912.

Bassili, J. N., & Brown, R. D. (2005). Implicit and explicit attitudes: Research, challenges, and theory. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 543–574). Mahwah, NJ: Lawrence Erlbaum Associates.

Bassili, J. N., & Smith, M. C. (1986). On the spontaneity of trait attribution: Converging evidence for the role of cognitive strategy. *Journal of Personality & Social Psychology*, *50*(2), 239–245.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly, 18*, 271–278.

Bendig, A. W. (1954). Reliability and the number of rating scale categories. *The Journal of Applied Psychology, 38*, 38–40.

Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology*, *12*(5), 409–415.

Berg, I. A, & Rapaport, G. M. (1954). Response bias in an unstructured questionnaire. *Journal of Psychology, 38*, 475–481.

Birkett, N. J. (1986). Selecting the number of response categories for a likert-type scale. *Proceedings of the American Statistical Association*, 488–492.

Bishop, G. F., Hippler, H. J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massedy, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 321–334). New York: Wiley.

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly, 51*, 220–232.

Bishop, G. F. (1990). Issue involvement and response effects in public opinion surveys. *Public Opinion Quarterly, 54*, 209–218.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1979). Effects of opinion filtering and opinion floating: Evidence from a secondary analysis. *Political Methodology, 6*, 293–309.

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*(1), 27.

Blaison, C., Imhoff, R., Hühnel, I., Hess, U., & Banse, R. (2012). The affect misattribution procedure: Hot or not? *Emotion*, *12*, 403–412.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (STIAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*, 977–997.

Bodenhausen, G. V., Schwarz, N., Bless, H., & Wänke, M. (1995). Effects of atypical exemplars on racial beliefs: Enlightened racism or generalized appraisals?. *Journal of Experimental Social Psychology, 31*, 48–48.

Bogart, L. (1972). *Silent politics: Polls and the awareness of public opinion*. New York: Wlley-Interscience.

Bornstein, R. F., & D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, *12*, 103–128. doi:10.1521/soco.1994.12.2.103

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.

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1993). Pictures as prepulse: Attention and emotion in startle modification. *Psychophysiology*, *30*(5), 541–545.

Brewer, M. B. (2001). Ingroup identification and intergroup conflict: When does ingroup love become outgroup hate? In R. D. Ashmore & L. Jussim (Eds.), *Social identity, intergroup conflict, and conflict reduction. Rutgers series on self and social identity* (Vol. 3, pp. 17–41). London, UK: Oxford University Press.

Cacioppo, J. T., Crites, S. L., Berntson, G. G., & Coles, M. G. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science*, *4*(2), 108–112.

Cacioppo, J. T., Crites, S. L., Gardner, W. L., & Berntson, G. G. (1994). Bioelectrical echoes from evaluative categorizations: I. A late positive brain potential that varies as a function of trait negativity and extremity. *Journal of Personality & Social Psychology*, *67*(1), 115–125.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality & Social Psychology*, *50*(2), 260–268.

Cacioppo, J. T., & Sandman, C. A. (1981). Psychophysiological functioning, cognitive responding, and attitudes. In R. E. Petty, T. M. Ostrom & T. C. Brock (Eds.), *Cognitive responses in persuasion* (pp. 81–103). Hillsdale, NJ: Lawrence Erlbaum Associates.

Calsyn, R. J., Roades, L. A., & Calsyn, D. S. (1992). Acquiescence in needs assessment studies of the elderly. *The Gerontologist, 32*, 246–252.

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.

Cameron, J. A., Alvarez, J. M., & Bargh, J. A. (2000). Examining the validity of implicit measures of prejudice. Paper presented at the First meeting of the Society for Personality and Social Psychology, Nashville, TN.

Campbell, B. A. (1981). Race-of-interviewer effects among southern adolescents. *Public Opinion Quarterly*, *45*, 231–244.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campbell, D. T., & Mohr, P. J. (1950). The effect of ordinal position upon responses to items in a checklist. *Journal of Applied Psychology, 34*, 62–67.

Carp, F. M. (1974). Position effects on interview responses. *Journal of Gerontology, 29*, 581–587.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario symposium* (Vol. 3, pp. 143–177). Hillsdale, NJ: Lawrence Erlbaum Associates.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23*, 323–337.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement, 51*, 531–540.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality & Social Psychology Bulletin*, *25*(2), 215–224.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy versus impression motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*, 262–275.

Cheng, S. (1988). Subjective quality of life in the planning and evaluation of programs. *Evaluation and Program Planning*, *11*, 123–134.

Clancy, K. J., & Wachsler, R. A. (1971). Positional effects in shared-cost surveys. *Public Opinion Quarterly, 35*, 258–265.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186–190.

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.

Converse, J. M. (1976). Predicting no opinion in the polls. *Public Opinion Quarterly, 40*, 515–530.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage.

Converse, P. E. (1964). The nature of belief systems in the mass public. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.

Coombs, C. H., & Coombs, L. C. (1976). "Don't know": Item ambiguity or respondent uncertainty? *Public Opinion Quarterly, 40*, 497–514.

Cotter, P., Cohen, J., & Coulter, P. B. (1982). Race of interviewer effects in telephone interviews. *Public Opinion Quarterly*, *46*, 278–294.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*, 151–174.

Crites, S. L., Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1995). Bioelectrical echoes from evaluative categorization: II. A late positive brain potential that varies as a function of attitude registration rather than attitude report. *Journal of Personality & Social Psychology*, *68*(6), 997–1013.

Cronbach, W. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3–31.
Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Multifacet studies of generalizability*. New York: Wiley.
Culpepper, I. J, Smith, W. R., & Krosnick, J. A. (1992). The Impact of Question Order on Satisficing in Surveys. Paper presented at the Midwestern Psychological Association Annual Meeting, Chicago, Illinois.
Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163–170.
Cuthbert, B. N., Bradley, M. M., & Lang, P. J. (1996). Probing picture perception: Activation and emotion. *Psychophysiology*, *33*(2), 103–111.
Dabbs, J. M., Jr., Bassett, J. F., & Dyomina, N. V. (2003). The Palm IAT: A portable version of the Implicit Association Test. *Behavior Research Methods, Instruments, and Computers, 35*, 90–95.
Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy build on myth*. New York: The Free Press.
Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 497–548). Boston, MA: McGraw-Hill.
Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 1, pp. 509–566). Hillsdale, NJ: Lawrence Erlbaum.
De Houwer, J. (2003a). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Lawrence Erlbaum Associates.
De Houwer, J. (2003b). The extrinsic affective Simon task. *Experimental Psychology*, *50*, 77–85.
De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology*, *43*, 972–978.
De Houwer, J., & De Bruycker, E. (2007a). 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., & De Bruycker, E. (2007b). The Implicit Association Test outperforms the Extrinsic Affective Simon Task as a measure of interindividual differences in attitudes. *British Journal of Social Psychology*, *46*, 401–421.
De Houwer, J., Hermans, D., & Eelen, P. (1998). Affective and identity priming with episodically associated stimuli. *Cognition & Emotion, 12*(2), 145–169.
De Houwer, J., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorization responses. *Cognition and Emotion*, *16*, 643–666.
De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 179–194). New York, NY: Guilford Press.
Degner, J. (2009). On the (un-)controllability of affective priming: Strategic manipulation is feasible but can possibly be prevented. *Cognition and Emotion*, *23*, 327–354.
DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*, 979–995.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.
Dickinson, J. R., & Kirzner, E. (1985). Questionnaire item omission as a function of within-group question position. *Journal of Business Research, 13*, 71–75.
Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating mixed standard scale formats. *Journal of Applied Psychology, 65*, 147–154.
Donnerstein, E., & Donnerstein, M. (1975). The effect of attitudinal similarity on interracial aggression. *Journal of Personality*, *43*(3), 485–502.
Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality & Social Psychology*, *82*(1), 62–68.
Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*, 510–540.
Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Forth Worth, TX: Harcourt Brace.
Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*, 267–278.
Eifermann, R. R. (1961). Negation: A linguistic variable. *Acta Psychologica, 18*, 258–273.
England, L. R. (1948). Capital punishment and open-end questions. *Public Opinion Quarterly, 12*, 412–416.
Eurich, A. C. (1931). Four types of examinations compared and evaluated. *Journal of Educational Psychology, 22*, 268–278.

Evans, R. I., Hansen, W. B., & Mittlemark, M. B. (1977). Increasing the validity of self-reports of smoking behavior in children. *Journal of Applied Psychology*, *62*, 521–523.

Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). Definition, identification, and reliability of measurement the P300 component of the event-related brain potential. In P. K. Ackles, J. R. Jennings, & M. G. Coles (Eds.), *Advances in psychophysiology* (Vol. 2, pp. 1–78). Greenwich, CT: JAI Press.

Faulkenberry, G. D., & Mason, R. (1978). Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly, 42*(4), 533–543.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). San Diego, CA: Academic Press.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*(2), 115–141.

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.

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*, 229–238.

Ferber, R. (1966). Item nonresponse in a consumer survey. *Public Opinion Quarterly, 30*, 399–415.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*, 307–316.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*, 74–147.

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *Public Opinion Quarterly*, *55*, 313–330.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Folsom, J. K. (1931). *Social psychology*. New York: Harper & Brothers.

Fonda, C. P. (1951). The nature and meaning of the Rorschach white space response. *Journal of Abnormal and Social Psychology, 46*, 367–377.

Forehand, G. A. (1962). Relationships among response sets and cognitive behaviors. *Educational and Psychological Measurement, 22*, 287–302.

Förster, J., Higgins, E., & Idson, L. C. (1998). Approach and avoidance strength during goal attainment: Regulatory focus and the "goal looms larger" effect. *Journal of Personality & Social Psychology*, *75*(5), 1115–1131.

Fridlund, A. J., Schwartz, G. E., & Fowler, S. C. (1984). Pattern recognition of self-reported emotional state from multiple-site facial EMG activity during affective imagery. *Psychophysiology*, *21*(6), 622–637.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of white racism, arousal, and helping behavior. *Journal of Personality and Social Psychology*, *35*, 691–707.

Gage, N. L., Leavitt, G. S., & Stone, G. C. (1957). The psychological meaning of acquiescence set for authoritarianism. *Journal of Abnormal and Social Psychology, 55*, 98–103.

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., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, *24*, 218 225.

Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? *Personality and Social Psychology Bulletin*, *40*, 3–15.

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly, 52*, 365–371.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). New York: Guilford Press.

Gilljam, M., & Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly, 57*, 348–357.

Giner-Sorolla, R., Garcia, M. T., & Bargh, J. A. (1999). The automatic evaluation of pictures. *Social Cognition*, *17*(1), 76–96.

Givon, M. M., & Shapira, Z. B. (1984). Response to rating scales: A theoretical model and its application to the number of categories problem. *Journal of Marketing Research, 21*, 410–419.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 311–325.

Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Doubleday/Anchor Books.
Goldsmith, R. E. (1987). Two studies of yeasaying. *Psychological Reports, 60*, 239–244.
Gordon, R. A. (1987). Social desirability bias: A demonstration and technique for its reduction. *Teaching of Psychology*, *14*, 40–42.
Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by redefining the category labels. *Journal of Experimental Social Psychology*, *40*, 357–365.
Gove, W. R., & Geerken, M. R. (1977). Response bias in surveys of mental health: An empirical investigation. *American Journal of Sociology, 82*, 1289–1317.
Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Violence viewed by psychopathic murderers: Adapting a revealing test may expose those psychopaths who are most likely to kill. *Nature, 423*, 497–498.
Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery—How many scales and response categories to use? *Journal of Marketing, 34*, 33–39.
Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. *Advances in Experimental Social Psychology*, *29*, 61–139.
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., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553–561.
Greenwald, A. G., Klinger, M. R., & Liu, T. J. (1989). Unconscious processing of dichoptically masked words. *Memory & Cognition*, *17*(1), 35–47.
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., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality & Social Psychology*, *85*(2), 197–216.
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.
Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess attitudinal components. *British Journal of Social Psychology, 37*(2), 129–149.
Hanley C. (1959). Responses to the wording of personality test items. *Journal of Consulting Psychology, 23*, 261–265.
Hanley, C. (1962). The "difficulty" of a personality inventory item. *Educational and Psychological Measurement, 22*, 577–584.
Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport*, *11*(11), 2351–2355.
Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*, 46–54.
Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *The psychology of attitudes* (pp. 23–87). Fort Worth, TX: Harcourt Brace Jovanovich.
Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, *43*, 710–717.
Hippler, H. J., & Schwarz, N. (1989). "No-opinion" filters: A cognitive perspective. *International Journal of Public Opinion Research, 1*, 77–87.
Hoffman, P. J. (1960). Social acquiescence and "education." *Educational and Psychological Measurement, 20*, 769–776.
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.
Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone vs. face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*, 79–125.
Holbrook, A. L., Krosnick, J. A., Visser, P. S., Gardner, W. L., & Cacioppo, J. T. (2001). Attitudes toward presidential candidates and political parties: Initial optimism, inertial first impressions, and a focus on flaws. *American Journal of Political Science, 45*, 930–950.
Holmes, C. (1974). A statistical evaluation of rating scales. *Journal of the Market Research Society, 16*, 86–108.
Hough, K. S., & Allen, B. P. (1975). Is the "women's movement" erasing the mark of oppression from the female psyche? *Journal of Psychology*, *89*, 249–258.
Houston, M. J., & Nevin, J. R. (1977). The effects of source and appeal on mail survey response patterns. *Journal of Marketing Research, 14*, 374–378.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change*. New Haven, CT: Yale University Press.

Hurd, A. W. (1932). Comparisons of short answer and multiple choice tests covering identical subject content. *Journal of Educational Psychology, 26*, 28–30.

Isard, E. S. (1956). The relationship between item ambiguity and discriminating power in a forced-choice scale. *Journal of Applied Psychology, 40*, 266–268.

Israel, G. D., & Taylor, C. L. (1990). Can response order bias evaluations? *Evaluation and Program Planning, 13*, 365–371.

Jackman, M. R. (1973, June). Education and prejudice or education and response-set? American *Sociological Review, 38*, 327–339.

Jackson, D. N. (1959). Cognitive energy level, acquiescence, and authoritarianism. *Journal of Social Psychology, 49*, 65–69.

Jacoby, J., & Matell, M. S. (1971). Three-point likert scales are good enough. *Journal of Marketing Research, 7*, 495–500.

Jajodia, A., & Earleywine, M. (2003). Measuring alcohol expectancies with the implicit association test. *Psychology of Addictive Behaviors, 17*, 126–133.

Jenkins, G. D., & Taber, T. D. (1977). A monte carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392–398.

Johanson, G. A., Gips, C. J., & Rich, C. E. (1993). If you can't say something nice, a variation on the social desirability response set. *Evaluation Review, 17*, 116–122.

Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. *Journal of Social Psychology, 113*, 21–27.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed, Vol. 1, pp. 180–232). Cambridge: Cambridge University Press.

Kahn, D. F., & Hadley, J. M. (1949). Factors related to life insurance selling. *Journal of Applied Psychology, 33*, 132–140.

Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion questions. *Applied Statistics, 27*, 149–161.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *The Statistician, 29*, 65–79.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 774–788.

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.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.

Kenny, D. A., & Kashy, D. A. (1992). The analysis of the multitrait – multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. Morf, & A. Panter (Eds.), *The Sage handbook of methods in social psychology* (pp. 195–212). Thousand Oaks, CA: Sage.

Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of White opinion on affirmative action for Blacks. *Social Cognition, 8*, 73–103.

Klare, G. R. (1950). Understandability and indefinite answers to public opinion questions. *International Journal of Opinion and Attitude Research, 4*, 91–96.

Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the Implicit Association Test effect. *Personality and Social Psychology Bulletin, 31*, 208–217.

Klauer, K. C., & Musch, J. (2002). Goal-dependent and goal-independent effects of irrelevant evaluations. *Personality & Social Psychology Bulletin, 28*(6), 802–814.

Klauer, K. C., Roßnagel, C., & Musch, J. (1997). List-context effects in evaluative priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 246 255.

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.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis-testing. *Psychological Review, 94*, 211–228.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

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.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*, 85–96.

Klopfer, F. J., & Madden, T. M. (1980). The middlemost choice on attitude items: Ambivalence, neutrality, or uncertainty. *Personality and Social Psychology Bulletin, 6*, 97–101.

Komorita, S. S. (1963). Attitude context, intensity, and the neutral point on a likert scale. *Journal of Social Psychology, 61*, 327–334.

Komorita, S. S., & Graham, W. K. (1965). Number of scale pints and the reliability of scales. *Educational and Psychological Measurement, 25*, 987–995.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin*, *21*, 58–75.

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.

Krosnick, J. A. (1990). Americans' perceptions of presidential candidates: A test of the projection hypothesis. *Journal of Social Issues, 46*, 159–182.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236.

Krosnick, J. A. (1992). The impact of cognitive sophistication and attitude importance on response order effects and question order effects. In N. Schwarz & S. Sudman (Eds.) *Order effects in social and psychological research* (pp. 203–218). New York: Springer.

Krosnick, J. A. (1999). Survey methodology. *Annual Review of Psychology, 50*, 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201–219.

Krosnick, J. A., & Berent, M. K. (1990). The impact of verbal labeling of response alternatives and branching on attitude measurement reliability in surveys. Paper presented at the American Association for Public Opinion Research Annual Meeting, Lancaster, Pennsylvania.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science, 37*, 941–964.

Krosnick, J. A., & Fabrigar, L. R. (forthcoming). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 70*, 29–44.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty and J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences*. Hillsdale, NJ: Erlbaum.

Krosnick, J. A., & Schuman, H. (1988). Attitude intensity, importance, and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology, 54*, 940–952.

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology, 65*, 1132–1151.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of no opinion response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66*, 371–403.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*, 395–409.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19*, 448–468.

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement, 33*, 545–563.

Kuncel, R. B. (1977). Ordering items by endorsement value and its effect upon text validity. *Educational and Psychological Measurement, 37*, 897–905.

Kunst-Wilson, W. R., & Zajonc, R. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558.

Laird, J. D. (1974). Self-attribution of emotion: The effects of expressive behavior on the quality of emotional experience. *Journal of Personality & Social Psychology, 29*, 475–486.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377–395.

Larkins, A.G., & Shaver, J. P. (1967). Matched-pair scoring technique used on a first-grade yes-no type economics achievement test. *Utah Academy of Science, Art, and Letters: Proceedings, 44*(I), 229–242.

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon and Schuster.

Leech, G. N. (1983). *Principles of pragmatics*. London, New York: Longman.

Lehmann, D. R., & Hulbert, J. (1972). Are three-point scales always good enough? *Journal of Marketing Research, 9*, 444–446.

Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology, 65*, 463–467.

Likert, R. (1932). A technique for the measurement of attitudes. *Archive of Psychology*, *140*, 44–53.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10–13.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality & Social Psychology*, *82*(1), 5–18.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality & Social Psychology, 47*, 1231–1243.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement: Vol. 3. Representation, axiomatization, and invariance*. San Diego, CA: Academic.

Maass, A., Salvi, D., Arcuri, L., & Semin, G. R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality & Social Psychology*, *57*(6), 981–993.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

Marsh, K. L., Johnson, B. T., & Scott-Sheldon, L. A. (2001). Heart versus reason in condom use: Implicit versus explicit attitudinal predictors of sexual behavior. *Zeitschrift Fuer Experimentelle Psychologie, 48*, 161–175.

Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality and Social Psychology, 51*, 493–504.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research, 10*, 316–318.

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research, 15*, 304–308.

Masters, J. R. (1974). The relationship between number of response categories and reliability of likert-type questionnaires. *Journal of Educational Measurement, 11*, 49–53.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657–674.

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology, 56*, 506–509.

Mathews, C. O. (1927). The effect of position of printed response words upon children's answers to questions in two-response types of tests. *Journal of Educational Psychology, 18*, 445–457.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly, 67*, 205–211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research, 20*, 60–103.

McClendon, M. J., & Alwin, D. F. (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods and Research, 21*, 438–464.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435–442.

McHugo, G., & Lanzetta, J. T. (1983). Methodological decisions in social psychophysiology. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 630–665). New York, NY: Guliford Press.

McKelvie, S. J. (1978). Graphic rating scales – How many categories? *British Journal of Psychology, 69*, 185–202.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S., & Frederiksen N. (1958). Ability, acquiescence, and "authoritarianism." *Psychological Reports, 4*, 687–697.

Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Michaelis, W., & Eysenck, H. J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in real-life motivation. *Journal of Genetic Psychology*, *118*, 223–234.

Milgram, S., Mann, L., & Harter, S. (1965). The lost-letter technique: A tool of social research. *Public Opinion Quarterly*, *29*, 437–438.

Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurement. *Journal of Abnormal and Social Psychology, 59*, 1–9.

Miller, W. E. (1982). *American national election study, 1980: Pre and post election surveys*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Mingay, D. J., & Greenwell, M. T. (1989). Memory bias and response-order effects. *Journal of Official Statistics, 5*, 253–263.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 x 2 index. *Social Psychology Quarterly, 54*(2), 127–145.

Mitchell, G., & Tetlock, P. (2006). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal*, *67*, 1023–1121.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469.

Morin, R. (1993, December 6–12). Ask and you might deceive: The wording of presidential approval questions might be producing skewed results. *The Washington Post National Weekly Edition*, p. 37.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723–739.

Murray, D. M., & Perry, C. L. (1987). The measurement of substance use among adolescents: When is the bogus pipeline method needed? *Addictive Behaviors*, *12*, 225–233.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58–88.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology—General*, *106*, 226–254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newcomb, T. E. (1943). *Personality and social change*. New York: Dryden Press.

Norman, D. A. (1973). Memory, knowledge, and the answering of questions. In R. L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola Symposium*. Washington, D.C.: Winston.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition*, *19*, 625–666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*, 166–180.

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*, 553–594.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*, 413–417.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*, 636–639.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test. *Journal of Personality and Social Psychology*, *86*, 653–667.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, Florida.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Ostrom, T. M., & Gannon, K. M. (1996). Exemplar generation: Assessing how respondents give meaning to rating scales. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 293–441). San Francisco, CA: Jossey-Bass.

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.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598–609.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. Wiggins (Eds.), *Personality assessment via Questionnaires: Current issues in theory and measurement*. New York, NY: Springer-Verlag.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightman (Eds.), *Measures of personality and social psychological attitudes. Volume 1 in Measures of social psychological attitudes series*. San Diego, CA: Academic Press.

Pavlos, A. J. (1972). Racial attitude and stereotype change with bogus pipeline paradigm. *Proceedings of the 80th Annual Convention of the American Psychological Association*, 7, 292–292.

Pavlos, A. J. (1973). Acute self-esteem effects on racial attitudes measured by rating scale and bogus pipeline. *Proceedings of the 81st Annual Convention of the American Psychological Association*, *8*, 165–166.

Payne, B. K., Cheng, S. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277–293.

Payne, J. D. (1971). The effects of reversing the order of verbal rating scales in a postal survey. *Journal of the Marketing Research Society, 14*, 30–44.

Payne, S. L. (1949/1950). Case study in question complexity. *Public Opinion Quarterly, 13*, 653–658.

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly, 14*, 687–696.

Petty, R. E., & Cacioppo, J. T. (1983). The role of bodily responses in attitude measurement and change. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 51–101). New York, NY: Guilford Press.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York, NY: Springer-Verlag.

Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength: Antecedents and consequences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E., Gatenby, J., Gore, J. C., Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738.

Poe, G. S., Seeman, I., McLaughlin, J., Mehl, E., & Dietz, M. (1988). Don't know boxes in factual questions in a mail questionnaire. *Public Opinion Quarterly, 52*, 212–222.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 55–85). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality & Social Psychology*, *61*(3), 380–391.

Quigley-Fernandez, B., & Tedeschi, J. T. (1978). The bogus pipeline as lie detector: Two validity studies. *Journal of Personality and Social Psychology*, *36*, 247–256.

Quinn, S. B., & Belson, W. A. (1969). *The effects of reversing the order of presentation of verbal rating scales in survey interviews*. London: Survey Research Centre.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513–532.

Rankin, R. E., & Campbell, D. T. (1955). Galvanic skin response to Negro and white experimenters. *Journal of Abnormal & Social Psychology*, *51*, 30–33.

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9*, 45–48.

Remmers, H. H., Marschat, L. E., Brown, A., & Chapman, I. (1923). An experimental study of the relative difficulty of true-false, multiple-choice, and incomplete-sentence types of examination questions. *Journal of Educational Psychology, 14*, 367–372.

Risko, E. F., Stolz, J. A., & Besner, D. (2005). Basic processes in reading: Is visual word recognition obligatory? *Psychonomic Bulletin and Review*, *12*, 119–124.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1999). *Measures of political attitudes*. San Diego, CA: Academic Press.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, *114*, 363–375.

Rohr, M., Degner, J., & Wentura, D. (2015). The "Emotion Misattribution" Procedure: Processing beyond good and bad under masked and unmasked presentation conditions. *Cognition and Emotion*, *29*, 196–219.

Rosenberg, N., Izard, C. E., & Hollander, E. P. (1955). Middle category response: Reliability and relationship to personality and intelligence variables. *Educational and Psychological Measurement, 15*, 281–290.

Rosenstone, S. J., Hansen, J. M., & Kinder, D. R. (1986). Measuring change in personal economic well-being. *Public Opinion Quarterly, 50*, 176–192.

Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: Attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of Personality & Social Psychology, 63*, 198–211.

Rothenberg, B. B. (1969). Conservation of number among four- and five-year-old children: Some methodological considerations. *Child Development, 40*, 383–406.
Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Eliminating the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology*, *62*, 84–98.
Rothermund, K., & Wentura, D. (2010). It´s brief, but is it better? An evaluation of the Brief Implicit Association Test (BIAT). *Experimental Psychology*, *57*, 233–237.
Rubin, H. K. (1940). A constant error in the Seashore test of pitch discrimination. Master's thesis. University of Wisconsin.
Ruch, G. M., & DeGraff, M. H. (1926). Corrections for chance and "guess" vs. "do not guess" instructions in multiple-response tests. *Journal of Educational Psychology, 17*, 368–375.
Rugg, D., & Cantril, H. (1944). The wording of questions. In H. Cantril (Ed.), *Gauging public opinion*. Princeton: Princeton University Press.
Rundquist, E. A., & Sletto, R. F. (1936). *Personality in the depression*. Minneapolis: University of Minnesota Press.
Salancik, G. R., & Conway, M. (1975). Attitude inferences from salient and relevant cognitive content about behavior. *Journal of Personality and Social Psychology, 32*, 829–840.
Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology, 59*, 614–622.
Saris, W., & Krosnick, J. A. (2000). The damaging effect of acquiescence response bias on answers to agree/disagree questions. Paper presented at the American Association for Public Opinion Research Annual Meeting, Portland, Oregon.
Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010, May). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61–79.
Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association, 84*, 402–413.
Scherpenzeel, A. (1995). Meta-analysis of a European comparative study. In W. Saris & A. Munnich (Eds.), *The multitrait-multimethod approach to evaluatemeasurement instruments*. Budapest, Hungary: Eotvos University Press.
Schlenker, B. R., & Weingold, M. F. (1989). Goals and the self-identification process: Constructing desires identities. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology* (pp. 243–290). Hillsdale, NJ: Lawrence Erlbaum Associates.
Schuman, H. (2008). *Method and meaning in polls and surveys*. Harvard University Press.
Schuman, H., & Converse, J. M. (1971). The effect of black and white interviewers on black responses. *Public Opinion Quarterly, 35*, 44–68.
Schuman, H., Ludwig, J., & Krosnick, J. A. (1986). The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly, 50*, 519–536.
Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*, 638–656.
Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgement* (pp. 217–245). Hillsdale, NJ: Erlbaum.
Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 433–465). New York: Guilford Press.
Schwarz, N., & Hippler, H. J. (1991). Response alternatives: The impact of their choice and presentation order. In P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 41–56). New York: Wiley.
Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research*. New York: Springer-Verlag.
Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570–582.
Schwarz, N., & Strack, F. (1991). Context effects in attitude surveys: Applying cognitive theory to social research. *European Review of Social Psychology, 2*, 31–50.
Schwarz, N., & Wyer, R.S. (1985). Effects of rank ordering stimuli on magnitude ratings of these and other stimuli. *Journal of Experimental Social Psychology, 21*, 30–46.
Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology*, *29*, 209–269.

Shaffer, J. W. (1963). A new acquiescence scale for the MMPI. *Journal of Clinical Psychology, 19*, 412–415.
Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.
Sherif, C. W., Sherif, M., & Nebergall, R. E. (1965). *Attitude and attitude change*. Philadelphia: W. B. Saunders.
Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. New Haven: Yale University Press.
Sherman, S.J., Castelli, J., & Hamilton, D.L. (2002). The spontaneous use of a group typology as an organizing principle in memory. *Journal of Personality & Social Psychology, 82*, 328–342.
Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.
Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. Journal of validation of the personality research form in Zimbabwe. *International Journal of Psychology*, *25*, 1–12.
Sigelman, C. K., & Budd, E. C. (1986). Pictures as an aid in questioning mentally retarded persons. *Rehabilitation Counseling Bulletin, 29*, 173–181.
Simon, J. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Stimulus-response compatibility: An integrated perspective*. *Advances in psychology* (Vol. 65, pp. 31–86). Amsterdam: North-Holland.
Singal, J. (2017, January 11). Psychology's favorite tool for measuring racism isn't up to the job. Almost two decades after its introduction, the implicit association test has failed to deliver on its lofty promises. *New York Magazine*.
Smith, E. R., Fazio, R. H., & Cejka, M. A. (1996). Accessible attitudes influence categorization of multiply categorizable objects. *Journal of Personality & Social Psychology*, *71*(5), 888–898.
Smith, T. W. (1994a). A comparison of two confidence scales. GSS Methodological Report No. 80, National Opinion Research Center, Chicago, IL.
Smith, T. W. (1994b). A comparison of two governmental spending scales. GSS Methodological Report No. 81, National Opinion Research Center, Chicago, IL.
Smith, T. W., & Peterson, B. L. (1985, August). The impact of number of response categories on inter-item associations: Experimental and simulated results. Paper presented at the American Sociological Association Meeting, Washington, DC.
Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly, 73*(2), 325–337.
Snyder, M. (1979). Self-monitoring processes. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 86–128). New York: Academic Press.
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.
Srinivasan, V., & Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science, 8*, 205–230.
Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*, 283–294.
Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*(3), 165–179.
Stember, H., & Hyman, H. (1949/1950). How interviewer effects operate through question form. *International Journal of Opinion and Attitude Research, 3*, 493–512.
Stieger, R. S., Göritz, A. S., Hergovich, A., & Voracek, M. (2011). Intentional faking of the single category implicit association test and the implicit association test. *Psychological Reports*, *109*, 219–230.
Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 249–275). Hillsdale, NJ: Lawrence Erlbaum Associates.
Strack, F., & Martin, L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H. J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123–148). New York: Springer Verlag.
Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*, 429–442.
Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality & Social Psychology, 54*, 768–777.
Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Sussman, B. (1978). President's popularity in the polls is distorted by rating questions. *Washington Post*.
Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: A meta-analysis of the racial attitude-behavior literature shows that emotions, not beliefs, best predict discrimination. *Social Justice Research*, *21*, 263–296.
Tamulonis, V. (1947). The effects of question variations in public opinion surveys. Masters thesis. Denver: University of Denver.
Teige-Mocigemba, S., & Klauer, K. C. (2013). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition and Emotion*, *27*, 632–657.
Teige-Mocigemba, S. Penzl, B., Becker, M., Henn, L., & Klauer, K. C. (2016). Controlling the "uncontrollable": Faking effects on the Affect Misattribution Procedure. *Cognition and Emotion*, *30*, 1470–1484.
Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 289–338). New York, NY: Academic Press.
Tesser, A., Whitaker, D., Martin, L., & Ward, D. (1998). Attitude heritability, attitude change and physiological responsivity. *Personality and Individual Differences, 24*, 89–96.
Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, *45*, 74–83.
Thomas, W. I., & Znaniecki, F. (1918). *The Polish peasant in Europe and America* (Vol. 1). Boston, MA: Badger.
Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34*, 251–259.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.
Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology*, *26*, 249–269.
Tourangeau, R, & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299–314.
Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Carryover effects in attitude surveys. *Public Opinion Quarterly, 53*, 495–524.
Trott, D. M., & Jackson, D. N. (1967). An experimental analysis of acquiescence. *Journal of Experimental Research in Personality, 2*, 278–288.
Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *8*(4), 336–342.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.
Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public Opinion Quarterly, 37*, 373–387.
Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, *71*, 941–959.
Visser, P. S., Krosnick, J. A., Marquette, J. F., & Curtin, M. F. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. L. Lavrakas, & M. Traugott (Eds.), *Election polls, the news media, and democracy*. New York: Chatham House.
von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1997). The linguistic intergroup bias as an implicit indicator of prejudice. *Journal of Experimental Social Psychology*, *33*(5), 490–509.
Waenke, M., Bohner, G., & Jurkowitsch, A. (1997). There are many reasons to drive a BMW: Does imagined ease of argument generation influence attitudes? *Journal of Consumer Research, 24*, 170–177.
Wallaert, M., Ward, A., & Mann, T. (2010). Explicit control of implicit responses: Simple directives can alter IAT performance. *Social Psychology*, *41*, 152–157.
Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.
Warr, P., Barter, J., & Brownridge, G. (1983). On the interdependence of positive and negative affect. *Journal of Personality and Social Psychology, 44*, 644–651.
Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133–142.
Watson, D. (1988). The vicissitudes of mood measurement: Effects of varying descriptors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology, 55*, 128–141.
Watson, D. R., & Crawford, C. C. (1930). Four types of tests. *The High School Teacher, 6*, 282–283.
Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Non-reactive research in the social sciences*. Chicago, IL: Rand McNally.
Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*, 341–356.
Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology, 23*, 230–249.

Wedell, D. H., Parducci, A., & Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors. *Journal of Personality and Social Psychology, 58*, 319–329.

Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–487). Hillsdale, NJ: Erlbaum.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology, 68*, 36–51.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141–208). San Diego, CA: Academic Press.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247.

Weitz, S. (1972). Attitude, voice, and behavior: A repressed affect model of interracial interaction. *Journal of Personality and Social Psychology*, *24*, 14–21.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956–972.

Wesman, A. G. (1946). The usefulness of correctly spelled words in a spelling test. *Journal of Educational Psychology, 37*, 242–246.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*, 41–78.

Wigboldus, D. H. J., Holland, R. W., & van Knippenberg, A. (2004). Single target implicit associations. Unpublished manuscript.

Wilson, T. D., & Hodges, S. D. (1992). Attitudes as temporary constructions. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 37–65). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*, 101–126. doi:10.1037//0033–295X.107.1.101

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*, 555–561.

Wiseman, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, *36*, 105–108.

Wittenbrink, B. (2004). Ordinary forms of prejudice. *Psychological Inquiry*, *15*, 306–310.

Wittenbrink, B. (2007). Measuring attitudes through priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 17–58). New York, NY: Guilford Press.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*, 262–274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001a). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality & Social Psychology*, *81*(5), 815–827.

Wittenbrink, B., Judd, C. M., & Park, B. (2001b). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology*, *37*(3), 244–252.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.

Ying, Y. (1989). Nonresponse on the center for epidemiological studies-depression scale in Chinese Americans. *International Journal of Social Psychiatry, 35*, 156–163.

Yzerbyt, V. Y., & Leyens, J. (1991). Requesting information to form an impression: The influence of valence and confirmatory status. *Journal of Experimental Social Psychology, 27*, 337–356.

Zajonc, R. B. (1960). The process of cognitive tuning and communication. *Journal of Applied Social Psychology, 61*, 159–167.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science, 36*, 579–616.