In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes. San Diego: Academic Press, 1991.

CHAPTER 2

Measurement and Control of Response Bias

Delroy L. Paulhus

A *response bias* is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (i.e., what the items were designed to measure). For example, a respondent might choose the option that is most extreme or most socially desirable. A response bias might be a *response set,* that is, a temporary reaction to a situational demand, for example, time pressure or expected public disclosure. Alternatively, a bias may be induced by context effects such as the item format or the nature of previous items in the questionnaire (for a review, see Tourangeau & Rasinski, 1988). To the extent that an individual displays the bias consistently across time and situations, the bias is said to be a *response style* (Jackson & Messick, 1958; Wiggins, 1973).

Response biases continue to be a disturbing issue in psychological assessment, particularly with self-report measures such as those in the present volume. People's reports of their own traits, attitudes, and behavior may involve systematic biases that obscure measurement of content variables. For example, there is evidence that standard self-report methodologies distort the reporting of racist attitudes (Sigall & Page, 1971), nonnormative sexual attitudes (Knudson, Pope, & Irish, 1967), desirable behaviors (Phillips & Clancy, 1972), deviant behaviors (Clark & Tifft, 1966), and abortion (Wiseman, 1972). Another repercussion of response bias is that an assessment instrument may confound content with style. That is, the instrument may simultaneously assess the individual's response style as well as the content variable. If so, every observed correlation with this instrument will be open to (at least) two possible explanations.

Among the response biases cited in the literature are deviant responding (Berg, 1967), careless responding (Meehl & Hathaway, 1946), consistent responding (Dillehay & Jernigan, 1970), and omitting items (Cronbach, 1946). In this chapter, however, attention will be restricted to the three most prominent response biases: (1) socially desirable responding (including lying, faking bad, etc.), (2) acquiescence (tendency to agree), and (3) extremity bias (tendency to use extreme ratings).

Socially Desirable Responding

The most frequently studied response bias is socially desirable responding (SDR), the tendency to give answers that make the respondent look good. Over 50 years ago,

psychometricians had already raised the issue of SDR effects on the validity of questionnaires (e.g., Bernreuter, 1933; Vernon, 1934). Ten years later, Meehl and Hathaway (1946) were able to cite eight measures specifically developed to index SDR in self-reports.[1] Since the 1950s, SDR has been a prominent concern in measuring personality (e.g., Edwards, 1953), psychopathology (e.g., Gough, 1947; McKinley, Hathaway, & Meehl, 1948), attitudes (Lenski & Leggett, 1960), and self-reports of sensitive behavior (Goode & Hart, 1952).

The first section below reviews methods developed to control SDR in each of these realms. Where SDR cannot be controlled, it might still be measured: Accordingly, the second section below reviews the best available measures of SDR.

Controlling the Influence of SDR

Available methods for controlling SDR are varied and often complex [for reviews, see DeMaio (1984), Nederhof (1985), or Paulhus (1981)]. Only a brief summary can be rendered here. Four types of methods may be distinguished: rational, factor analytic, covariate, and demand reduction methods.

1. Rational techniques are control features that are built into the self-report instrument. Their purpose is to prevent the subject from responding in a socially desirable fashion. For example, one may use a forced-choice format in which the two statements are equated for social desirability. If single statements are used, they might be restricted to those that are roughly neutral with respect to social desirability. Alternatively, one could select statements with high content saturation, that is, statements in which the relative influence of content over desirability is high (see Jackson, 1967).

2. Factor analytic techniques may be applied during test construction if the procedure involves choosing the highest loading items in a factor analysis. If one principal component appears to represent SDR, it may be deleted before the factors are rotated (Paulhus, 1981). Even better, if a measure of SDR is administered along with the content items, then the first factor may be rotated to the SDR measure (e.g., Morf & Jackson, 1972; Norman, 1969). In both techniques the highest loading items on the remaining factors necessarily tap individual differences above and beyond the effects of SDR. Hence, these items may be used to construct SDR-free content measures.

3. In covariate techniques, no attempt is made to prevent respondents from answering in a desirable fashion. Instead, a measure of SDR, for example, one of the measures provided at the end of this chapter, is administered along with the content measures. SDR may then be partialed out of correlations between two content scales to control for spurious correlation.

Alternatively, to improve the validity of individual scores, the raw score may be adjusted by an amount commensurate with the contamination by SDR. This adjustment may be effected by regressing the content score on SDR: The residual then represents the content score corrected for SDR (e.g., Norman, 1967). This procedure is systematized in scoring the MMPI, in which certain clinical scale scores are adjusted using the K scale as a measure of SDR (McKinley *et al.*, 1948). A similar procedure is now used in adjusting several scales of the 16PF (Karson & O'Dell, 1976).

4. A wide-ranging form of control, demand reduction, includes those methods that

[1]Precedence should be granted to Hartshorne and May (1928), who in the twenties had already developed a lie scale to directly assess dishonesty in school children.

reduce the situational press for desirable responding. Perhaps the most obvious strategy is to assure respondents of anonymity. In classroom administrations, perceived anonymity is increased by (1) physically separating respondents (especially acquaintances), (2) insisting that they put no identifying marks on the questionnaire (Becker, 1976), and (3) telling them beforehand that they will seal the completed questionnaire in an envelope and drop it in a box on the way out. Presumably because of perceived anonymity, mail surveys appear to be less susceptible to SDR (Wiseman, 1972). Computerized assessments show lower SDR than face-to-face interviews, but higher SDR than paper-and-pencil tests (Martin & Nagao, 1989; Davis & Cowles, 1989; Lautenschlager & Flaherty, in press).

If it is necessary to match responses across two administrations, respondents could be asked to give their birthdates, or to use a consistent pseudonym for all administrations. Despite all such precautions, at least some subjects may still suspect that their identity can be determined by the experimenters.

A more aggressive technique, the bogus pipeline (Jones & Sigall, 1971), is essentially a pseudo lie detector. Respondents are hooked up to electronic equipment that the operator claims can assess their attitudes directly through physiological measures. The equipment is said to be, in effect, "a pipeline to the soul." As a putative test of their own self-insight, subjects are asked to guess the machine's reading for each attitude question. The rationale is that subjects want to avoid being embarrassed by the machine: Hence, their guesses should be less contaminated with SDR than are ordinary self-reports.

The efficacy of the technique is documented by its ability to increase admissions of such undesirable behavior as (a) racist attitudes (Sigall & Page, 1971), (b) sexist attitudes (Faranda, Kaminski, & Giza, 1979), (c) inconsistent attitudes (Paulhus, 1982), (d) dislike for a handicapped confederate (Sigall & Page, 1972), and (e) having prior knowledge of test answers (Quigley-Fernandez & Tedeschi, 1978).

Some related techniques are milder versions of the bogus pipeline. For example, before respondents complete a questionnaire they are warned that the measure contains methods for detecting faking. In a complex variant of this technique, respondents are given an especially transparent lie scale, which is then scored, and they are advised whether their score indicates faking (Montag & Comrey, 1982).

A technique useful for face-to-face interviews is the Randomized Response Method (Greenberg, Abdula, Simmons, & Horvitz, 1969; Warner, 1965). The desirability-loaded question (e.g., Have you ever had an abortion?) is posed along with an innocuous question (Are your mother's eyes blue?). The respondent is instructed to flip a coin and (without telling the interviewer how the flip came out) answer question 1 if heads; question 2, if tails. Since the interviewer is not aware which question is being answered, the respondent is under less pressure to respond desirably. Nonetheless, the abortion rate in the sample can ultimately be estimated from three group statistics: the observed proportion of "yes" responses, the assumption that .50 of the respondents answered the abortion question, and the known rate of blue eyes in the population. The bulk of the evidence suggests that the randomized response technique does increase reports of sensitive behaviors, although not quite up to their true rates (Dawes & Smith, 1985).

Psychological assessment through biographical data is a unique approach that tends to reduce SDR demand (Shaffer, Saunders, & Owens, 1986). The rationale is that, when reporting on concrete, verifiable facts, subjects are less tempted to dissimulate.

Finally, some researchers have tried to reduce SDR by employing proxy subjects: Instead of the target person, a close acquaintance is questioned about the target's behavior. According to evidence collected by Sudman and Bradburn (1974), this technique is satisfactory for measuring publicly observable types of behavior, but not for attitudes. The

evidence for its efficacy in measuring personality is mixed (Kane & Lawler, 1978; McCrae, 1982).

5. The last category, stress minimization, refers to basic guidelines for reducing tension during the test administration. Among the factors known to increase desirable responding are: (a) speed instructions (Sutherland & Spilka, 1964), (b) emotional arousal (Paulhus & Levitt, 1987), and distraction (Paulhus, Graf, & Van Selst, 1989). Hence, these factors should all be minimized.

Measurement of SDR

There are a number of reasons why a researcher might want a direct measure of SDR such as those included in this chapter. The most common usage is for supporting the discriminant validity of a content instrument. To ensure that the content instrument is not confounded with SDR, the researcher would administer both measures to the same sample, hoping to see a low intercorrelation. An SDR measure would also be necessary to use the covariate and target rotation techniques described above.

In some cases, rather than correcting a biased score, one might prefer to discard an individual's data if an accompanying SDR measure detected a high degree of SDR. Optimal cutoff scores may be derived for detecting faking good and faking bad (e.g., Helmes & Holden, 1986; Karson & O'Dell, 1976; Lanning, 1989).

Such cutoff scores may then be useful in evaluating the fakeability of content instruments. A content scale is resistant to faking if, under fake-good instructions, the SDR scale exceeds the cutoff point but the content scale does not change. Note that demonstrating that one can fake good on an instrument does not prove that it is confounded with SDR under ordinary conditions (Furnham, 1986). Moreover, faking good can be idiosyncratic (Lautenschlager, 1986).

Situational demands can also be assessed with SDR measures. For example, organizational researchers have used such measures to determine the extent to which various work environments encourage SDR (see Zerbe & Paulhus, 1987, for a review). SDR measures may also assist in determining the best interviewer characteristics (Weiss, 1968) or test conditions (e.g., Martin & Nagao, 1989; Paulhus, 1984; Schriesheim, 1979) to promote disclosure. Computerized testing, for example, has been evaluated with several measures of SDR (Davis & Cowles, 1989; Lautenschlager & Flaherty, in press).

In practice, the primary use for SDR measures has been to assess consistent individual differences, that is, response styles. There is a longstanding concern that response styles interfere with accurate assessment of content variables (e.g., Edwards, 1953; Goode & Hart, 1952; Gough, 1947; McKinley *et al.*, 1948). Such styles have also been claimed to reflect deeper psychological constructs of interest in their own right (Block, 1965; Damarin & Messick, 1965; Sweetland & Quay, 1953). The classic example is Crowne and Marlowe's work (1964) on the need for approval construct, which emerged from studies on their 1960 measure of SDR (Crowne & Marlowe, 1960). Other personality constructs postulated to underlie response styles include repression–sensitization (Byrne, 1964), censure avoidance (Allaman, Joyce & Crandall, 1972; Millham & Jacobson, 1978), and self-deception (Paulhus, 1984; Sackeim & Gur, 1978). Measures of several of these constructs are included at the end of this chapter.

The widespread concern about SDR in test responses is reflected in the fact that major personality batteries invariably include an SDR measure. The MMPI includes two such scales: the Lie scale to detect blatant dissimulation and the K scale to tap more subtle

distortions (McKinley *et al.*, 1948). The Eysenck Personality Inventory (Eysenck & Eysenck, 1964) and the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) both contain a rationally developed lie scale. Comrey (1980) includes the Response Bias scale in the Comrey Personality Scales. The Personality Research Form (Jackson, 1967) contains the Desirability scale assembled from diverse items of extreme desirability. The Differential Personality Questionnaire (Tellegen, 1982) contains the Unlikely Virtues scale as well as a Desirability Inconsistency scale. The California Psychological Inventory contains the Good Impression and Well-Being scales (Gough, 1987). Measures of both faking good and faking bad have been developed for the 16-PF (Winder, O'Dell, & Karson, 1975).

With the exception of the MMPI validity scales, little direct research has been conducted on these commercially published SDR measures. One reason is that these instruments are often tailored to be similar in format and content to the inventory as a whole and hence may have little application outside the inventory. Because of the limitations in empirical evidence, such measures are not presented in detail here; rather, the reader is referred to the manuals cited above for each of the inventories.

Varieties of SDR

A disturbing feature of SDR measures is the low intercorrelation among several of the more well-known instruments, for example, Edwards's SD scale, the Marlowe–Crowne scale and Wiggins's Sd scale. The frustration of nonspecialists is epitomized in the title of one article: "Will the real social desirability scale please stand up?" (Strosahl, Linehan, & Chiles, 1984). Factor analyses of SDR instruments have consistently revealed two primary factors (Borkenau & Ostendorf, 1989; Edwards & Walsh, 1964; Jackson & Messick, 1962; Paulhus, 1984; Wiggins, 1964). One cluster is associated with *Alpha,* the general anxiety factor of the MMPI (Block, 1965). The second cluster is associated with the another MMPI factor called *Gamma* (Wiggins, 1964), which is linked to agreeableness and traditionalism. Paulhus (1984, 1986) provided evidence that these two SDR factors represent (a) self-deceptive positivity (an honest but overly positive self-presentation) and (b) impression management (self-presentation tailored to an audience). This distinction was outlined many years before by Damarin and Messick (1965). As depicted in Fig. 1, the gamut of SDR measures may usefully be characterized in terms of their relative weighting of self-deception and impression management. The Desirability scale of the PRF and Edwards's SD scale load primarily on the self-deception factor; Wiggins's Sd scale and Eysenck's Lie scale load primarily on the impression management factor. The Marlowe–Crowne scale loads on both factors, although more so on impression management.

The term "impression management" was chosen to represent one traditional view of SDR: that some subjects are purposefully tailoring their answers to create the most positive social image. Of the many impressions that one may try to present, this factor represents only one: a socially conventional, dependable persona. The label "impression management" is preferable to "lying," which is an overly harsh and sweeping indictment. After all, such individuals may misrepresent themselves only to avoid social disapproval (Crowne, 1979). Whatever the label,[2] this tendency will vary according to situational

[2]This factor has also been given such diverse interpretations as moralistic hypocrisy (Cattell, Pierson, & Finkbeiner, 1976), interpersonal sensitivity (Holden & Fekken, 1989), defensiveness (Weinberger, Schwartz, & Davidson, 1979), extraverted adjustment (McCrae & Costa, 1983), and test-taking intelligence (P. Borkenau, personal communication).

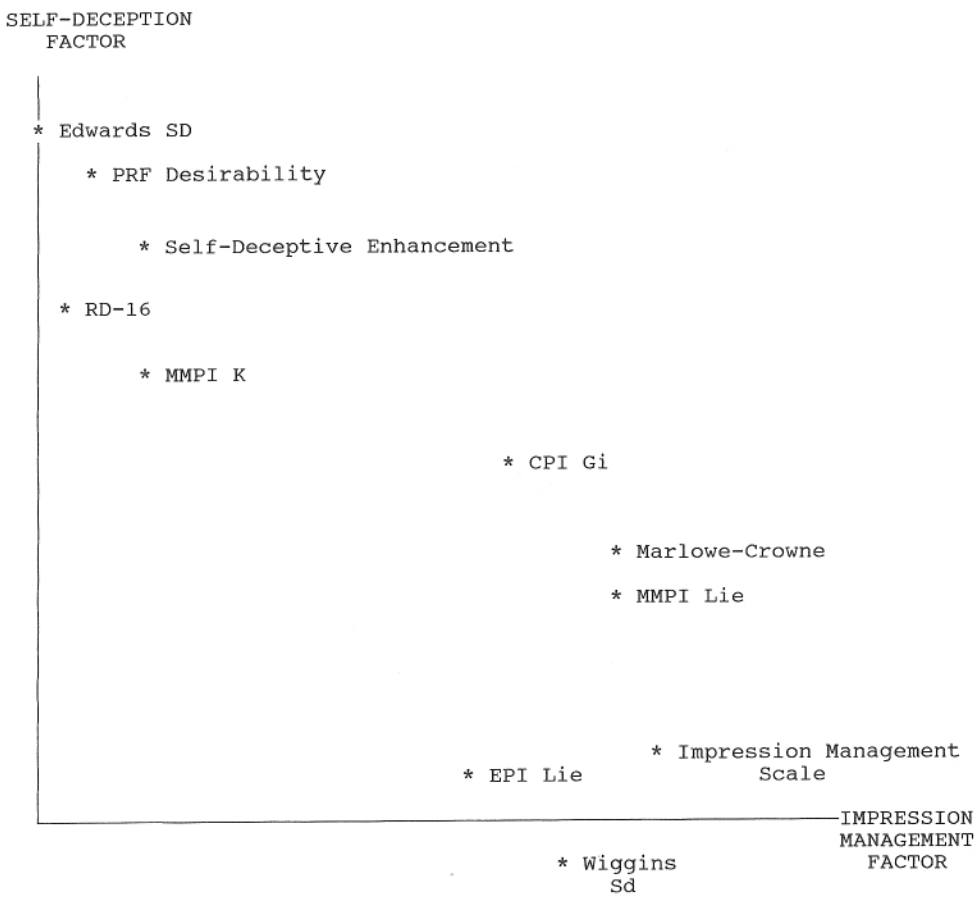

Fig. 1 Typical factor loadings of SDR measures.

demands and transient motives and that variation may obscure the validity of the respondent's self-reports.

The concept of self-deceptive positivity, on the other hand, appears to be intrinsically linked to such personality constructs as adjustment (Taylor & Brown, 1988), optimism (Scheier & Carver, 1985), self-esteem (M. Rosenberg, personal communication), and a sense of general capability (Holden & Fekken, 1989). These conceptual links are supported by significant correlations between measures of self-deception and measures of mental health (Linden, Paulhus, & Dobson, 1986; Sackeim & Gur, 1979), self-esteem (Paulhus & Reid, in press; Winters & Neale, 1985), and various cognitive biases (Paulhus, 1988; Paulhus & Reid, in press). The label "self-deceptive" was chosen in reference to the verifiable distortion by high scorers on certain forms of self-information (Paulhus, 1988).

This two-factor distinction helps resolve several issues in the SDR literature. Research typically shows a positive relation between SDR and adjustment, whereas traditional theories (as well as common sense) presume a negative relation. Self-deceptive SDR bears a strong positive relation with adjustment whereas impression management bears little relation. Thus the relation with adjustment depends on the type of SDR.

The second issue clarified by the self-deception versus impression management distinction is what should be done about the SDR that pervades such central personality variables as perceived control, social dominance, and adjustment. Norman (1967, 1990) argued that removing a general index of SDR from personality ratings clarifies the content dimensions. However, most relevant studies have shown that controlling SDR, if anything, actually *reduces* the predictive validity of content measures (Borkenau & Amelang, 1985; Kozma & Stones, 1988; McCrae, Costa, Dahlstrom, Barefoot, Siegler, & Williams, 1989; McCrae & Costa, 1983; Ruch & Ruch, 1967). Indeed, it now appears that controlling any SDR measure that taps self-deception (e.g., the Edwards scale, the Borkenau–Ostendorf SD scale, the K scale, and to a lesser degree, the Marlowe–Crowne scale) will lower the predictive validity of measures involving a self-deceptive positivity (e.g., anxiety, achievement motivation, dominance, well-being, perceived control, self-esteem). This form of SDR is inextricably linked to content variance and should not be controlled.

In contrast, impression management should be controlled under certain circumstances (Paulhus, 1986), namely, when impression management is conceptually independent of the trait being assessed but still contributes to the self-report scores of that trait. In personnel selection, for example, transient motives ("I'll say anything to get this job because my business just burned down") satisfy those criteria. Many of the techniques listed earlier were designed to control such conscious impression management.

The ideal procedure would involve establishing the distribution of impression management scores in the specific job-selection situation. In addition, the distribution of subjects asked to fake good would be established. If the scores of faking subjects sat several standard deviations above the mean of the no-fake group, then an efficient cutoff point would be easy to specify.

Finally, a set of measures with uncertain status must be mentioned—the so-called honesty or integrity scales. These are questionnaires administered in industrial–organizational settings to determine whether employees or job candidates can be trusted. They fall into two groups. One type contains direct questions about the respondent's integrity, of the form: "Are you now, or have you ever been, a crook?" The second type targets more general traits, such as conscientiousness and impulsiveness, that bear on employee trustworthiness (e.g., Hogan & Hogan, 1989).

Although widely used in business, the direct measures have had a poor reputation among psychologists (e.g., Sackett & Harris, 1984). This reputation may be due in part to the difficulty in obtaining the measures for research purposes. In addition, some honesty scales show positive correlations with lie scales. Thus the same individuals could be labeled honest or dishonest depending on whether honesty scales or lie scales are used. Nonetheless, the more recent evidence about the validity of certain honesty scales is more encouraging (Cunningham, 1989; Sackett, Burris, & Callahan, 1989).

Guidelines

The comments above suggest several practical implications for researchers. Caution must be exercised in interpreting reports of high correlations between self-report instruments and SDR: Such relations must not blithely be assumed to represent contamination and, therefore, deficits in the instruments. Indeed, the SDR component may be a legitimate aspect of the construct being measured.

If individual differences in SDR are evidenced in one's data, one must carefully consider why this has occurred. In a personnel selection situation, for example, a correlation between self-reported motivation and a measure of impression management has a number of plausible interpretations. A stylistic interpretation would suggest that chronic

impression managers are faking high motivation. Alternatively, the nature of the position (e.g., public relations) may be such that chronic impression managers would continue to be motivated and are, therefore, ideal candidates. A third interpretation is that the observed relation is due, not to stylistic impression management, but to a temporary response set: By chance, some respondents are temporarily motivated and are, therefore, presenting a good impression that is unlikely to predict future behavior.

Measures Reviewed Here

In selecting an SDR measure from the eight provided below, the researcher must consider which form of SDR is relevant and select an appropriate measure. The eight measures are

1. Edwards (1957) Social Desirability Scale,
2. Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960),
3. MMPI Lie Scale (Hathaway & McKinley, 1951),
4. MMPI K Scale (Meehl & Hathaway, 1946),
5. Balanced Inventory of Desirable Responding (Paulhus, 1984, 1988),
6. RD-16 (Schuessler, Hittle, & Cardascia, 1978), and
7. Children's Social Desirability Scale (Crandall, Crandall, & Katkovsky, 1965).

It is difficult to order the measures in terms of their value because they each have specialized applications. Instead, they will be presented in descending order of the amount of research published on the measure. No implication should be drawn that SDR measures not included here are deficient. There is simply less information available on other measures (except for the Eysenck Lie scale, which is reviewed by Furnham [1986]).[3]

The Edwards Social Desirability (SD) scale contains 39 items from the MMPI that have extremely high or low desirability ratings. The scale falls squarely on the first factor of SDR: hence it correlates highly with many standard measures of adjustment (e.g., anxiety, self-esteem, depression) and personality (e.g., perceived control, assertiveness). A high correlation with the SD does not invalidate an individual difference measure but suggests that the measure may involve some positivistic bias.

The Marlowe–Crowne (MC) scale contains 33 True–False items about behaviors that are desirable but rare or undesirable but common. The behaviors concern everyday events, not psychopathology. The scale loads on both factors of SDR but more highly on the impression management factor. The MC scale generally correlates less highly than the SD with measures of adjustment.

The MMPI Lie scale contains 15 statements about attitudes and practices that are socially undesirable, but common. Saying "false" to 8 or more is considered evidence that the respondent is faking good on the MMPI. The measure falls toward the second factor of SDR, highly correlated with other lie scales and the Marlowe–Crowne scale. The scale can detect faking good only in naive test-takers.

The MMPI K scale contains 30 items designed to identify abnormal persons whose MMPI scores appear normal. In contrast to the MMPI Lie scale, it is a more subtle

[3]Nor is there sufficient information concerning two new SDR measures: (1) the Borkenau–Ostendorf Social Desirability scale (Borkenau & Ostendorf, 1989), and (2) the Self-Presentation Scale (Roth, Harris, & Snyder, 1988). A promising technique that requires standardization is Phillips and Clancy's (1972) overclaiming index.

measure of SDR. Its status is rather complex, given that it is said to index psychological health in normal samples, but defensiveness in maladjusted samples.

The CPI Good Impression scale contains 40 items designed to measure faking good. The high scorer wants to present the impression of being dependable, cooperative, and moral. A good deal of validity data are available for this measure (Gough, 1987; Tellegen, 1982).

The Balanced Inventory of Desirable Responding (BIDR) contains separate measures of impression management (audience-driven self-presentation) and self-deceptive enhancement (an honest positivistic bias). The sum of the two measures correlates highly with the MC scale.

The RD-16 instrument was specifically designed to detect SDR in surveys of attitudes and opinions on the general population. An impressive set of norms is available from a national probability sample. The items were screened to preclude differences across such subgroups as race and education.

The Children's Social Desirability Scale was designed for children from grades 6–12. The item content largely follows that of the Marlowe–Crowne scale but is worded in children's language. Additional items involve child-specific content.

Edwards Social Desirability Scale (SD)

(Edwards, 1957)

Variable

Edwards (1957) described the SD as measuring "the tendency to give socially desirable responses in self-description" (p. 35), more specifically, an individual's characteristic level of self-presentation without special instructions or motivation to do so (p. 230).

Description

Edwards (1957) asked ten judges to rate whether "True" or "False" was the most desirable response to 79 items assembled from the K, F, and Lie scales of the MMPI. The 39 items on which the judges unanimously agreed formed the SD. Most of the items (30 of 39) are keyed negatively. As on the MMPI, respondents must answer "True" or "False," with one point added for each response that matches the key. Hence, possible scores range from 0 to 39, higher scores indicating more socially desirable responding. Given their source, it is not surprising that item content is heavily laden with references to psychological distress.

Samples

Edwards (1957) reported means of 28.6 (s.d. = 6.5) and 27.1 (s.d. = 6.5) for males and females in a sample of 192 college students. Edwards and Walsh (1964) found a mean of 28.8 in a sample of 130 paid students. More recently, Paulhus (1984) reported means of 28.4 (s.d. = 5.5) and 30.3 (s.d. = 4.9) for students in anonymous (n = 60) and public disclosure conditions (n = 40), respectively. In a sample of 503 students, Tanaka–Matsumi and Kameoka (1986) found means of 26.7 (s.d. = 6.3) and 20.9 (s.d. = 5.3) in normal and depressed samples, respectively.

Reliability

Internal Consistency

Alpha coefficients range from .83 to .87 in the samples reported above.

Test–Retest

Test–retest reliabilities of .66 (males) and .68 (females) after 2 weeks were reported by Rorer and Goldberg (1965).

Validity

Convergent

The ESD is robust in that it correlates highly with scales from a variety of content areas if they too were assembled from items with extreme desirability ratings (Edwards, 1970; Jackson & Messick, 1962). For example, the SD correlates .71 with the Desirability scale of Jackson's (1967) PRF, which comprises items of extreme desirability chosen from diverse personality domains (Holden & Fekken, 1989).

Discriminant

Early critics alleged that, because 22 of the 39 items overlapped with Taylor's (1953) anxiety scale, the SD was simply another anxiety measure. Similarly, Crowne and Marlowe (1960) complained that the scale was intrinsically confounded with psychopathology because many of the items referred to psychological distress. Edwards and Walsh (1964) responded by showing that the pattern of correlations with other measures was unchanged when the psychopathology items were replaced.

Edwards distinguished the construct measured by the SD from tendencies to lie deliberately as measured by impression management scales (Edwards, 1957, 1970). This conceptual distinction has been clearly sustained by the data (Edwards & Walsh, 1964; Paulhus, 1984; Wiggins, 1964).

A major controversy was stirred by reports of a very high correlation between SD and the first factor of the MMPI (e.g., Jackson & Messick, 1962). To some observers this correlation suggested that the MMPI assesses SDR instead of psychopathology. Block (1965), however, argued forcefully that the SD reflects a substantive trait, namely, ego resiliency. Although most commentators agree that the SD reflects a more general disposition, the precise nature of the disposition remains moot (Paulhus, 1986).

Location

Because the items are taken directly from the MMPI (a copyrighted instrument), only a few sample items can be provided. However, the 39 MMPI booklet numbers are listed so that the full scale may be assembled by the reader. The MMPI is available from University of Minnesota Press, Minneapolis, Minnesota 55455.

Comments

Despite his many reports on the SD, Edwards has said very little about the nature of the underlying construct. Even in his most recent comment (Edwards *et al.*, 1988), he has

adhered closely to an operational definition: the tendency to respond desirably to the sort of item in his scale. Because the construct is not well defined, it is difficult to marshall evidence for it. As noted in the literature review, the SD falls clearly on the first factor of SDR, indicating an honest form of positivistic bias.

Edwards's original implication that a high correlation with SD invalidates a self-report measure is no longer tenable. In fact, controlling for SD may actually reduce the validity of adjustment-related measures (e.g., Kozma & Stones, 1988; McCrae, 1986).

Fortunately, Edwards's colleagues (L. K. Edwards & Clark, 1987) have finally published his alternative version (Edwards, 1963)—one comprising nonpsychopathology items. The scale appears to perform similarly to the original, thereby challenging certain critiques going back 30 years (Edwards, Edwards, & Clark, 1988). Nevertheless, the new version has yet to be subjected to scrutiny by other researchers.

Edwards Social Desirability Scale

Sample Items

True	False	1. I am happy most of the time. (T)
True	False	2. My hands and feet are usually warm enough. (T)
True	False	3. No one cares much what happens to you. (F)
True	False	4. I sometimes feel that I am about to go to pieces. (F)

Complete Scale

The MMPI booklet numbers for the 9 items keyed "True" are as follows: 7, 18, 54, 107, 163, 169, 257, 371, 528. The 30 items keyed "False" are 32, 40, 42, 43, 138, 148, 156, 158, 171, 186, 218, 241, 245, 247, 252, 263, 267, 269, 286, 301, 321, 335, 337, 352, 383, 424, 431, 439, 549, 555.

Marlowe–Crowne Social Desirability Scale (MCSD)

(Crowne and Marlowe, 1960)

Variable

Although Crowne and Marlowe (1960) originally constructed the MCSD to be a measure of SDR in self-reports, their subsequent research on the construct convinced them that the scale was tapping a more general motive: They dubbed it *need for approval* (Crowne & Marlowe, 1964).[4] In the most recent statement, Crowne (1979) refined the concept to be an avoidance of disapproval.

Description

Crowne and Marlowe (1960) set out to build an SDR measure that improved upon the Edwards scale. Noting that Edwards's items were largely pathological in content, they

[4]Although Crowne was the senior author on both reports, the scale itself was labeled the Marlowe–Crowne scale, presumably to balance the credit.

focused instead on ordinary personal and interpersonal behaviors. Fifty such items were assembled and reduced to 33 by item analyses and ratings of experienced judges. The correlations with MMPI scales were still sizable, but not as high as those shown by the Edwards scale (e.g., Katkin, 1964).

The 33 items describe either (a) desirable but uncommon behaviors (e.g., admitting mistakes) or (b) undesirable but common behaviors (e.g., gossiping). Respondents are asked to respond "True" or "False" to 18 items keyed in the true direction and 15 in the false direction. Hence, scores range from 0 to 33, with higher scores representing higher need for approval.

Samples

Crowne and Marlowe (1964) reported a mean of 15.5 (s.d. = 4.4) in a sample of 300 college students. In a more recent study of 100 students, Paulhus (1984) reported means of 13.3 (s.d. = 4.3) and 15.5 (s.d. = 4.6) in anonymous and public disclosure conditions, respectively. In a sample of 503 students, Tanaka-Matsumi and Kameoka (1986) reported means of 14.0 and 12.3 for normal and depressed respondents, respectively. In a sample of 650 Peace Corps volunteers (90% college graduates), Fisher (1967) found means of 16.1 (s.d. = 6.8) and 16.4 (s.d. = 6.5) for males and females, respectively.

Reliability

Internal Consistency

Alpha coefficients ranged from .73 to .88 in the samples reported above.

Test–Retest

Crowne and Marlowe (1964) reported a test–retest correlation of .88 over 1 month. Fisher (1967) reported a value of .84 over a 1-week interval.

Validity

Convergent

The scale, as published in 1960, was intended as a measure of SDR in self-reports. A series of studies, summarized in Crowne and Marlowe (1964), uncovered a broad range of correlates suggesting the existence of an underlying motivational construct, namely, need for approval. For example, evidence showed that, compared to low scorers, high scorers on the MCSD respond more to social reinforcement, inhibit aggression, and are more amenable to social influence. Their task performance is more influenced by the evaluations of others. They prefer low-risk behaviors and avoid the evaluations of others, even when there is as much possibility for positive as for negative evaluation (for reviews of the research, see Crowne, 1979; Millham & Jacobson, 1978; Strickland, 1977).

Discriminant

As noted in the introduction, the MCSD falls primarily on the second SDR factor, showing only low to moderate correlations with such measures as Edwards SD and Self-Deceptive Enhancement.

Location

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, **24,** 349–354.

Comments

The MCSD continues to sustain a dual existence as an SDR scale and a measure of the approval-dependent personality. Both interpretations are consistent with analyses showing the scale taps predominantly the second factor of SDR, that is, impression management (Paulhus, 1984).

One review (Strickland, 1977) was generally supportive of the need for approval construct, but recommended the label "approval motivation." Millham and Jacobson (1978) seem to prefer "evaluative dependence." The original prefix "need" was fashionable when the scale was developed but now seems presumptuous.

In addition, the weight of evidence has gradually shifted the interpretation to avoiding disapproval, rather than seeking approval, as implied by the original label (Allaman *et al.*, 1972; Crowne, 1979; Millham & Jacobson, 1978). Finally, some work has suggested that the attribution and denial items may be tapping distinct constructs (Millham, 1974; Paulhus & Reid, in press; see also Roth, Snyder, & Pace, 1986).

Use of the MCSD scale as a measure of situational demand is well-supported: Several studies have demonstrated its sensitivity to various audience effects (Davis & Cowles, 1989; Paulhus, 1984). Such effects, however, do not prove that subjects *consciously* modified their self-presentations.

More controversial is the question of whether high MCSD scores predict a proneness to dissimulation. A classic supporting example is Kiecolt-Glaser and Murray (1980): After an assertiveness training program, high MCSD scorers rated themselves as more assertive than low scorers although the program trainers rated them as *less* assertive. Other evidence suggests that high scorers will actually lie for reasons related to social approval (Jacobson, Berger, & Millham, 1970), but there is no clear evidence that they will lie for other reasons.

A complicating factor in interpreting certain studies is that, according to their spouses, high MCSD scorers actually do possess such desirable qualities as good adjustment, friendliness, and openness to experience (McCrae & Costa, 1983). Nonetheless, correlations in that study suggest that high MCSD scorers may further exaggerate their claims to such good qualities. A further complication is that high MCSD scorers also possess an honest demeanor: That is, judges tend to believe them and trust them even when they are instructed to lie (Riggio, Salinas, & Tucker, 1988). Indeed there is some evidence for a self-deceptive component (Millham & Kellogg, 1980; Weinberger, in press).

Marlowe–Crowne Scale

Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you.

T F 1. Before voting I thoroughly investigate the qualifications of all the candidates.

T	F	2. I never hesitate to go out of my way to help someone in trouble.
T	F	*3. It is sometimes hard for me to go on with my work if I am not encouraged.
T	F	4. I have never intensely disliked anyone.
T	F	*5. On occasion I have had doubts about my ability to succeed in life.
T	F	*6. I sometimes feel resentful when I don't get my way.
T	F	7. I am always careful about my manner of dress.
T	F	8. My table manners at home are as good as when I eat out in a restaurant.
T	F	*9. If I could get into a movie without paying and be sure I was not seen, I would probably do it.
T	F	*10. On a few occasions, I have given up doing something because I thought too little of my ability.
T	F	*11. I like to gossip at times.
T	F	*12. There have been times when I felt like rebelling against people in authority even though I knew they were right.
T	F	13. No matter who I'm talking to, I'm always a good listener.
T	F	*14. I can remember "playing sick" to get out of something.
T	F	*15. There have been occasions when I took advantage of someone.
T	F	16. I'm always willing to admit it when I make a mistake.
T	F	17. I always try to practice what I preach.
T	F	18. I don't find it particularly difficult to get along with loudmouthed, obnoxious people.
T	F	*19. I sometimes try to get even, rather than forgive and forget.
T	F	20. When I don't know something I don't at all mind admitting it.
T	F	21. I am always courteous, even to people who are disagreeable.
T	F	*22. At times I have really insisted on having things my own way.
T	F	*23. There have been occasions when I felt like smashing things.
T	F	24. I would never think of letting someone else be punished for my wrongdoings.
T	F	25. I never resent being asked to return a favor.
T	F	26. I have never been irked when people expressed ideas very different from my own.
T	F	27. I never make a long trip without checking the safety of my car.
T	F	*28. There have been times when I was quite jealous of the good fortune of others.
T	F	29. I have almost never felt the urge to tell someone off.
T	F	*30. I am sometimes irritated by people who ask favors of me.

T	F	31. I have never felt that I was punished without cause.
T	F	*32. I sometimes think when people have a misfortune they only got what they deserved.
T	F	33. I have never deliberately said something that hurt someone's feelings.

Note: Items marked with an asterisk are keyed negatively.

MMPI Lie (L) Scale

(Meehl & Hathaway, 1946)

Variable

The L scale was designed to identify respondents who are deliberately trying to appear socially desirable while completing the MMPI.

Description

The scale comprises 15 statements about attitudes and practices that are socially undesirable but common. Topic areas include minor dishonesties, aggression, bad thoughts, and weaknesses of character. As on the MMPI as a whole, the response format is True–False. For all items, "False" is the response scored as a lie. In current usage, scores of 8 or above are considered suggestive of purposeful self-presentation (Greene, 1980).

Samples

Hathaway and McKinley (1951) reported means of 4.2 (s.d. = 2.6) and 4.5 (s.d. = 2.6) for males and females, respectively. In a more recent sample of 765 college students, Goldberg (1972) reported means of 2.5 (s.d. = 1.9) and 2.7 (s.d. = 1.8) for males and females, respectively. In a massive sample of 50,000 medical outpatients, Swenson, Pearson, and Osbourne (1973) reported means of 4.2 (s.d. = 2.3) and 4.8 (s.d. = 2.3) for males and females, respectively.

Reliability

Internal Consistency

Gocka (1965) reported an alpha coefficient of .72 on a patient sample. Paulhus (1984) reported an alpha value of .60 on a student sample.

Test–Retest

Test–retest correlations for intervals of up to 1 week range from .70 to .85. For intervals of 1 year or more, correlations range from .35 to .60 (Greene, 1980; Rorer & Goldberg, 1965).

Validity

Convergent

Evidence for concurrent validity is available from studies showing high correlations with similar constructs, for example, Eysenck's Lie scale (Paulhus, 1986) and the Marlowe–Crowne scale (e.g., Edwards & Walsh, 1964). As noted in the introduction, the L scale loads on both factors of SDR but primarily on the second factor, impression management.

No claim has been made that high scores should predict lying outside of self-reports. In the only known laboratory study, high scorers performed better than lows in a stressful situation (Burish & Houston, 1976).

Discriminant

The measure shows low correlations with measures loading on the first factor of SDR. For example, correlations with the K scale, the alternative SDR measure from the MMPI, range from low to moderate (Dahlstrom, Welsh, & Dahlstrom, 1972).

Location

Hathaway, S. R., & McKinley, J. C. (1951). *MMPI manual.* New York: Psychological Corporation.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook* (Vol. 1). Minneapolis: University of Minnesota Press.

Comments

It was soon apparent to the constructors of the L scale that the desirability demand of the items was less than subtle: "The L score was a trap for the naive subject but easily avoided by more sophisticated subjects" (Meehl & Hathaway, 1946). It is obvious to a sophisticated test-taker that, even if one is trying to appear desirable, it is unrealistic to deny such ubiquitous attributes.

Given its negative correlation with intelligence, some have construed the L scale as a measure of psychological sophistication. College-educated persons and those of higher

MMPI Lie Scale

Sample Items

1. At times I feel like swearing. (F)

   TRUE FALSE

2. I get angry sometimes. (F)
3. Sometimes when I am not feeling well I am cross. (F)

Complete Scale

The MMPI booklet numbers for the 15-item Lie scale are 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165, 195, 225, 255, 285. All items are keyed negatively, that is, one point is assigned for each "False."

socioeconomic status rarely score above 4 (Dahlstrom *et al.*, 1972). When they do, it may suggest a gross lack of insight into their own behavior (Greene, 1980).

MMPI K Scale

(Meehl & Hathaway, 1946)

Variable

The K scale was designed as a subtle measure of SDR on the MMPI. It is used to identify persons with psychopathology whose MMPI protocols appear normal.

Description

The 30 items were selected empirically. First, MMPI responses from normals were compared to those of persons with known psychopathology who scored as normals on the clinical scales. This procedure yielded 22 discriminating items. Depressed and schizophrenic patients, however, scored low on these items. Therefore, eight items were added to differentiate these two groups from normals. In current usage, scores of 16 or above are said to suggest invalid MMPI protocols (Greene, 1980).

Samples

In the original MMPI sample of 610 normals (a cross section of Minnesota residents), McKinley *et al.* (1948) reported means of 12.8 (s.d. = 5.6) and 12.1 (s.d.=5.1) for males and females, respectively. In their mixed sample of 968 psychiatric cases, the means were 14.6 (s.d. = 5.9) and 14.3 (s.d. = 5.2) for males and females, respectively. In their sample of 100 university students, the means were 16.1 (s.d. = 5.1) and 15.7 (s.d. = 5.0) for males and females, respectively.

In a more recent sample of college students, Goldberg (1972) reported means of 15.4 (s.d. = 4.7) and 15.5 (s.d. = 4.3) for males and females, respectively. In a massive sample of 50,000 medical outpatients, Swenson *et al.* (1973) reported means of 15.4 (s.d. = 4.9) and 15.5 (s.d. = 4.8) for males and females, respectively.

Reliability

Internal Consistency

Gocka (1965) reported an alpha coefficient of .82 on a patient sample.

Test–Retest

Correlations range from .78 to .92 for an interval up to 2 weeks and range from .52 to .67 for intervals from 8 months to 3 years (Greene, 1980; Rorer & Goldberg, 1965).

Validity

Convergent

According to the original test constructors, the scale "was not assumed to be measuring anything which in itself is of psychiatric interest" (Meehl & Hathaway, 1946, p. 544). The

validation of the scale was considered to rest on its value as a correction factor. That is, controlling other measures for K should improve their predictive validity.

In this respect, the validational evidence is weak. A few studies have shown improved validity of the MMPI scales after correcting scores as recommended in the MMPI manual (e.g., Wooten, 1984). Most studies have shown, if anything, decreases in validity (e.g., Heilbrun, 1963; McCrae *et al.*, 1989; Yonge, 1966).

The validity of K as a measure of defensiveness is better supported but appears to vary according to the type of respondent. In maladjusted college students, there is evidence that K indexes defensiveness (e.g., Heilbrun, 1961; Reis, 1966). In normal college students, K appears to tap a healthy positive self-image (e.g., McCrae *et al.*, 1989; Yonge, 1966).

Location

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology*, **30**, 525–564.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook* (Vol. 1.) Minneapolis: University of Minnesota Press.

Comments

As noted in the introduction, the K scale falls on the first factor of SDR. This association is consistent with the original conception of the K scale as a subtle measure of desirable responding. At the same time, this association suggests that at least some high scores result from a positive bias in self-image.

MMPI K Scale

Sample Items

1. I like to let people know where I stand on things. (T)
   TRUE FALSE
2. I have very few quarrels with members of my family. (T)
3. People often disappoint me. (F)

Complete Scale

The MMPI booklet numbers for the items keyed "False" are 30, 39, 71, 89, 124, 129, 134, 138, 142, 148, 160, 170, 171, 180, 183, 217, 234, 267, 272, 296, 316, 322, 374, 383, 397, 398, 406, 461, 502. The sole item keyed "True" is MMPI booklet number 96.

CPI Good Impression (Gi) Scale

(Gough, 1952)

Variable

The Gi scale was designed to measure what people say about themselves when trying to create an extremely favorable impression (Gough, 1987, p. 36).

Description

Following Ruch (1942), test development involved contrast groups. Subjects first took an experimental booklet of items under normal circumstances and then repeated the testing with "good impression" instructions: "Try to give just as favorable an impression of yourself as you would if you were actually applying for an important position, or were trying to create a very favorable impression. . ."

The items tested included some adopted from Ruch (1942) and others newly written to measure impression management. The 40 best-differentiating items were included on Gough's (1957) CPI and five were modified for the revised CPI (Gough, 1987). Scores can range from 0 to 40 with scores above 30 suggestive of faking good.

Samples

Gough (1987) reported means and standard deviations for a wide variety of samples including 4126 college students (18.5, s.d. = 5.9), 100 nurses (18.6, s.d. = 5.5), and 345 prison inmates (17.9, s.d. = 7.0).

Reliability

Internal Consistency

Gough (1987) reported alpha coefficients of .77 for both male and female college students.

Test–Retest

Gough (1987) reported test–retest correlations of .68 after 1 year for both male and female high school students.

Validity

Convergent

A total of 400 CPI respondents were rated by their spouses using Q-sorts. The four Q-sort items showing the largest positive correlations with respondents' Gi scores were: (1) A conscientious and serious-minded person, (2) Well-organized, capable, patient, and industrious; values achievement, (3) Gentle, considerate, and tactful in dealing with others, and (4) Gets along well with others; able to "fit in" easily in most situations. In addition, 793 respondents were rated on Q-sorts by trained assessors. The four highest correlating

items were (1) Is fastidious, (2) Favors conservative values in a variety of areas, (3) Is a genuinely dependable and responsible person, (4) Tends toward overcontrol of needs and impulses; binds tensions excessively; delays gratification unnecessarily. These and other validity data suggest that the high scorer is a highly controlled individual who behaves in a socially conventional manner.

Location

Gough, H. G. (1987). *California Psychological Inventory administrator's guide*. Palo Alto, CA: Consulting Psychologists Press.

Comments

The substantial validity data on the Gi highlight the dualistic nature of SDR measures. According to people who should know (spouses, peers, interviewers), the high scorer actually is a socially desirable person in being dependable, industrious, and cooperative.

However, the raters also see the high scorer as overcontrolled, suggesting an unwillingness to acknowledge undesirable qualities. This defensiveness is borne out by the fact that to score high on the Gi the respondent had to claim desirable qualities well beyond those validated by the judges (e.g., cultured interests, social skills). Other evidence of the high scorer's tendency to put the best foot forward is that interviewers rated him/her as well groomed, well dressed, and polite.

This dualism in the target construct is handled by giving a substantive interpretation to scores up to the cutoff point of 30, after which respondents are assumed to be faking good (Lanning, 1989).

The qualities measured by a scale developed through role-playing instructions depend wholly on the instructions given to the experimental group. As the reader may see above, the instructions used to select Gi items mentioned the job interview, thereby targeting the conventional, dependable, industrious, and cooperative types. This persona is, of course, only one of many possible good impressions.

CPI Gi Scale

Sample Items

1. I always follow the rule: business before pleasure. (T)

   TRUE FALSE

2. I have never deliberately told a lie. (F)
3. I enjoy hearing lectures on world affairs. (T)

Complete Scale

The CPI booklet numbers for the items keyed "True" are 14, 103, 127, 133, 140, 165, 195, 222, 254. The items keyed "False" are 10, 30, 34, 38, 42, 44, 48, 56, 66, 70, 78, 81, 91, 101, 102, 109, 120, 150, 153, 159, 170, 178, 203, 207, 231, 238, 248, 262, 268, 273, 289, 293.