Accepted for publication in forthcoming

“The ITC International Handbook of Testing and Assessment".

Response biases

Eunike Wetzel[1], Jan R. Böhnke[2], & Anna Brown[3]

[1]University of Konstanz, Germany

[2]Mental Health and Addiction Research Group (MHARG), Hull York Medical School and Department of Health Sciences, University of York, UK

[3]University of Kent, UK

Corresponding author:

Eunike Wetzel, University of Konstanz, Department of Psychology, Box 31, 78457 Konstanz, Germany. Email: eunike.wetzel@uni-konstanz.de

Response biases

Most assessments of psychological constructs such as personality traits, interests, or attitudes rely on questionnaires where respondents are required to describe themselves (self-report) or others (other-report) by responding to a set of items. The responses to the items are often given on dichotomous or polytomous rating scales which consist of a fixed number of categories. Examples of common rating scales include *strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree* or *true* and *false*, or *never*, *sometimes*, and *always*. The assumption researchers and practitioners make when using scores from these questionnaires to draw inferences is that the underlying latent trait level and the random error are the only factors influencing participants' responses to the items. For example, a person who endorses *strongly agree* on the item "I keep my paperwork in order" is assumed to be higher on conscientiousness than a person who endorses *agree* on the same item. These types of comparisons between respondents on the item or test score level are only valid when there are no other systematic influences on item responses. However, often this is not the case and there are additional factors that influence item responses. A response bias can be defined as the "systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991; p.17).

This broad definition of response biases includes different types of "non-test-relevant response determinants" (Crowne & Marlowe, 1960; p. 349): Response styles that reflect a differential use of the response options independent of the items' content (e.g., a tendency to agree with statements) and response biases that reflect a tendency to distort responses in order to align them with contextual demands or one's self-concept such as the tendency to give socially desirable responses. In the following, we will first introduce different types of response biases that commonly occur in questionnaire data. Second, we will summarize research on the assessment and third, on the correlates of several important response biases.

Fourth, we will describe different methods that can be applied at the stage of test construction to avoid or minimize the occurrence of response biases. Fifth, methods developed for correcting for the effects of response biases when they have been detected in the data will be depicted. We end with conclusions and suggestions for further research on response biases.

Types of Response Biases

Different types of response biases occur in self-report and other-report questionnaires. Table 1 summarizes the definition and relevant characteristics of the most important response biases. One class of response biases occurring in self-report and other-report questionnaires are response styles. Response styles reflect systematic tendencies of respondents to prefer certain response categories over others. These include extreme response style (ERS), the tendency to prefer extreme categories, midpoint response style, the tendency to prefer the midpoint of a response scale, acquiescence response style (ARS), the tendency to prefer categories stating agreement, and disacquiescence response style, the tendency to prefer categories stating disagreement. While these response styles reflect individual differences in interacting with the response scale, careless responding (sometimes also called random responding or non-contingent responding) refers to inattentive responding that does not reflect a preference for certain response categories (Meade & Craig, 2012). Thus, participants with a systematic response style will show response patterns dominated by one or two response categories (e.g., *strongly disagree* and *strongly agree* for ERS). In contrast, participants using careless responding might show a seemingly random response pattern or they might repeat certain responses (e.g., *agree*) or response sequences (e.g., *agree*, *disagree*, *agree*, *disagree*) over and over again (J. A. Johnson, 2005; Meade & Craig, 2012). Both response styles and careless responding are mainly independent of item content.

Another response bias differs from the previous two in that it requires reading and comprehending the item's content for it to occur: socially desirable responding (SDR) is characterized by a "tendency to give positive self-descriptions" (Paulhus, 2002; p. 47). Thus,

it involves distorting responses in a way to make them more in line with social norms and expectations. According to Paulhus' two-factor model of SDR (Paulhus, 1984, 2002), it contains both an intentional component (impression management; sometimes called faking) and an unconscious component (self-deception). Other response biases that belong in this category are simulation and dissimulation (e.g., overreporting or underreporting certain symptoms or behaviors; Meehl & Hathaway, 1946).

Responses to describe other individuals, experiences or services (other-report questionnaires) can be affected by the same response biases as self-report data. Response styles, inattentiveness, and socially (or politically) desirable representation of the rated target all can occur. In addition, there are biases unique to observer-reported data – halo effects and leniency/severity biases. The halo effect is the tendency to like or dislike all features of the assessment target including those one has not observed (Kahneman, 2011; Thorndike, 1920). This powerful cognitive bias creates a false sense of coherence in judgment, whether the assessed features indeed tend to co-occur or not.

Leniency bias describes an observer's tendency to be lenient in all of his/her assessments; and severity is the opposite tendency to be harsh/severe. If left uncontrolled for, this bias can render assessments made by different observers incomparable, since some observers consistently elevate and some depress their ratings. The following sections will discuss the most researched response biases, response styles, SDR, and rater biases, in turn.

Assessment of Response Styles

Methods developed for detecting the presence of response styles can be described by a combination of two aspects: 1) manifest or latent variable modeling and 2) the use of trait items or separate items. The first aspect differentiates manifest approaches such as frequency indices (e.g., counting the number of extreme responses) from latent variable modeling approaches such as item response models for the measurement of response styles. The second aspect refers to whether the same set of items are used to assess both the trait of interest and

one or more response styles, or whether a separate set of items that are not related to the trait of interest is applied specifically for the purpose of assessing response styles.

The most popular manifest approach which was used from the beginnings of research on response styles involves the computation of frequency indices (Bachman & O'Malley, 1984; Baumgartner & Steenkamp, 2001; Berg & Collier, 1953; Cronbach, 1946; Greenleaf, 1992a; Lorge, 1937). For example, for ERS, the number of extreme categories a respondent endorsed in all questionnaire items is counted. For ARS, the response categories stating agreement are counted. These indices are included in several instruments as validity checks in the scoring of test results. For example, the manual of the NEO-Personality Inventory (Costa & McCrae, 1992) advises practitioners to check each test protocol for ARS and disacquiescence response style by counting the number of items a respondent agreed or disagreed to, respectively.

In early research on response styles, the items measuring the trait or traits of interest were used to compute the response style indices (Bachman & O'Malley, 1984; Berg & Collier, 1953; Hui & Triandis, 1985; Lorge, 1937). Later, the notion was put forward that a separate heterogeneous item set, consisting of conceptually unrelated items, should be used to assess response styles in order to "cleanly separate stylistic variance from substantive variance" (Baumgartner & Steenkamp, 2001; p. 154). This has been implemented in recent research using frequency indices such as the representative indicators for response styles (RIRS) method, which uses a random sample of items from multiple scales that are not related to the trait of interest to assess response styles (deBeuckelaer, Weijters, & Rutten, 2010; Weijters, Schillewaert, & Geuens, 2008). However, the application of additional items (Weijters et al. (2008) recommended 14 items per response style indicator) that are not of substantive interest is often not feasible due to testing time constraints and considerations for test-taker motivation.

A different line of research has explored using latent variable models to separate response style variance from true trait variance. One model type applied for this purpose are latent class models (Lazarsfeld, 1950) or mixed Rasch models (Rost, 1990), which analyze participants' response patterns with the goal of differentiating subgroups (latent classes) that differ systematically in their response behavior. These models have, for example, been used to show the presence of ERS and non-ERS in personality questionnaire data (Austin, Deary, & Egan, 2006; Meiser & Machunsky, 2008; Rost, Carstensen, & von Davier, 1997; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013) and organizational questionnaire data (Eid & Rauber, 2000). One distinguishing characteristic of the use of latent class models to detect response styles is that they assume the resulting subgroups to differ qualitatively, i.e., the response style is purported to be a categorical variable with the response style either being present or not present. This approach – unlike frequency indices or latent trait models – therefore does not allow for individual differences in the degree to which respondents employ a response style. This approach has been extended to combine latent class analysis with Rasch models (i.e, mixed Rasch models, Rost, 1990; Rost et al., 1997), regression analysis (Moors, 2010) or with confirmatory factor analysis (Moors, 2003, 2012; Morren, Gelissen, & Vermunt, 2012) in order to simultaneously model response style groups and individual differences in the traits of interest.

Another latent variable modeling method was proposed by Billiet and McClendon (2000) who specified a latent method factor in a confirmatory factor analysis (CFA) fashion to model ARS. They used the questionnaire items for modeling the trait factor(s) and the method factor. Importantly, both positively and negatively worded items loaded positively on the method factor (see also Maydeu-Olivares & Coffman, 2006). A disadvantage of this method is that, for model identification reasons, it can only be applied to "balanced" scales (i.e. scales containing both positively and negatively worded items). The use of multidimensional item response models to model response styles is based on a similar

conceptual background as the CFA approach: Additional factors (dimensions) are defined to model the trait and one or more response styles simultaneously. Bolt and Johnson (2009) and Bolt and Newton (2011) modeled traits using the original item responses given on a polytomous rating scale. ERS was modeled using the same items, but the item responses were recoded to reflect whether an extreme category had been endorsed or not. In a three-dimensional item response model, the combined items from two substantive traits were used for modeling the response style dimension (Bolt & Newton, 2011). This method was recently extended by Bolt, Lu, and Kim (2014) who also estimated a multidimensional item response model, but used anchoring vignettes instead of the trait items to measure response styles. The advantages of this new approach are that the anchoring vignettes can be used to model any type of response style and that responses styles are not confounded with traits. The approaches based on CFA and multidimensional item response models generally allow a more differentiated analysis of response styles than approaches based on latent class analysis since the response style factors are modeled as continuous variables on which respondents can vary.

Another method presented by Böckenholt (2012) is based on the idea that the response process of endorsing a certain category can be separated into sub-processes that are either related to the trait or related to response styles. By defining pseudo-outcomes of different sub-processes (e.g., whether to endorse the midpoint or make a positive or negative decision), the two factors (response styles and traits) can be distinguished (see also Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014).

In sum, a multitude of methods have been proposed for the assessment of response styles that differ with respect to their underlying assumptions (response styles as categorical or continuous variables), their approach of using manifest or latent variables, their use of separate items or trait items for modeling response styles, and with respect to how they separate trait variance from response style variance. However, a comprehensive comparison

of the ability of these different methods to detect response styles and thus an investigation of their convergent validity is yet missing.

Assessment of Socially Desirable Responding

In contrast to response styles, the SDR bias has mainly been assessed by applying questionnaires designed to measure the propensity to respond in a socially desirable fashion. These questionnaires usually include two types of items: 1) items that describe infrequent but socially approved behavior (e.g., "I always pick up my litter on the street") and 2) items that describe frequent but socially disapproved behavior (e.g., "I like to gossip at times"). Participants with high scores on infrequent, approved behaviors and low scores on frequent, disapproved behaviors are assumed to exhibit strong SDR. A large number of social desirability scales has been developed such as the Edwards Social Desirability scale (Edwards, 1957) and the Self-Deception Questionnaire (Sackeim & Gur, 1978; for an overview see Paulhus, 1991), though the two most popular instruments appear to be the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984) and the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960). The BIDR is based on Paulhus' two-factor model and contains subscales to measure both impression management and self-deception. The factor structure of the MCSDS is rather disputed, though it largely appears to measure impression management (Uziel, 2010). Some instruments assessing substantive traits also include validity scales to check for SDR such as the L ("lie") scale of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1989). However, the validity of these scales has been subject to debate (e.g., Costa & McCrae, 1992), though one meta-analysis found them to have reasonable differential power in laboratory settings (Baer & Miller, 2002).

The use of social desirability scales to measure respondents' tendency for SDR has been called into question for several reasons. First, like any other self-report questionnaires, instruments assessing SDR may themselves be affected by SDR or other response biases.

Second, SDR (or more specifically impression management) scales appear to measure a trait which can be interpreted as interpersonally oriented self-control, not SDR (Uziel, 2010). The idea that SDR scales measure substance rather than style has already been put forward by several researchers including Borkenau and Ostendorf (1992) and Diener, Sandvik, Pavot, and Gallagher (1991). Thus, using scores on SDR scales to "correct" scores on scales assessing substantive traits for SDR might actually remove valid trait variance, thereby distorting any results found for example with respect to correlates of the trait of interest (Ones, Viswesvaran, & Reiss, 1996). Third, social desirability is also an item characteristic that can be manipulated during test construction (Backstrom, Bjorklund, & Larsson, 2009). Fourth, the rated desirability of the response options often shows a non-linear relationship to the trait and is strongly context dependent (Kuncel & Tellegen, 2009). That is, the highest response option is not necessarily the most desirable one and which response option is rated as the most desirable depends on the instruction (e.g., applying for a job as a nurse vs. as a salesperson; Dunlop, Telford, & Morrison, 2012). Approaches using SDR scores to adjust substantive scores assume a linear relationship between the trait and SDR and may therefore be inappropriate in many instances. Thus, methods to adequately assess SDR are at the moment still lacking. A promising new method to assess a specific aspect of SDR, namely the tendency to self-enhance, has been proposed by Paulhus, Harms, Bruce, and Lysy (2003): The over-claiming technique requires respondents to rate their knowledge of facts including famous persons, scientific findings, historical events, or the like. However, about 20% of the presented "facts" are made-up: they do not refer to real persons, real scientific findings, real historical events etc. The underlying rationale is that respondents who claim to (confidently) recognize these non-existing facts may have a stronger tendency to self-enhance. This questionnaire design allows the tendency to over-claim to be measured using signal detection theory.

Assessment of Rater Biases

Directly following from its definition, the halo effect manifests itself in high correlations between all, even conceptually unrelated, items, sometimes leading to effective redundancy of assessment facets. One consequence is spuriously high estimates of internal consistency (high Cronbach's alpha). A multitude of methods to assess the halo effect has been proposed. One obvious method involves the comparison of the observed correlations between variables with the theoretically expected correlations (Thorndike, 1920). Any significant discrepancy might indicate halo. The caveat to relying on this method is that very rarely in psychology we have error-free measures. Scores derived from questionnaires are usually subject to non-trivial measurement error, and consequently to attenuation in correlation coefficients due to unreliability. Thus, without any halo bias, observed correlations between conceptually overlapping traits are usually smaller than the theoretical correlations between them, and a direct comparison of the two types of correlations to assess halo may be misleading.

Another method involves computing the variance across all variables within each assessed individual (ratee), again comparing the result with some expected standard. Smaller than expected intra-ratee variances would suggest the presence of halo effects. Fisicaro and Vance (1994) suggested obtaining and comparing both measures – the correlations and the intra-ratee variances – because neither method is superior to the other in all conditions. Factor analytic procedures can also be used to assess the presence of halo in assessments by external raters. If a dominant general factor is present where a multidimensional structure was expected, the lack of differentiation between constructs is evident, for which a halo effect is a plausible explanation, although it might not be the only one (e.g. response styles such as acquiescence might also be present).

Despite the logical suitability of the above methods to assessments of halo effects, none of them can separate the actual halo bias from the theoretically expected overlap between various traits assessed (Murphy, Jako, & Anhalt, 1993). For example, are the frequently

observed high correlations between ratings of workplace competencies due to cognitive bias of observers, or due to a substantive factor causing all competencies to be rather similar (for example, overall job competence)[1]? Generally, it is impossible to tell these two causes apart unless special research designs are employed. One such design would include in a questionnaire several conceptually unrelated items. Substantial positive relationships between them, clearly, would indicate the cognitive bias of exaggerated emotional coherence. This relationship is then used as a baseline to estimate the extent of halo bias in remaining, "valid" items. Brown, Ford, Deighton, and Wolpert (2014) discuss this approach with the use of a bifactor model – a CFA model whereby a response to any "valid" item is underlain by two factors, the substantive trait factor and the halo factor, while a response to any "distractor" (or theoretically unrelated to the rest) item is underlain by the halo factor only. To identify such a model, the halo factor is assumed uncorrelated with the substantive factor(s).

Similar to the halo effect, the leniency bias can be assessed by employing special research designs. Thus, to identify elevation in ratings due to overall leniency of the rater, as opposed to truly high trait levels of the assessed target, multiple assessments by the same rater must be observed. Then the overall rater effect can be assessed by computing the Intra-Class Correlation (ICC). The ICC assesses the proportion of variance due to ratings coming from a certain observer (essentially, the observers having different means). Direct modelling of the leniency effect can be carried out, by assuming that leniency is a random variable varying between raters. Incorporating this variable into a measurement model for the data allows assessing the extent of the effect (specifically, by assessing the variance of this variable). One such method is the "many-faceted conjoint measurement" (Linacre, Engelhard Jr., Tatum, & Myford, 1994), whereby the rater leniency is incorporated into a Rasch model describing the

[1] To differentiate between the two causes of similarity in observer ratings, Cooper (1981) introduced so-called "illusory" halo – the cognitive bias that we call the "halo effect" in this article, and "true" halo – the theoretically expected positive manifold in correlations between measured traits.

probability of endorsing items conditional on the attribute the test is designed to measure, and also item difficulty and rater overall leniency. Generally, two-level modelling techniques where individual assessments (level 1) are nested within raters (level 2), are suitable to assess this bias. Furthermore, a framework for assessing rater biases is provided by generalizability theory (Brennan, 2001; Hoyt & Kerns, 1999).

Characteristics of Respondents and Response Biases

Contextual variables can have important effects on response biases. For example, the stakes of assessments are known to have a large effect on impression management (see above). The relevant contextual variables can be manipulated to reduce response biases as discussed in the next section. However, while it may be possible to manipulate context, it is not possible to manipulate characteristics of respondents. These, nevertheless, can have effects on response biases. Numerous studies have investigated the relationships between responses biases and socio-demographic variables and personal characteristics. However, due to inconsistencies in the findings of these studies – which may in part be attributed to differences in measuring response biases – there are not always clear results.

Socio-demographic variables. There is mixed evidence regarding the relationship of age to ERS: Some studies found an increase in ERS with age (Greenleaf, 1992b; Weijters, Geuens, & Schillewaert, 2010b) whereas others found a decrease (Austin et al., 2006; Light, Zax, & Gardiner, 1965) or even no effect at all (T. Johnson, Kulesa, Cho, & Shavitt, 2005; Moors, 2008). Similarly, Weijters, Geuens, et al. (2010b) found that ARS was positively related to age while Eid and Rauber (2000) did not find an association between the two variables. Austin et al. (2006) and Weijters, Geuens, et al. (2010b) reported that women showed higher ERS than men, though other studies did not find any gender differences in ERS levels (Eid & Rauber, 2000; Naemi, Beal, & Payne, 2009). The results for ARS are similarly inconclusive with some studies showing that women use ARS more than men (Weijters, Geuens, et al., 2010b) while other studies report no gender differences in ARS

(Light et al., 1965; Marin, Gamba, & Marin, 1992). A less disputed finding is that socio-economic status is negatively related to both ARS and ERS (Carr, 1971; Greenleaf, 1992a, 1992b; Ross & Mirowsky, 1984). The amount of variance in response styles explained by socio-demographic variables generally appears to be low (e.g., between 1 and 8% in Weijters, Geuens, et al., 2010b), indicating that other individual differences variables such as personality traits or differences at the cross-cultural level may play a more important role in influencing the degree to which respondents use response styles.

When assessing others, greater observer severity in assessments within exam contexts (i.e. when the ratee is unknown to the rater) has been associated with greater experience and being from an ethnic minority, but not with gender (McManus, Thompson, & Mollon, 2006). However, when raters and ratees are known to each other, it is likely that interactions between their characteristics might take place and cause biases. For example, Landy and Farr (1980) concluded from many studies that greater leniency is observed when rater and ratee are of the same race. In terms of gender, male raters tend to be more severe towards females than they are towards other males.

Personal characteristics. A consistent finding across studies is that ARS is negatively related to intelligence (Forehand, 1962; Gudjonsson, 1990). The relationship between ERS and intelligence is unclear with Light et al. (1965) and Das and Dutta (1969) reporting a negative relationship and Naemi et al. (2009) reporting no relationship. ERS has been shown to be related to intolerance of ambiguity, simplistic thinking, and decisiveness (Naemi et al., 2009) as well as extraversion and conscientiousness (Austin et al., 2006). ARS appears to be related to impulsiveness and extraversion (Couch & Keniston, 1960). As for the rater effects, high rater agreeableness has been shown to predict leniency in ratings of others (Randall & Sharples, 2012).

Consistency and Stability. Respondents appear to be largely consistent in their use of response styles over the course of a questionnaire; a result that generalizes across two

modeling approaches. Weijters, Geuens, and Schillewaert (2010a) showed the consistency of ARS and ERS using a path model with a tau-equivalent factor for the respective response style. Wetzel, Carstensen, and Böhnke (2013) confirmed this finding for ERS applying a second-order latent class analysis. Furthermore, participants' tendency to employ certain response styles is relatively stable over time as shown by Weijters, Geuens, et al. (2010b) for ARS, ERS, midpoint response style, and disacquiescence response style over a period of one year, by Billiet and Davidov (2008) for ARS over a period of four years, and by Wetzel, Lüdtke, Zettler, and Böhnke (2015) for ERS and ARS over a period of eight years.

Cross-cultural differences. Several studies show that cultural differences in the use of response styles exist. For example, higher ARS and ERS has been reported for African Americans and Hispanics compared with White Americans (Bachman & O'Malley, 1984; Marin et al., 1992). Furthermore, van Herk, Poortinga, and Verhallen (2004) found higher ARS and ERS in less individualistic societies (Mediterranean countries) compared with more individualistic societies (Western European countries). According to findings by T. Johnson et al. (2005), ARS is lower in countries that are higher on Hofstede's (2001) dimensions individualism, uncertainty avoidance, masculinity, and power distance whereas ERS is more prevalent in countries higher on power distance and masculinity. The relationship between ERS and masculinity was confirmed by de Jong, Steenkamp, Fox, and Baumgartner (2008). In addition, they found a positive correlation between ERS and individualism and uncertainty avoidance. Culture has also been found to play a significant role in rater biases. Thus, raters with high power distance and collectivistic values tend to be more lenient and exhibit more halo bias in judgment (Ng, Koh, Ang, Kennedy, & Chan, 2011).

Correlates of SDR. While numerous studies exist that investigate the relationships between scores on SDR scales with socio-demographic variables, personality traits, and outcome variables, most report relationships that point to the conclusion that SDR scales measure meaningful variance related to a predisposition to interpersonally oriented behavior

(Uziel, 2010). For example, high scores on SDR scales correlate positively with agreeableness and conscientiousness in both self-report and other-report (Borkenau & Ostendorf, 1992). Furthermore, SDR scores are positively related to the probability of getting married and staying married (Harker & Keltner, 2001), having greater job satisfaction and organizational commitment (Moorman & Podsakoff, 1992), and being described as possessing more positive traits (e.g., emotional stability) by spouses (Diener et al., 1991) and acquaintances (Pauls & Stemmler, 2003). Thus, without a valid method to measure SDR, not much can be said about individual differences in the tendency to use SDR.

Research on Methods for Avoiding Response Biases

Characteristics of the testing situation and the instrument can invite or discourage response biases. In this section, we first summarize research on the effects of the testing situation such as the mode of data collection and then turn to research on the effects of the instrument (e.g., the response format).

Testing situation. One important factor that influences the occurrence of response biases is the mode of data collection. ARS and ERS appear to be more common when data are collected using telephone interviews compared to face-to-face interviews (Jordan, Marcus, & Reeder, 1980). Regarding the comparison of online to paper and pencil data collection modes, Weijters et al. (2008) found that participants in the online condition used ERS and disacquiescence response style slightly less than participants in the paper and pencil condition. ARS was used more often in telephone interviews compared to paper and pencil or online data collections (Weijters et al., 2008). The effects of the interviewer in data collections based on interviews are unclear: Olson and Bilgen (2011) found that ARS was increased with more experienced reviewers though Hox, De Leeuw, and Kreft (1991) did not find an effect of the interviewer on ARS.

Test-taking motivation is another important factor that can trigger different response biases in low-stakes versus high-stakes assessment contexts. In low-stakes contexts, the main

challenge is to engage participants with the assessment process to minimize careless responding and overly relying on response styles. When respondents feel that the questionnaire is personally relevant to them (high topic involvement) and they think that the results of the research are important and useful to society, this usually increases their motivation to respond accurately (Gibbons, Zellner, & Rudek, 1999; Krosnick, 1991).

In high-stakes contexts, the main challenge is to minimize the impression management component of SDR. Motivated by the desire to get a job, or to be admitted to an educational program, respondents can and do engage in impression management behaviors (Viswesvaran & Ones, 1999). Although it is hard to see how the obvious motivators can be counteracted, it may be possible to manipulate the assessment context to reduce rater biases. Research shows that shorter intervals from observation to assessment tend to produce less halo effect (Wirtz, 2001). Also, calibration of one's ratings against other observers or other targets of assessment (e.g. assessing one competency for several colleagues sequentially, and then moving on to another competency, rather than assessing all competencies for one colleague and then moving on to another colleague) may substantially reduce the halo effect (Kahneman, 2011).

Instrument characteristics. Important characteristics of the instrument that influence response biases (and response behavior in general; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Schwarz, 1999) are the response format and item wording. Most research investigating the effects of the response format has focused on rating scales with differing numbers of categories and differences in the labeling of response options. The effects of the number of response options on ERS is unclear: Kieruj and Moors (2010) found no differences in ERS between response scales with five to 11 response categories whereas Weijters, Cabooter, and Schillewaert (2010) found that ERS was reduced with longer scales (e.g., 7 vs. 4 response options). According to Weijters, Cabooter, et al. (2010), the number of response options does not have an effect on ARS. In a large survey experiment Moors, Kieruj, and Vermunt (2014) found that bipolar scales with numerical values from -3 to +3 evoked more ERS than scales

with verbal labels from *totally disagree* to *totally agree*. In general, an even number of categories might be advisable since midpoint response style and other problems associated with the middle category (Hernández, Drasgow, & González-Romá, 2004) could be avoided this way. Weijters, Cabooter, et al. (2010) as well as Moors et al. (2014) found that fully labeled scales reduced ERS compared to only labeling the end points of the scale. A possible reason for this result is that verbally labeling all response categories clarifies their meaning for respondents and thereby increases reliability and validity (Krosnick & Berent, 1993; Peters & McCormick, 1966). However, fully labeling all response options may also increase ARS (Weijters, Cabooter, et al., 2010). In addition, full verbal anchoring of response options is preferable to numerical labelling since the choice of numbers (e.g., from 1 to 5 versus from -2 to 2) influences endorsement of the options (Schwarz, Knauper, Hippler, Noelleneumann, & Clark, 1991).

Avoiding the popular rating scale format and instead applying a unidimensional or multidimensional forced-choice format could also be an option to reduce response styles. In the forced-choice format, several items (e.g., 3 or 4) are presented simultaneously to respondents and their task is to rank the items with respect to how well they describe their own behavior and feelings in self-reports or that of the focal person in other-reports, respectively. Research shows that rankings may lead to data with higher discriminant validity (Bass & Avolio, 1989) than ratings. The effect of comparative judgment is particularly strong for rater biases. Thus, leniency, being a uniform effect, is eliminated when a forced-choice format is employed (Brown & Maydeu-Olivares, 2013). Halo effects are also counteracted by explicitly forcing differentiation between different characteristics (Kahneman, 2011), which can lead to an increase in operational validities of measures by as much as 50% (Bartram, 2007).

Another important factor is item wording. Test constructors should avoid ambiguous and/or complex language in items since response styles intensify when respondents are

uncertain about item content (Cronbach, 1946; Podsakoff et al., 2003) and when cognitive load increases (Knowles & Condon, 1999). A frequently posited idea is that ARS can be controlled by balancing the scale with respect to positively and negatively worded items (Baumgartner & Steenkamp, 2001; Nunnally, 1978). However, this may come at the cost of impairing response accuracy and therefore validity (Schriesheim & Hill, 1981). Regarding the influence of item wording on SDR, Backstrom et al. (2009) demonstrated that item responses are influenced by SDR less when items on a Big Five inventory were rephrased to be more neutral as opposed to socially desirable or socially undesirable.

In sum, several steps can be taken during test construction to minimize or avoid the occurrence of response biases. These steps involve careful item wording, using an adequate response format, increasing participants' motivation in low-stakes assessments, and carefully managing the process in high-stakes assessments.

Correcting for Response Biases

Variance in questionnaire data is assumed to be due to true variance (variance due to individual differences on the traits the test is designed to measure) and error variance (due to all item-by-person random influences). When response biases are present in the data, however, the variance they contribute is not random but systematic, and unless modelled appropriately, can mask itself as true variance. Response biases can affect the results of statistical analyses for example by causing spurious relationships between variables (Moors, 2012) or by overestimating relationships (Wetzel, Carstensen, et al., 2013). A number of post-hoc methods of correcting questionnaire data for the effects of response biases (mainly response styles) have been developed. Some of these were already described above in the context of the assessment of response biases, since they allow assessing and correcting response styles simultaneously.

A correction method based on linear regression was suggested by Webster (1958). It involves first computing frequency indices to quantify the degree to which a respondent

employed a certain response style. Then, the trait score is regressed on the response style index and the regression residual is computed by subtracting the expected trait score from the observed trait score. The regression residual is then argued to be free from response style variance. Extensions of the regression method include regressing the trait score on several response style indices at the same time and using a set of heterogeneous items that do not overlap with the trait items to compute the response style indices (Baumgartner & Steenkamp, 2001; Weijters et al., 2008). One advantage of this method is that it is simple to implement since it only requires sum scores as opposed to model-based trait estimates. One disadvantage of this method is that it assumes the relationship between the trait and the response style to be linear, which may not be the case.

Mixed Rasch models can be applied to differentiate subgroups of participants that differ systematically in their response behavior. Estimates of participants' trait levels are then obtained separately within each subgroup, which according to Rost et al. (1997) implies a correction for response style effects. The resulting trait estimates should then be on the same scale and therefore comparable across subgroups. One benefit of this method is that it does not require additional items to be administered since the model is run using only the trait items. However, this method only differentiates between ERS and non-ERS groups and it does not allow quantitative differences in the response style behavior.

When a method factor is modeled in addition to the trait factor to account for ARS, both positively and negatively worded items are assumed positive and equal indicators of the method factor (Maydeu-Olivares & Coffman, 2006). However, this method has the drawback that it only works for ARS and that it requires a balanced scale. Multidimensional item response models allow modeling ERS in addition to the trait. Participants' trait levels are then estimated taking into account information from both dimensions (ERS and trait). Therefore, Bolt and Johnson (2009) argue, the trait estimate is corrected for ERS. When the items from more than one trait are used to model ERS, the differentiation between variance due to the

trait and variance due to ERS is facilitated (Bolt & Newton, 2011). While this method has mainly been applied for ERS, it can easily be extended to modeling other response styles such as ARS or midpoint response style and to modeling different response styles simultaneously (Wetzel & Carstensen, in press). An advantage of using multidimensional item response models to correct trait estimates for response style effects is that the same items can be used for modeling the different dimensions. However, this is at the same time a drawback since – especially when there are only one or two traits – there are dependencies between the dimensions. For example, with one trait and one ERS dimension it is not possible to separate respondents with high trait levels from respondents with high ERS levels. A second drawback is that the response style captured by this approach is scale-specific rather than generalizable across scales (Wetzel & Carstensen, in press). A recent extension by Bolt et al. (2014) may overcome this problem by using data from anchoring vignettes to model response style behavior and rating scale data to model the traits.

Finally, item response models based on differentiating sub-processes related to the trait versus sub-processes related to response styles can also be applied (Böckenholt, 2012). For this purpose, the original item responses on a polytomous rating scale are recoded into dichotomously scored pseudo-items that indicate sub-processes related to the trait or sub-processes related to ERS or midpoint response style, respectively. Khorramdel and von Davier (2014) demonstrate the application of this model with data from a Big Five questionnaire and show how the traits of interest can be measured free from response styles by analyzing data from the pseudo-items related to the trait components of the response process. One disadvantage of this method is that the information obtained from a polytomous rating scale is essentially dichotomized into the categories agree/disagree and more differentiated information on the trait is therefore lost.

While various methods have been proposed to correct estimates of respondents' trait levels for response style effects, as with the assessment methods, a comprehensive

comparison of how effective the methods are at removing response style variance has not been conducted. A recent simulation study indicates that the multidimensional approach may be more suitable for obtaining trait estimates that are corrected for response styles than the mixed Rasch approach. Interestingly, in this study analyzing data that contained ERS with a unidimensional model (i.e., no correction for ERS) did not result in substantially worse recovery of true latent trait levels (Wetzel, Böhnke, & Rose, in press). More research on how to effectively correct trait estimates for response style effects is clearly needed.

Correcting for rater biases can also be accomplished by applying models that were used for the assessment of rater biases. For example, a bifactor model can be employed within some special designs to control for a halo factor as described above (Brown et al., 2014). Similarly, leniency can be modelled as a rater-level effect within a multilevel framework, which allows controlling for this effect and even estimating the extent of leniency bias for every rater. For example, with multi-faceted conjoint measurement (Linacre et al., 1994), a rater leniency parameter can be estimated.

Conclusion and Outlook

In the past decades, the topic of response biases has generated a lot of research investigating individual differences in the use of response biases, methods to assess response biases, and methods to correct for the effects of response biases. The trans-disciplinary nature of this research with studies for example from psychology (Podsakoff et al., 2003), economics (Baumgartner & Steenkamp, 2001; Steenkamp, De Jong, & Baumgartner, 2010), and sociology (Van Vaerenbergh & Thomas, 2013) shows that response biases are a concern for any field applying questionnaires. In this chapter, we followed a broad conceptualization of response biases, addressing both general response tendencies that are independent of item content and traits (response styles, rater biases) and response tendencies that are related to traits and depend strongly on the context (SDR). A lot has been learned though there are still plenty of open questions for example with respect to the convergence of different methods in

their ability to assess and correct for response styles. Future research could examine individual differences in response styles using consistent methods with high convergent validity in order to elucidate the currently heterogeneous findings. Studies show that response styles are used fairly consistently over the course of a questionnaire by most respondents (Weijters, Geuens, et al., 2010a; Wetzel, Carstensen, et al., 2013) and are also stable to a great extent over time (Weijters, Geuens, et al., 2010b; Wetzel et al., 2015). Thus, post-hoc corrections of trait estimates are feasible, making it even more important to understand which methods achieve estimates that can provide unbiased results in substantive research.

More research is still needed on methods to assess SDR that are not confounded with trait variance, and to differentiate between the two components of SDR. This research should take into account the assessment context since the prevalence of one component or the other is expected to vary in low-stakes versus high-stakes assessments. A large body of research shows the conditions under which response biases are more likely to occur. Thus, when designing new instruments, these findings should be taken into account since it is preferable to prevent the occurrence of response biases as opposed to correcting the data post-hoc.

References

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235-1245. doi: 10.1016/j.paid.2005.10.018

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

Backstrom, M., Bjorklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335-344. doi:10.1016/j.jrp.2008.12.013

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment, 14*(1), 16-26. doi: 10.1037//1040-3590.14.1.16

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263-272. doi:10.1111/j.1468-2389.2007.00386.x

Bass, B. M., & Avolio, B. J. (1989). Potential biases in leadership measures - how prototypes, leniency, and general satisfaction relate to ratings and rankings of transformational and transactional leadership constructs. *Educational and Psychological Measurement, 49*(3), 509-527. doi: 10.1177/001316448904900302

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 28*, 143-156.

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement, 13*, 164-169. doi: 10.1177/001316445301300202

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*(4), 542-562. doi: 10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balances sets of items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608-628. doi: 10.1207/S15328007SEM07045

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychol Methods, 17*(4), 665-678. doi: 10.1037/a0028111

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*(5), 335-352. doi: 10.1177/0146621608329891

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528-541. doi: 10.1037/met0000016

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814-833. doi: 10.1177/0013164410388411

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6*(3), 199-214. doi: 10.1002/per.2410060303

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Brown, A., Ford, T., Deighton, J., & Wolpert, M. (2014). Satisfaction in child and adolescent mental health services: Translating users' feedback into measurement. *Administration and Policy in Mental Health and Mental Health Services Research, 41*(4), 434-446. doi: 10.1007/s10488-012-0433-9

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36-52. doi: 10.1037/a0030641

Carr, L. G. (1971). Srole items and acquiescence. *American Sociological Review, 36*(2), 287-293. doi: 10.2307/2094045

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218-244. doi: 10.1037//0033-2909.90.2.218

Costa, P. T., & McCrae, R. R. (1992). Reply to Ben-Porath and Waller. *Psychological Assessment, 4*(1), 20-22.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers - Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*(2), 151-174. doi: 10.1037/H0040372

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494. doi: 10.1177/001316444600600405

Crowne, D. P., & Marlowe, D. (1960). A New Scale of Social Desirability Independent of Psychopathology. *J Consult Psychol, 24*(4), 349-354. doi:10.1037/H0047358

Das, J. P., & Dutta, T. (1969). Some correlates of extreme response set. *Acta Psychologica, 29*(1), 85-92. doi: 10.1016/0001-6918(69)90005-5

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 65*, 104-115.

deBeuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity, 44*, 761-775. doi: 10.1007/s11135-009-9225-z

Diener, E., Sandvik, E., Pavot, W., & Gallagher, D. (1991). Response artifacts in the measurement of subjective well-being. *Social Indicators Research, 24*(1), 35-56. doi: 10.1007/Bf00292649

Dunlop, P. D., Telford, A. D., & Morrison, D. L. (2012). Not too little, but not too much: The perceived desirability of responses to personality items. *Journal of Research in Personality, 46*(1), 8-18. doi:10.1016/j.jrp.2011.10.004

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden Press.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*(1), 20-30.

Fisicaro, S. A., & Vance, R. J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement, 54*(2), 366-371. doi: 10.1177/0013164494054002010

Forehand, G. A. (1962). Relationships among response sets and cognitive behaviors. *Educational and Psychological Measurement, 22*, 287-302. doi: 10.1177/001316446202200204

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research, 33*(4), 369-381.

Greenleaf, E. A. (1992a). Improving rating-scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*(2), 176-188. doi: 10.2307/3172568

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly, 56,* 328-351.

Gudjonsson, G. H. (1990). The relationship of intellectual skills to suggestibility, compliance and acquiescence. *Personality and Individual Differences, 11*(3), 227-231. doi: 10.1016/0191-8869(90)90236-K

Harker, L. A., & Keltner, D. (2001). Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology, 80*(1), 112-124. doi: 10.1037//0022-3514.80.1.112

Hathaway, S. R., & McKinley, J. C. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89*(4), 687-699. doi: 10.1037/0021-9010.89.4.687

Hofstede, G. (2001). *Culture's consequences*. Thousand Oaks, CA: Sage.

Hox, J. J., De Leeuw, E., & Kreft, I. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 439-461). New York: Wiley.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychol Methods, 4*(4), 403-424. doi: 10.1037//1082-989x.4.4.403

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly, 49*(2), 253-260. doi: 10.1086/268918

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129. doi: 10.1016/j.jrp.2004.09.009

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles - Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*(2), 264-277. doi: 10.1177/0022022104272905

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response Styles in Telephone and Household Interviewing - a Field Experiment. *Public Opinion Quarterly, 44*(2), 210-222. doi:10.1086/268585

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Allen Lane.

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(2), 161-177. doi: 10.1080/00273171.2013.866536

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research, 22*(3), 320-342. doi: 10.1093/Ijpor/Edqool

Knowles, E. S., & Condon, C. A. (1999). Why people say "Yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology, 77*(2), 379-386. doi: 10.1037//0022-3514.77.2.379

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5*(3), 213-236. doi: 10.1002/acp.2350050305

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*(3), 941-964. doi: 10.2307/2111580

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*(2), 201-228. doi: 10.1111/j.1744-6570.2009.01136.x

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107. doi: 10.1037/0033-2909.87.1.72

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star & J. A. Clausen (Eds.), *Measurement and prediciton* (pp. 362-412). Princeton, NJ: Princeton University Press.

Light, C. S., Zax, M., & Gardiner, D. H. (1965). Relationship of age, sex, and intelligence level to extreme response style. *Journal of Personality and Social Psychology, 2*(6), 907-909.

Linacre, J. M., Engelhard Jr., G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569-577.

Lorge, I. (1937). Gen-like: Halo or reality? *Psychological Bulletin, 34*, 545-546.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics - the role of acculturation and education. *Journal of Cross-Cultural Psychology, 23*(4), 498-509. doi: 10.1177/0022022192234006

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344-362. doi: 10.1037/1082-989x.11.4.344

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education, 6*, 42. doi: 10.1186/1472-6920-6-42

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods, 17*(3), 437-455. doi: 10.1037/A0028085

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*(5), 525-564.

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure - A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment, 24*(1), 27-34. doi: 10.1027/1015-5759.24.1.27

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65*, 131-149.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity, 37*(3), 277-302. doi: 10.1023/A:1024472110002

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42*(6), 779-794. doi: 10.1007/s11135-006-9067-x

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research, 22*(1), 93-119. doi: 10.1093/Ijpor/Edp036

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology, 21*(2), 271-298. doi: 10.1080/1359432x.2010.550680

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology, 44*(1), 369-399. doi: 10.1177/0081175013516114

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology*. doi: 10.1027/1614-2241/a000048

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error - A critical analysis. *Journal of Applied Psychology, 78*(2), 218-225. doi: 10.1037/0021-9010.78.2.218

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response styles. *Journal of Personality, 77*(1), 261-286. doi: 10.1111/j.1467-6494.2008.00545.x

Ng, K. Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K. Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology, 96*(5), 1033-1044. doi: 10.1037/A0023368

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Olson, K., & Bilgen, I. (2011). The role of interviewer experience on acquiescence. *Public Opinion Quarterly, 75*(1), 99-114. doi: 10.1093/Poq/Nfq067

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660-679. doi: 10.1037//0021-9010.81.6.660

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598-609. doi: 10.1037//0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. *Measures of Personality and Social Psychological Attitudes, 1*, 17-59.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*(4), 890-904. doi: 10.1037/0022-3514.84.4.890

Pauls, C. A., & Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences, 35*(2), 263-275. doi: 10.1016/S0191-8869(02)00187-3

Peters, D. L., & McCormick, E. J. (1966). Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology, 50*(1), 92-&. doi: 10.1037/H0022935

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement, 74*(5), 875-899. doi: 10.1177/0013164413514998

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. doi:10.1037/0021-9101.88.5.879

Randall, R., & Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. *Journal of Occupational and Organizational Psychology, 85*(1), 42-59. doi: 10.1348/2044-8325.002002

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health Social Behavior, 25*(2), 189-197.

Rost, J. (1990). Rasch models in latent classes - an integration of 2 approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282. doi: 10.1177/014662169001400305

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Retrieved from http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm.

Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self confrontation, and consciousness. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research* (pp. 139-197). New York: Plenum Press.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals - the effect on questionnaire validity. *Educational and Psychological Measurement, 41*(4), 1101-1114.

Schwarz, N. (1999). Self-reports - How the questions shape the answers. *American Psychologist, 54*(2), 93-105. doi: 10.1037/0003-066x.54.2.93

Schwarz, N., Knauper, B., Hippler, H. J., Noelleneumann, E., & Clark, L. (1991). Rating scales - Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582. doi: 10.1086/269282

Steenkamp, J. B. E. M., De Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research, 47*(2), 199-214.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29.

Uziel, L. (2010). Rethinking Social Desirability Scales: From Impression Management to Interpersonally Oriented Self-Control. *Perspectives on Psychological Science, 5*(3), 243-262. doi:10.1177/1745691610369465

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360. doi: 10.1177/0022022104264126

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*(2), 195-217. doi: 10.1093/Ijpor/Eds021

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. doi:10.1177/00131649921969802

Webster, H. (1958). Correcting personality scales for response sets or suppression effects. *Psychological Bulletin, 55*(1), 62-64. doi: 10.1037/H0048031

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format in response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*, 236-247. doi: 10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105-121. doi: 10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychol Methods, 15*(1), 96-110. doi: 10.1037/a0018721

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*, 409-422. doi: 10.1007/s11747-007-0077-6

Wetzel, E., Böhnke, J. R., & Rose, N. (in press). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences, 34*(2), 69-81. doi: 10.1027/1614-0001/A000102

Wetzel, E., & Carstensen, C. H. (in press). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment.*

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality, 47*(2), 178-189. doi: 10.1016/j.jrp.2012.10.010

Wetzel, E., Lüdtke, O., Zettler, I., & Bohnke, J. R. (2015). The stability of extreme response style and acquiescence over 8 years. *Assessment*. doi: 10.1177/1073191115583714

Wirtz, J. (2001). Improving the measurement of customer satisfaction: A test of three methods to reduce halo. *Managing Service Quality, 11*(2), 99-111.

Table 1

Common response biases and their characteristics

	Response bias	Characteristics	Representative studies
Self-report	Acquiescence response style	Preference for categories stating agreement (e.g., agree, strongly agree)	Baumgartner and Steenkamp (2001); Knowles & Condon (1999)
	Disacquiescence response style	Preference for categories stating disagreement (e.g., disagree, strongly disagree)	Baumgartner and Steenkamp (2001)
	Careless responding	Inattentive responding	Meade & Craig, 2012
	Extreme response style	Preference for extreme categories (e.g., strongly disagree, strongly agree)	Baumgartner & Steenkamp (2001); Greenleaf (1992b)
	Midpoint response style	Preference for the midpoint of a rating scale (e.g., neutral)	Hernández et al. (2004)
	Socially desirable responding	Tendency to describe oneself positively and in accordance with social norms and rules	Paulhus (2002); Uziel (2010)
Other-report	Halo	Tendency to exaggerate coherence in judgments of multiple characteristics	Kahneman (2011); Murphy et al. (1993)
	Leniency/severity	Tendency to be lenient/harsh in assessments of all objects	Podsakoff et al. (2003)