Trends in Psychology / Temas em Psicologia
ISSN 2358-1883 (online edition)

DOI: 10.9788/TP2019.2-02

Article

Methods for the Control of Extreme Response Styles in Self-Report Instruments: A Review

Ariela Raissa Lima Costa[*, 1]
Orcid.org/0000-0002-5942-6466
Nelson Hauck Filho[1]
Orcid.org/0000-0003-0121-7079

[1]*Universidade São Francisco, Campinas, SP, Brasil*

Abstract

Response styles are systematic ways of responding to self-report items that may impact the validity and the precision of scores from instruments. One of these biases is extreme responding (ER), which occurs when a person tends to use only extreme rating categories from a response scale (e.g., *totally disagree* or *totally agree*), irrespective of item content. Many different methods were developed that aim to identify and control extreme responses to provide a more accurate assessment of an individual's trait. The aim of this study is to perform a systematic review of these main techniques for statistical control of extreme responses in psychometric instruments of self-report. We identified several analytical approaches, which we organized into seven clusters, from simple count of the numbers of extreme response to the use of modern statistics methods, as Item Response Theory uni and multidimensional. Advantages and limitations of each method are discussed. We also present a general diagram that summarizes the distinct available methods we found.

Keywords: Response style, Likert, response bias, measurement.

Métodos de Controle de Respostas Extremas em Instrumentos de Autorrelato: Uma Revisão

Resumo

Estilos de respostas são formas sistemáticas de responder a itens de autorrelato, que podem interferir na validade e precisão dos escores de instrumentos. Um tipo específico é o estilo de respostas extremas (RE), em que a pessoa tende a usar categorias extremas de resposta (por exemplo, *concordo totalmente* e *discordo totalmente*), em detrimento do conteúdo do item. Na literatura, há uma miscelânea de métodos que se propõem a identificar e corrigir as respostas extremas para favorecer uma avaliação mais apurada do sujeito. O objetivo do presente estudo é proporcionar uma revisão sistematizada das principais técnicas de controle estatístico de respostas extremas em instrumentos psicométricos de autorrelato. Foram identificadas diversas abordagens analíticas, agrupadas em sete categorias, desde contagem do

[*] Mailing address: Rua Waldemar César da Silveira, 105, Campinas – SP, Brazil 13045-510. Phone: (11) 94042-8629. E-mail: arielalima10@gmail.com and hauck.nf@gmail.com
Support: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

número de respostas extremas até o uso de métodos estatísticos mais modernos, como Teoria de Resposta ao Item uni e multidimensional. Vantagens e limitações de cada método são discutidas. Apresenta-se também um diagrama geral da modelagem latente de RE no cenário atual.

Palavras-chaves: Estilo de resposta, Likert, viés de resposta, mensuração.

Métodos de Control de Respuestas Extremas en Instrumentos de Autorrelato: Una Revisión

Resumen

Estilos de respuestas son formas sistemáticas de responder a ítems de autorrelato, que pueden interferir en la validez y precisión de instrumentos. Un tipo específico es el estilo de respuesta extrema, en el que la persona tiende a utilizar categorías extremas de respuesta (por ejemplo, *concordo totalmente* y *discuerdo totalmente*), en detrimento del contenido del elemento. En la literatura, hay una miscelánea de métodos que se proponen identificar y corregir las respuestas extremas para favorecer una evaluación más apurada del sujeto. El objetivo del presente estudio es proporcionar una revisión sistematizada de las principales técnicas de control estadístico de respuestas extremas en instrumentos psicométricos de autorrelato. Se identificaron diversos enfoques analíticos, agrupados en siete categorías, desde el simple contaje de respuestas extremas hasta el uso de modernas técnicas estadísticas, como Teoría de Respuesta al Ítem uni y multidimensional. Las ventajas y limitaciones de cada método se discuten. Se presenta también un diagrama general del modelado latente de RE en el escenario actual.

Palabras clave: Estilo de respuesta, Likert, sesgo de respuesta, mensuración.

Self-report instruments represent an invaluable data collection strategy and have supported the development of psychological science for many decades. The main advantages of this method include the collection of information directly from the examinees with no interference in between, their low cost and no need for specific training (Lilienfeld & Fowler, 2006). However, this does not imply that the self-report is a perfect method. Actually, self-report inventories for personality traits have received many criticisms due to their vulnerability to some biases, especially response styles.

Response styles are systematic ways of responding to self-report items irrespective of the content. These biases occur when a person tends to choose only some of the response categories from a response scale, even from a range of items that capture distinct psychological traits. Thus, response styles might affect the validity of scores from psychometric instruments, as they constitute a systematic source of variance independent from the trait in question in the testing (Plieninger, 2017).

One of the most common response styles in psychological data collected via self-report is extreme responding (ER). Extreme responding characterizes a preference for the options or categories located at the extremes of a response scale (Greenleaf, 1992). For example, when rating items on a five-point Likert scale, a respondent with this bias might be tempted to choose only 1 or 5, no matter what the descriptive content of the item is in each case. As an illustration of how influential ER can be, a study estimated that, on average, 25% of the common variance among items of self-report inventories is due to this style of responding (Wetzel & Carstensen, 2015). Accordingly, ER comprises a variance component that needs to be considered, as it can easily confound the trait variance in the statistical analyses performed on self-report data.

One remarkable feature is that ER apparently emerges as a manifestation of the broader underlying cognitive functioning of individuals. Naemi, Beal, and Payne (2009) found that people that exhibited ER were more likely to be intolerant of ambiguity and uncertainty and had less cognitive flexibility. Meisenberg and Williams (2008) discovered a negative relationship between ER and education, income, and a positive correlation to age. Smith et al. (2016) reported that ER more often occurs in self-reliant people and in cultures favoring this characteristic. Batchelor and Miao (2016) performed a meta-analysis, and found that women are more prone to ER than men, that individuals from Hispanic cultures are more prone to it than white people and that Mexican and Australian individuals respond more extremely than North-Americans. Furthermore, they found a negative correlation between ER and intelligence.

Extreme responding might distort the validity and reliability of self-report inventories and scales. More specifically, this response style can distort the means, standard deviations, internal consistency estimates and correlations to external variables of instruments, among other issues. Moors (2012) reported that the greater scores in passive leadership observed in women when compared to men disappear after controlling for ER, suggesting that this mean difference was entirely spurious and due to systematic error variance. Jin and Wang (2014) gave an illustration of how ER might distort the rank ordering of respondents. The authors exemplified the problem with a hypothetical situation in which three respondents A, B, and C exhibit the same score in a given item (2.80), but distinct frequencies of extreme responses across the remaining items (4, 6, and 18, respectively). Although the theta estimates for the respondents were −1.47, −1.54, and −1.70 when using the uncorrected raw data, they changed to −1.66, −1.42, and −1.32 when controlling for ER. That is, the true ordering of the individuals was distorted because of extreme responses in the raw data. This illustrates how the failure to control for ER can bias theta estimates, due to systematic measurement errors in the scores. This hypothetical situation can be generalized to cross-cultural studies, in which countries are compared regarding their means of a variable assessed via self-report. Because cultural differences in the likelihood of exhibiting ER exist, this bias might confound comparisons, as well as distorting the reliability and dimensionality of the variables in the analyses, thus leading to spurious conclusions about the phenomena under investigation (Chun, Campbell, & Yoo, 1974).

All these issues have inspired the development of statistical techniques for controlling ER. Available strategies can either be targeted toward the control of response styles in general (i.e., several response styles indiscriminately) or specifically toward ER. There are many approaches to the problem, which vary according to how the model is specified and identified, and how the parameters are estimated. Techniques can consist of simple frequencies of extreme responses to latent class factor models, mixed Rasch models, and even multidimensional item response theory models (Wetzel, Lüdtke, Zettler, & Böhnke, 2015). The common feature among the approaches is the attempt to isolate the influence of ER from trait variance (e.g. Bolt & Newton, 2011), which might well solve measurement invariance issues (e.g. Morren, Gelissen, & Vermunt, 2012).

Even though there are numerous strategies to control for ER, they are not necessarily popular among researchers, with most academics being unaware of their benefits and potential applications. There are many reasons for this situation. One issue is that methods are sometimes described in technical language that is not easily accessible to researchers less familiar with complex statistical modeling. Furthermore, the majority of the techniques require proficiency in statistical programs that do not resemble the traditional SPSS and that require a long and difficult learning process. These aspects and others contribute to the lack of popularity of ER control models, thus preventing the practical refinements from becoming real. Therefore, the present study was conducted to provide a brief description of and introduction to the main statistical techniques developed for controlling ER in psychometric instruments. For this, a systematized review of the literature was conducted, being a type of review in which only some of the criteria for a standard systematic review are fulfilled (Grant & Booth, 2009).

The techniques are described in a simple and conceptual way, with their main advantages and shortcomings highlighted.

Method

Search Strategy

The search for articles was performed in three electronic databases – PubMed, PsycINFO, and Science Direct, in August 2017. Abstracts recovered were downloaded and managed, using the software Mendeley, for detailed reading. A free search strategy was conducted to ensure a wide coverage of published works, with no limit for year or language of publication. However, given the potential small number of publications in the area, broad keywords were selected for the search, namely, "extreme response style" OR "extreme response bias".

Study Selection Criteria

The screening of the abstracts recovered was guided by three criteria. First, studies should be empirical, using either real (i.e., information collected from a sample of respondents) or simulated data (i.e., artificially produced using software). Second, if the study was carried out using real data, then it should contain analyses on self-report data. Third, if the study reported simulated data, then it should report the development or test of a model for the control of ER. Investigations were excluded that focused on the identification of correlates of ER or on illustrating the use of an available method but without presenting a refinement or improvement.

Selection and Extraction of Data

The first phase in the selection was the reading of the titles and abstracts of the studies recovered, with the exclusion of those that failed to meet the inclusion criteria or that met the exclusion criterion. The eligible articles were then read in full, allowing a more refined selection of the documents. As an additional strategy, a hand search was carried out on the reference lists of the selected documents. To provide a visual illustration of this entire process, Figure 1 presents a detailed flow diagram following the PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2015). The extracted metadata were authors, year of publication, journal, objective, type of data employed, theme of the investigation and a summary of the methods developed when applicable. Studies included in the review are highlighted with a "*" in the reference list at the end of this document.

Results

Descriptive Information of the Studies

The initial search returned 293 documents, from which nine duplicates were removed. A closer inspection of abstracts helped identify 260 cases that did not fulfill any of the inclusion criteria or that fulfilled the exclusion criterion, resulting in 24 documents eligible for the full reading. Among these selected documents, only 12 perfectly matched the inclusion criteria, these being retained for further evaluation. Eight further studies were recovered using the hand search strategy (see Figure 1).

The next step was the full reading of the documents, with the resulting models grouped into seven thematic categories: Counts of extreme responses, ER as a violation of item parameter invariance, ER as a categorical latent variable, ER as a continuous latent variable, hybrid models of ER, ideal point response models, and models for decomposing latent response processes. The following sections describe the conceptual elements of each approach, with an emphasis on some of their main advantages.

Counts of Extreme Responses

The method proposed by Greenleaf (1992) is one of the simplest available approaches and consists solely of a careful selection of items to form a composite score of ER. To do this, the researcher should select items related to different psychological variables (preferably, from different instruments), with a correlation close to zero, and with similar frequency distributions in the scale categories. Once items fulfilling these conditions are found, then their original

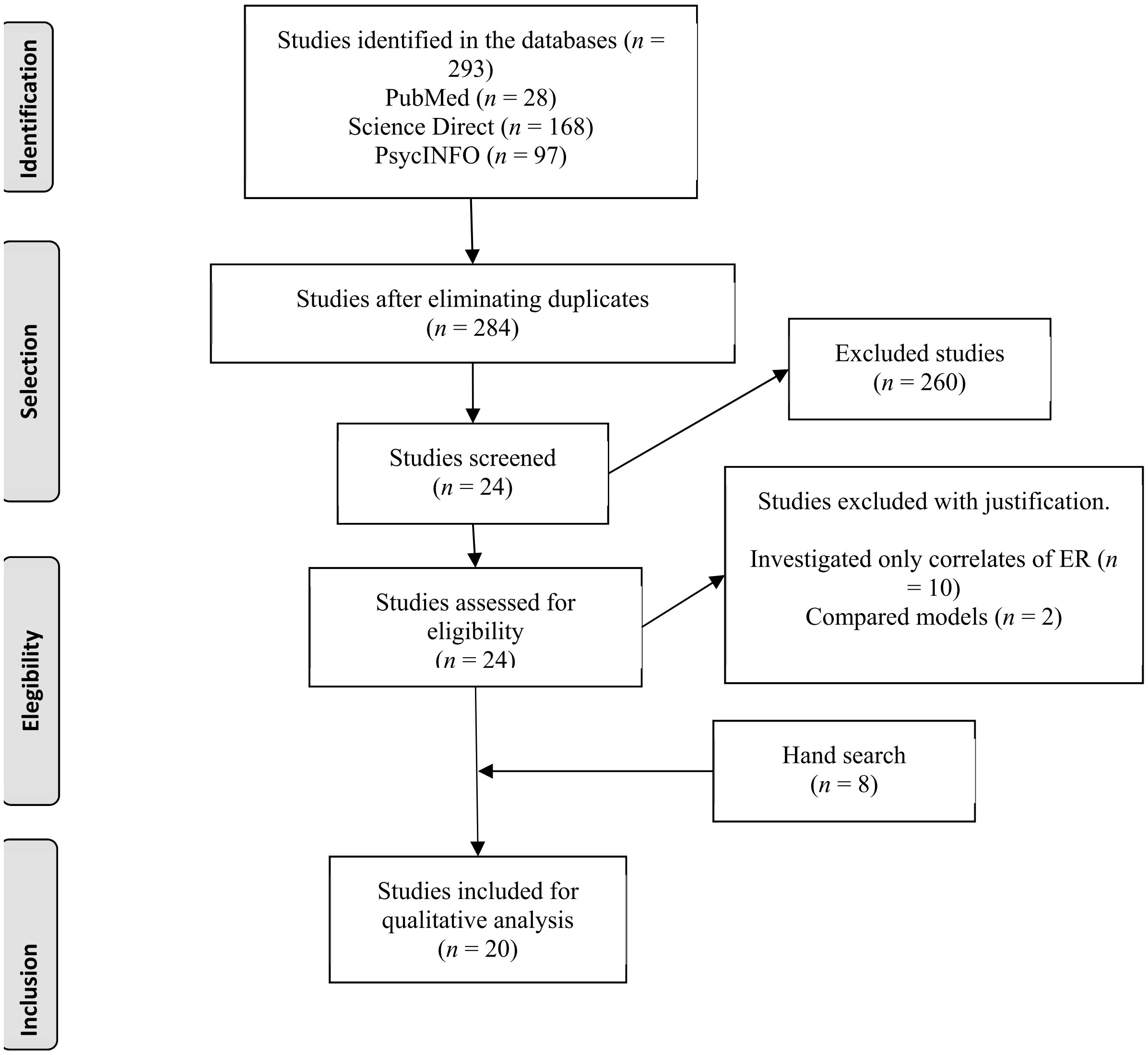
Identification
Selection
Elegibility
Inclusion
Studies identified in the databases (n = 293)
PubMed (n = 28)
Science Direct (n = 168)
PsycINFO (n = 97)
Studies after eliminating duplicates (n = 284)
Excluded studies (n = 260)
Studies screened (n = 24)
Studies excluded with justification.
Investigated only correlates of ER (n = 10)
Compared models (n = 2)
Studies assessed for eligibility (n = 24)
Hand search (n = 8)
Studies included for qualitative analysis (n = 20)

Figure 1. **PRISMA flow diagram of studies included in the review (adapted from Moher et al., 2015).**

scores are recoded so that they reflect counts of extreme responses, with them then aggregated. For example, if the original Likert scale is the type 1 2 3 4 5, then it would be transformed into 1 0 0 0 1, counting only responses 1 or 5 as “1”. Reliability of the resulting scale of binary items can be checked via internal consistency indices and its validity via correlations to external variables (Greenleaf, 1992). Another similar technique consists of calculating the general proportion of extreme responses (e.g., 1 or 5) relative to the total number of responses given to the items of the study questionnaire (Harzing, 2006). In both cases, the resulting score might be included as a control variable in regression or correlation analysis or in mean comparisons etc.

ER as a Violation of Item Parameter Invariance

This perspective treats ER is as a violation of invariance in item intercept thresholds (difficulty parameters). Models that deal with this situation are often called “mixed” or “random-effect models”. From this perspective, an ER score of latent estimate is not obtained, however, instead, the relative difficulty of endorsing item categories is allowed to vary among the respondents, which helps in dealing with extreme responses. For example, suppose two groups of individuals, A and B, with similar level of Extraversion (within and between groups) rated self-report items of Extraversion. If individuals from group A are more likely to choose extreme categories in the response scale, analyzing the data from each of these groups separately will return potentially distinct parameter estimates. If this situation exists or at least some individuals in the data are extreme respondents, then forcing items to have invariant intercept or threshold parameters among individuals might return biased theta estimates (e.g., Extraversion scores). To deal with this issue, many authors have developed models that relax the invariance assumption of intercepts or thresholds, allowing item parameters to have a distribution of values (i.e., they are treated as "random" instead of "fixed" effects). This flexibility makes it possible to model idiosyncratic uses of the response scale, such as systematically choosing extreme categories.

Mixed models that vary around this very theme were developed by Jin and Wang (2014), Johnson (2003), Rossi, Gilula, and Allenby (2011), Wang, Wilson, and Shih (2006) and Wang and Wu (2011)all rights reserved. In each of these works, the authors presented Item Response Theory (IRT) models that relax the assumption of threshold invariance. In some cases, the solution to the ER issue consisted of simply allowing the estimated distances between item categories to vary between individuals (e.g., Johnson, 2003). In other approaches, a variable was included that multiplied thresholds and designated a tendency toward choosing extreme categories (e.g., Jin & Wang, 2014). Modeling violations of invariance in the discrimination parameter (i.e., factor loading) is also useful in controlling for other response styles, such as careless responding. Nevertheless, the idea is that ER does not necessarily impact item validity (as careless responding does, for example), however, it does alter how difficult it is to choose categories for some respondents.

ER as a Categorical Latent Variable

A solution closely related to random threshold models is mixture models. The core feature in mixture models is that ER is a categorical latent variable: One or more "hidden" or non-observed groups of extreme respondents exist, which might be responsible for the observed extreme responses. In this case, an attempt is made to model threshold invariance by uncovering latent groups of individuals who display idiosyncratic use of the item response scale. Latent class models for ER are like random intercepts or threshold models, except that variability in these parameters are conceived as occurring among discrete classes of homogeneous respondents. For example, an item intercept can exhibit a larger estimate in the class of extreme respondents, which means that it is more easily endorsable by these individuals, irrespective of the trait level (i.e., the item presents differential item functioning – DIF).

Moors (2003)there in no single accepted methodological approach in dealing with this issue. This article aims at illustrating the flexibility of a latent class factor approach in diagnosing response style behavior and in adjusting findings from causal models with latent variables. We present a substantive example from the Belgian MHSM research project on integration-related attitudes among ethnic minorities. We argue that an extreme response style can be detected in analyzing two independent sets of Likert-type questions referring to 'gender roles' and 'feelings of ethnic discrimination'. If the response style is taken into account the effect of covariates on attitudinal dimensions is more adequately estimated. (PsycINFO Database Record (c developed a nominal response model specifying two descriptive (i.e., trait) factors and one ER factor. In this case, the ER factor was parameterized to present a discrete distribution (ordinal), similar to ordered latent classes of individuals. Including an ER factor returned better estimates of the relationship between descriptive factors, as well as better estimates of correlation to a series of external variables. Kankaraš and Moors (2011)these two issues have rarely been investigated together. In this article we demonstrate the flexibility of a multigroup latent-class factor approach in both analyzing measurement equivalence and detecting ERB. Using data from the European Values Survey from 1999/2000, we identified an ERB in answering Likert-type questions on attitudes toward morals of compatriots. Furthermore, we found measurement inequivalence in the form of direct effects of countries on the attitude items. The model that included both these issues resulted in quite distinct findings regarding country difference in the latent attitude compared to the models that only included one of these effects \u2013 either measurement inequivalence or extreme response. It is suggested that the all-inclusive model provides the more valid estimates of country differences in the latent attitude. (PsycINFO Database Record (c illustrated how this approach was useful in refining cross-cultural comparisons of several European countries when measuring attitudes toward morals of compatriots. Morren et al. (2012) then developed a more flexible version of Moors' (2003) model, allowing the descriptive factors to have an increasing monotonic relationship with the ordinal item scores, while the ER remains connected to items via a non-monotonic relationship (the factor increases the likelihood of responses only to the extreme categories). Furthermore, Moors (2012) extended this approach to accommodate the control of other response styles also in the model.

Another strategy that implements latent classes in the modeling of ER is the mixed Rasch model (Eid & Zickar, 2007; Rost, Carstensen, & von Davier, 1997). The difference relative to traditional Rasch-family models is that latent classes are extracted, so that item difficulty estimates can exhibit between-class variability. An inspection of estimates found for each class can help identify whether one of them reflects individuals using response styles. For example, a latent class might be characterized by items having threshold disorders (i.e., violations in the ordering of threshold parameters), especially in those that separate the extreme categories from their neighbors. Alternatively, the response style class can have ordered thresholds that are, however, less separated. A possible inference is that the members are extreme responders. If this interpretation is granted, then researchers can decide whether to exclude extreme responders from further inferential analysis using the data or to perform separate analysis for each group.

ER as a Continuous Latent Variable

Most modern modeling techniques assume response styles such as ER to be a continuous latent variable. This means that individual differences in response styles do not occur in the form of discrete groups of individuals, but as a variable with many possible levels among individuals. There are numerous examples of this. Jong, Steenkamp, Fox, and Baumgartner (2008)the authors present a new item response theory-based model for measuring ERS. This model contributes to the ERS literature in two ways. First, the method improves on existing procedures by allowing different items to be differentially useful for measuring ERS and by accommodating the possibility that an item's usefulness differs across groups (e.g., countries formulated a unidimensional IRT model that can account for ER. The modeling consists of selecting multiple unrelated items from different instruments in a database and then recoding them to capture the occurrence of extreme responses. For example, items with a Likert scale of 1 2 3 4 5 are recoded as 1 0 0 0 1 for the analysis. These binary items are then specified as indicators of a latent ER factor. Because many items might still share trait content, the authors suggested that all non-ER common variance could be modeled by including testlets for item clusters. In this case, the testlets are small factors that lie in between the ER factor and the binary indicators, which accommodate residual correlations between items that are not due to the ER factor, thus refining the measure of this response style. Furthermore, the model can also accommodate invariance violations by treating discrimination and difficulty parameters as random effects that vary among sample groups, which makes this approach excellent for refining cross cultural comparisons. It also includes covariates of the ER factor, allowing the investigation of causal antecedents of this response style. Using simulated data, Jong et al. (2008) demonstrated how the model can be identified and have its parameters recovered using standard estimation algorithms. Furthermore, the authors found a full ER model with testlets and random effects to have the best fit to real data when compared to alternative candidates that do not address the issue of testlets or parameter invariance violations.

Bolt and Johnson (2009) developed a strategy of modeling ER using Bock's nominal response model. This IRT model is useful when a latent trait explains items rated on a nominal scale. That is, no increasing monotonic relationship between the latent trait and the indicators is assumed. Relaxing this assumption is justifiable given that, if the Likert scale is 1 2 3 4 5, the latent ER factor will only increase the likelihood of choosing categories 1 or 5, the reason why no monotonic relationship is found in this case. Bolt and Johnson reported an illustrative analysis of a database containing nicotine dependence indicators regarding how an ER factor can be recovered after imposing some constraints for the sake of model identification. The proposed model consisted of an ER factor explaining nicotine dependence items rated on a 4-point Likert scale (1 = *strongly disagree*, 4 = *strongly agree*), so that category slopes (discrimination parameters) were fixed at 1, −1, −1 and 1. These constraints mean that the ER factor only increases the likelihood of responding 1 or 4, and decreases the likelihood of choosing 2 or 3. Another factor capturing nicotine dependence was included in the model, however, following standard IRT parameterization, where a monotonic relationship exist between the factor and items (category slopes fixed at −3, −1, 1 and 3). The hypothesized model achieved a better fit to the data when compared to alternative models. In a further study, Bolt and Newton (2011) extended their model to accommodate more than one trait factor. The idea was to control for ER in multiple instruments (or items from different factors) simultaneously.

A similar approach to Bolt and Johnson's (2009) was that of Wetzel and Carstensen (2015), derived from the Partial Credits model, a polytomous version of the standard Rasch model. The procedure is applicable for polytomous items and, in addition to the standard Partial Credits measurement model of the latent trait of interest, includes an ER factor connected to another set of Likert-type items (assembled from distinct instruments, as in Greenleaf's, 1992, or in Jong et al.'s study, 2008) recoded to capture extreme responses (e.g., 1 2 3 4 5 converted into 1 0 0 0 1). Once the measurement model for the ER factor has been established, items from the trait Partial Credit measurement model are allowed to crossload on the ER factor, which partially cancels out ER-related variance from the trait model. As the basis for this approach is that the ER factor is specified using items from an independent instrument (or more than one instrument), correlations between the trait and the ER factors can be estimated. Another benefit is that the model can be expanded to account for other response styles, such as acquiescence (i.e., agreeing with items irrespective of their content). The authors illustrated the use of the approach using real personality data (the Big Five factors).

Hybrid Models of ER

Some of the approaches for addressing the issue of ER combine features of the modeling of continuous and categorical latent variables. The model proposed by Huang (2016)interest, and personality to measure a variety of latent traits. Extreme response style (ERS, for example, allows researchers to include both a continuous ER factor that explains item variance beyond the trait factor of interest, and the latent classes of respondents. These latent groups of individuals are intended to capture violations of invariance in item intercepts or factor loadings. That is, the model accounts for ER by including a latent factor and latent classes, this being the reason why the approach is called "hybrid". The main idea is to isolate the contribution of an ER factor in the item responses, allowing each item to have a discrimination parameter on the ER factor, and another on the trait factor. Items can vary in the extent to which they elicit extreme responses and individuals might also vary in their propensity to give extreme responses. The model follows an intuitive notion that the more an item is discriminative for the ER factor, the less it is informative on the trait dimension of interest. One distinctive feature is that the model deals with measurement invariance by allowing item parameters to vary among latent classes of respondents, as well as modeling more than one trait factor (e.g., Extraversion and Neuroticism). Latent classes in the model are intended to reflect varying levels of ER. For example, in the simulated data analysis reported by the authors, three latent classes were recovered: extreme responders, mid-point responders, and "typical" responders.

It should be stressed that mixed Rasch models (Eid & Zickar, 2007; Rost et al., 1997) can also be included in this category, as they involve latent classes of ER. In sharp contrast with Huang's (2016) work, mixed Rasch models typically contain no ER factor, only a trait dimension. However, in this case, grouping of individuals with different difficulty estimates can arise so that latent classes in which items have lower threshold values (or threshold disorder) can be interpreted as comprising extreme respondents. Furthermore, in mixed Rasch models, only invariance violations in the difficulty parameters are accounted for. Huang's (2016) approach is more sophisticated, as it also deals with measurement invariance violations in the discrimination parameters.

Ideal Point Response Models

Although not very popular, ideal point models are an interesting alternative. Most IRT models assume a monotonic relationship between the latent trait and the indicators (the nominal response model, previously described, is an exception). In factor or graded response models, it is expected that the increases in the latent scores are followed by increases in the observed item scores (or the contrary, if the item is negatively scored). For example, the classical understanding is that someone with a high level in Extraversion might give a rating of 5 ("totally agree") to an item like "I would be happy to spend the afternoon with my friends in a Café", while another individual with a lower level may respond 1 ("totally disagree"). Nevertheless, an extraverted person may perceive the context "in a Café" as boring or providing little appealing stimulation, while an introverted individual might feel more comfortable talking to friends at home. If this happens, both extraverted and introverted individuals will disagree with this item. Thus, a high score (5 or totally agree) would be expected only from individuals with moderate levels of Extraversion. The pattern of the relationship between factor and item follows a "U" shape, this being the reason why models that address this issue are called "ideal point response" (Stark, Chernyshenko, Drasgow, & Williams, 2006)similar in spirit to Thurstone's work in the context of attitude measurement, can provide viable alternatives to the traditionally used dominance assumptions for personality item calibration and scoring. Item response theory methods were used to compare the fit of 2 ideal point and 2 dominance models with data from the 5th edition of the Sixteen Personality Factor Questionnaire (S. Conn & M. L. Rieke, 1994. Javaras and Ripley (2007)where responses fall into ordered categories ranging from disagreement to agreement. Social science and marketing researchers frequently use data of this type to measure attitudes toward an entity such as a policy or product. We focus on data on American and British attitudes toward their respective nations (\"national pride\" developed an ideal response model capable of accounting for acquiescent and extreme response styles. The approach consists of specifying random thresholds in an ideal response model (see "ER as a violation of item parameter invariance"), which allows variability in these parameters caused by response styles.

Models for Decomposing Latent Response Processes

Item response tree models constitute a recent strategy of controlling for ER (Böckenholt, 2012). Response tree models can decompose the latent response processes in several cognitive stages, then address the influence of distinct phenomena related to choosing one response category over the other options. The person must decide between choosing a neutral or a non-neutral response (if a neutral response is available), then whether the non-neutral response will be positive or negative and, finally, what the intensity of the agreement or disagreement with the item content will be. Von Davier and Khorramdel (2013) developed a bi-factor response tree model that is capable of controlling for a general factor of response style that impacts on item responses. Items are recoded binarily following Greenleaf's (1992) procedure, and then the fit of a unidimensional 2-parameter IRT model is compared with a bi-factor model and a hierarchical model, both containing a general factor. While response styles will tend to manifest as a general factor, trait dimensions will emerge as specific factors. Using Big Five data, the authors discussed the benefits and the ease of using the procedure in practical research situations.

Discussion

The present study introduced some of the main strategies for the statistical control of ER. A systematized review of the literature revealed that there are many statistical techniques available, ranging from simple sum or count scores of endorsement of extreme categories to modern modeling approaches that consider multiple sources that explain variance in the observed scores in self-report inventories.

Two features common to all these techniques should be highlighted. First, controlling for ER requires multiple items, irrespective of using simple counts of extreme responses (e.g., Greenleaf, 1992) or modern latent modeling (e.g., Jin & Wang, 2014). The issue is, even though response styles are undesirable biases, they account for only a relatively small fraction of between-item common variance (e.g., ER represent about 25%, as estimated by Wetzel & Carstensen, 2015). Thus, isolating this common variance parcel from the remaining trait factors requires an assemblage of (preferably many) items with mixed content. Second, ER is considered to be one of the causes for measurement invariance violations in self-report instruments. This means that ER can be regarded as an issue of differential function of the person, which can impair group comparisons. Not attending to this problem can, therefore, lead researchers to spurious inferences regarding the nature of psychological phenomena and the way they manifest in various social settings.

Despite each method having its merits and possibly being recommended as a potential solution for the control of ER, none of them are perfect. Counts and frequencies of endorsement of extreme categories (e.g., Greenleaf, 1992; Harzing, 2006) are simple, easy and intuitive techniques to compute an ER score. However, this approach assumes that all ER indicators are perfectly valid and reliable. No control of measurement error is offered and, therefore, no estimates of a latent variable of ER are provided. This is the reason why, although useful, the count strategy is less desirable and should not be chosen over the more refined models and techniques developed.

Random threshold models often offer no estimate of a latent ER variable (Jin & Wang's 2014 model being an exception). One issue is that, even though these techniques allow the control of ER, they do not provide an ER score that can be used in further inferential analyses (e.g., to compare groups or correlate with an external criterion). Furthermore, most of the models included in this category are unidimensional, which is less useful when the data reflects variability across multiple latent traits. By contrast, the IRT testlet model developed by Jong et al. (2008) accommodates more than one response style, however, no trait factors. A mixed effect approach might be useful in controlling for response styles, however, perhaps not very specific, as at this point only ER is controlled and not other styles such as acquiescence (the tendency to agree with items irrespective of their content) or socially desirable responding (the tendency to agree with items with socially acceptable content). Accordingly, random intercept models have also been proposed as a potential solution to the problem of acquiescence (see Maydeu-Olivares & Coffman, 2006)common for all participants.

Latent class models can establish a score representing ER in the form of membership of a latent group of individuals. The case of latent class factor analysis is even more interesting because it allows classes to be ordered according to their levels of ER (e.g., Moors, 2003, 2012; Morren et al., 2012). These models address measurement invariance issues, and can refine cross cultural comparisons. However, latent class models usually assume that classes are internally homogeneous, with no within-class variability between individuals, which is an unlikely assumption that should be tested against the alternative hypothesis that ER is a continuous latent variable. Furthermore, even though the recommendation is that multiple items are required when controlling for ER, latent class models tend to work at their best when using a small number of indicators (Eid, Langeheine, & Diener, 2003). Hybrid models such as Huang's (2016) require estimating even more parameters than standard latent class models, something that requires large sample sizes and computers with a high data processing capacity.

Another general shortcoming is that the nature of the latent processes that produce ER is barely known. Accordingly, it can only be hypothesized whether the ER latent variable has either a monotonic or a non-monotonic relationship with the model indicators. Despite the existence of alternative strategies to the classic IRT/factor models such as the ideal response models (e.g., Javaras & Ripley, 2007), little is known about whether these models accurately describe real data collected under most research conditions and whether these models are more representative of indicators with specific content (e.g., attitudes). Perhaps comparisons are still needed between monotonic models and nominal response (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011), ideal response (e.g., Javaras & Ripley, 2007) and tree decision models (von Davier & Khorramdel, 2013) to help researchers decide which one is most appropriate for specific situations. In any case, a model capable of integrating the entire available knowledge about the latent processes that produce ER is desirable, however, has not yet been proposed.

Mixture models are still under explored in psychometric studies (e.g., factor mixture models; Muthén, 2006). These models are flexible in that they can accommodate a latent factor of ER and latent groups with different score distributions in this factor. The only approach of the kind that it was possible to find was Huang's (2016), which is very recent and has not yet been popularized. Perhaps, the benefits of using mixture models would be even greater if researchers could manage to include covariates that are theoretically related to ER, to help identify and estimate the latent classes. Latent classes capture, in an exploratory fashion, similarities between individuals, and their recovery might benefit from the inclusion of covariates in the model (see https://www.statmodel.com/download/relatinglca.pdf). To trust that the statistical analysis per se will identify groups only differing in ER levels can be risky, as response styles are systematic sources of variance that are confounded by true trait variance. The use of external information could help achieve better parameter estimates for these complete models, as well as provide a deeper understanding of the causes of ER.

Even considering that shortcomings exist in each approach developed to control ER, together these strategies address a multitude of psychometric issues. A general framework of ER modeling can be established from the studies found by the present literature review, as shown in Figure 2. From this picture, it can be seen that extreme responding emerges from a latent variable that explains the way an individual chooses item categories irrespective of his or her level in a series of trait factors. The main strategies discussed here can be implemented assuming that ER is a latent variable that is categorical (qualitatively distinct groups), ordered categorical (ordered groups) or continuous (quantitative differences with no groups). In each case, model parameterization is independent from the type of indicators, which can be nominal (e.g. Bolt & Newton, 2011), counts (e.g. Greenleaf, 1992), categorical ordered variables (e.g. Moors, 2003)there in no single accepted methodological approach in dealing with this issue. This article aims at illustrating the flexibility of a latent class factor approach in diagnosing response style behavior and in adjusting findings from causal models with latent variables. We present a substantive example from the Belgian MHSM research project on integration-related attitudes among ethnic minorities. We argue that an extreme response style can be detected in analyzing two independent sets of Likert-type questions referring to 'gender roles' and 'feelings of ethnic discrimination'. If the response style is taken into account the effect of covariates on attitudinal dimensions is more adequately estimated. (PsycINFO Database Record (c, continuous variables (e.g. Jong et al., 2008)the authors present a new item response theory-based model for measuring ERS. This model contributes to the ERS literature in two ways. First, the method improves on existing procedures by allowing different items to be differentially useful for measuring ERS and by accommodating the possibility that an item's usefulness differs across groups (e.g., countries or even others. Connection between latents and their indicators can be estimated as linear, logistic, probit or other types. Furthermore, it is admissible to allow for variability in the model parameters, such as intercepts (e.g. Jin & Wang, 2014), thresholds (e.g. Javaras & Ripley, 2007)*where responses fall into ordered categories ranging from disagreement to agreement. Social science and marketing researchers frequently use data of this type to measure attitudes toward an entity such as a policy or product. We focus on data on American and British attitudes toward their respective nations (\"national pride\"* or factor loadings (discrimination parameters; e.g. Wetzel & Carstensen, 2015). By contrast, if the modeling is only focused on estimating a latent variable of ER, then including testlets for controlling trait factors is also possible. Outcomes and covariates can also be accommodated in distinct hierarchical levels of the model.

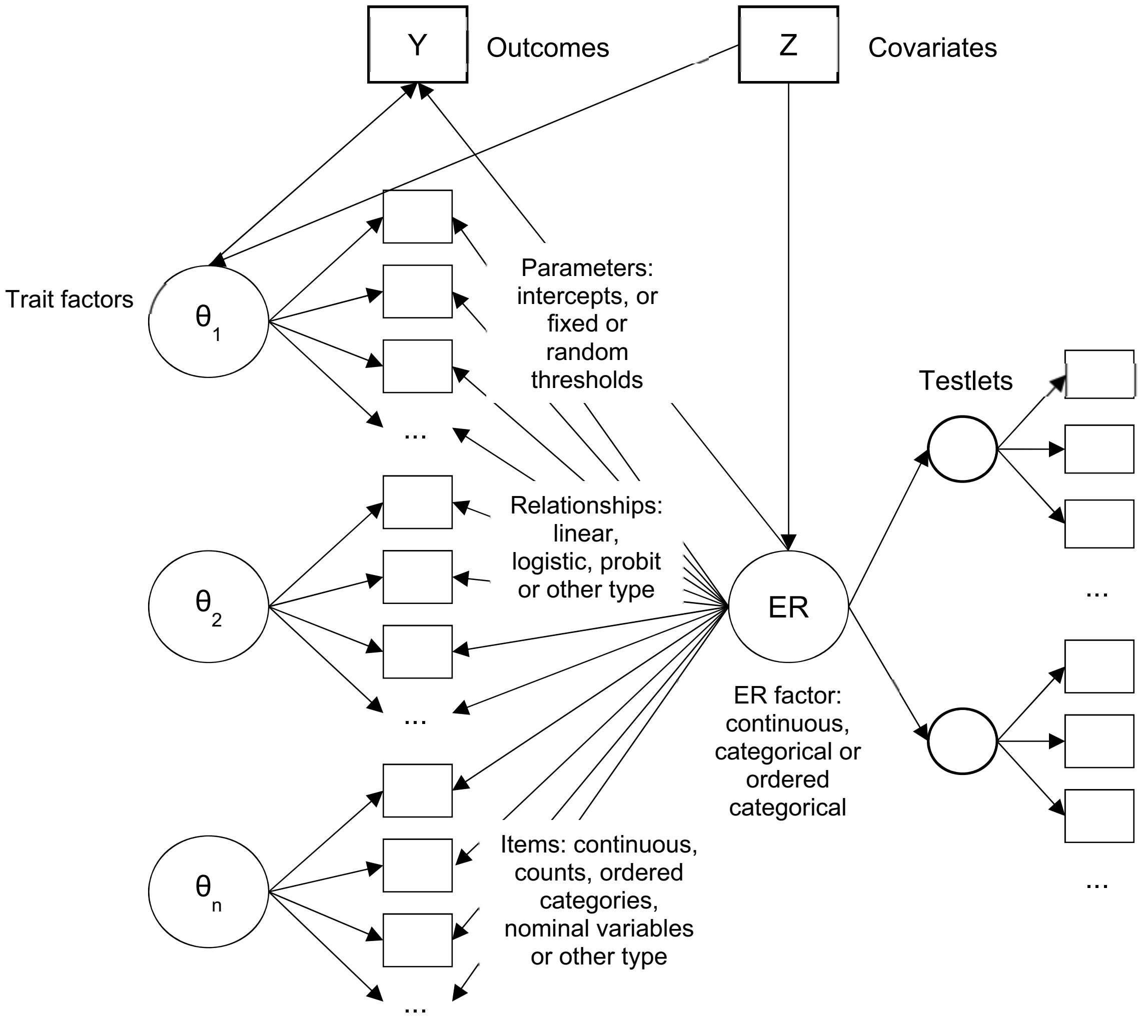
Y
Outcomes
Z
Covariates
Trait factors
θ1
θ2
θn
Parameters: intercepts, or fixed or random thresholds
Relationships: linear, logistic, probit or other type
Items: continuous, counts, ordered categories, nominal variables or other type
ER
ER factor: continuous, categorical or ordered categorical
Testlets
...

Figure 2. Synthesis of latent variable models for the control of ER.

Final Considerations

This study reviewed methods developed to control extreme response styles, to provide researchers with a broad perspective of the modeling strategies available and maybe inspire the use of these techniques in psychometric investigations. The study is also intended to be a guide or summary of the main options in the area, and an invitation to a more profound reflection about psychological research that is dependent on self-reports. The focus of this work was rather conceptual, so that more technical details were avoided, which can be accessed from the original publication of each model. A further step to help popularize models to control for ER might be the writing of tutorials instructing researchers in how to make practical use of the methods described here.

References

Batchelor, J. H., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology, 16*(2), 51-62.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665-678. https://doi.org/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. https://doi.org/10.1177/0146621608329891

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

Chun, K.-T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross cultural research. *Journal of Cross-Cultural Psychology*, *5*(4), 465-480. https://doi.org/10.1177/0146167299025006006

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. *Journal of Cross-Cultural Psychology*, *34*(2), 195-210. https://doi.org/10.1177/0022022102250427

*Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In C. H. Carstensen (Ed.), *Multivariate and Mixture Distribution Rasch Models* (pp. 255-270). New York,: Springer-Verlag.

Grant, M. J., & Booth, A. (2009). A typology of reviews : An analysis of 14 review types and associated methologies. *Health Information and Libraries Journal*, *26*(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

*Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*(3), 328. https://doi.org/10.1086/269326

*Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243-266. https://doi.org/10.1177/1470595806066332

*Huang, H.-Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*, 1706. https://doi.org/10.3389/fpsyg.2016.01706

*Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data. *Journal of the American Statistical Association*, *102*(478), 454-463. https://doi.org/10.1198/016214506000000960

*Jin, K., & Wang, W. (2014). Generalized IRT models for extreme response style. *Educa*, *74*(1), 116-138.https://doi.org/10.1177/0013164413498876

*Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*. https://doi.org/10.1007/BF02295612

*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*. https://doi.org/10.1509/jmkr.45.1.104

*Kankaraš, M., & Moors, G. (2011). Measurement equivalence and extreme response bias in the comparison of attitudes across Europe: A multigroup latent-class factor approach. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. https://doi.org/10.1027/1614-2241/a000024

Lilienfeld, S. O., & Fowler, K. A. (2006). The self-report assessment of psychopathy: Problems, pitfalls, and promisses. In C. J. Patrick (Ed.), *Handbook of Psychopathy* (pp. 107-132). New York: Guilford Press.

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

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*(7), 1539-1550. https://doi.org/https://doi.org/10.1016/j.paid.2008.01.010

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2015). Principais itens para relatar revisões sistemáticas e meta-análises: A recomendação PRISMA. *Epidemiologia e Serviços de Saúde*, *24*(2), 335-342. https://doi.org/10.5123/S1679-49742015000200017

*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: International Journal of Methodology*. https://doi.org/10.1023/A:1024472110002

*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*. https://doi.org/10.1080/1359432X.2010.550680

*Morren, M., Gelissen, J., & Vermunt, J. (2012). The Impact of Controlling for Extreme Responding on Measurement Equivalence in Cross-Cultural Research. *Metodology*, *8*(4), 159-170. https://doi.org/10.1027/1614-2241/a000048

Muthén, B. (2006). The potential of growth mixture modelling. *Infant and Child Development: An International Journal of Research and Practice, 15*(6), 623-625. https://doi.org/10.1002/icd.482

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261-86. https://doi.org/10.1111/j.1467-6494.2008.00545.x

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32-53. https://doi.org/10.1177/0013164416636655

*Rossi, P. E., Gilula, Z., & Allenby, G. M. (2011). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, *96*(453), 20-31. https://doi.org/10.1198/016214501750332668

*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). New York: Waxmann.

Smith, P. B., Vignoles, V. L., Becker, M., Owe, E., Easterbrook, M. J., Brown, R., …Harb, C. (2016). Individual and culture-level components of survey response styles: A multi-level analysis using cultural models of selfhood. *International Journal of Psychology*, *51*(6), 453-463. https://doi.org/10.1002/ijop.12293

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25-39. https://doi.org/10.1037/0021-9010.91.1.25

*von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (Vol. 66). New York: Springer.

*Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling Randomness in Judging Rating Scales with a Random-Effects Rating Scale Model. *Journal of Educational Measurement*. National Council on Measurement in Education. https://doi.org/10.2307/20461834

*Wang, W.-C., & Wu, S.-L. (2011). The Random-Effect Generalized Rating Scale Model. *Journal of Educational Measurement*, *48*(4), 441-456. https://doi.org/10.1111/j.1745-3984.2011.00154.x

*Wetzel, E., & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of Extreme Response Style and Acquiescence Over 8 Years. *Assessment*. https://doi.org/10.1177/1073191115583714

Received: 11/02/2018
1st revision: 21/05/2018
Accepted: 22/05/2018