What's the measure? An empirical investigation of self-ratings on response scales

Sven Hilbert, Florian Pargent, Elisabeth Kraus, Felix Naumann, Kathryn Eichhorn, Patrizia Ungar, Markus Bühner

Angaben zur Veröffentlichung / Publication details:

Hilbert, Sven, Florian Pargent, Elisabeth Kraus, Felix Naumann, Kathryn Eichhorn, Patrizia Ungar, and Markus Bühner. 2022. "What's the measure? An empirical investigation of self-ratings on response scales." *International Journal of Social Research Methodology* 25 (1): 59–78. https://doi.org/10.1080/13645579.2020.1839163.

What's the measure? An empirical investigation of self-ratings on response scales

Sven Hilbert,[1] Florian Pargent[2], Elisabeth Kraus[1], Felix Naumann[2], Kathryn Eichhorn[2], Patrizia Ungar[2], Markus Bühner[2]

[1] Faculty of Psychology, Educational Science and Sport Science, University of Regensburg, Regensburg, Germany
[2] Department of Psychology, Psychological Methods and Assessment, Ludwig-Maximilians-University, München, Germany

Correspondence can be addressed to Elisabeth Kraus: elisabeth.kraus@ur.de

Abstract

The present investigation comprises two studies. In Study 1, participants gave numerical information about demographic attributes (real-scores). They subsequently rated themselves regarding these attributes on a five-point Likert-type scale (5LTS). Items used different phrasings, inducing (1) a general, (2) a personal, and (3) an outsiders' perspective. By regressing these ratings on the real-scores, it was shown that information on centers and intervals of the real-scores were not readily reflected by the response scales. This led to different representations of the intervals and centers of the real-scores. The outsiders' perspective resulted in the most adequate representation of the real-score intervals. Study 2 used neutral item wording with a 5LTS and a four-point Likert-type scale (4LTS) to investigate the possible confound of positive wording. This increased the adequacy of the representations only slightly. Together, the findings indicate that, even on average, the investigated rating scales and items reflect the actual attributes only limitedly and that the self-ratings depend on the item phrasing instead of simply representing a coarse measure of the real-scores. All data and analysis scripts are available on https://osf.io/4pcdb/.

Keywords: Response format; level of measurement; scale; psychometrics; validity

What do we measure when we ask someone to rate him or herself on a rating scale regarding a particular attribute? The idea of rating scales is to obtain valid information about a criterion, which is not measured directly but by taking the detour and asking about it. More than a century of research on rating scales has produced a plethora of theories and investigations about the possibilities and impossibilities of validly measuring just about anything with rating scales and questionnaires. The present article approaches the issue empirically by measuring physical attributes with different self- rating scales and discusses the implications for validity, the central aspect of any measurement.

Introduction to Validity

Next to objectivity and reliability, validity poses the third main quality criterion in psychological measurement (Messick, 1993). Extensively discussed on a semantic level, this study explores validity from a statistical point of view by investigating the assignability of single item rating scale responses to real-scores of physical values.

Validity means measuring, what is intended to be measured. Thereby, the nature of validity is two-fold: on the one hand, validity means a semantic reference of item content and the associated construct, which concerns content validity in its narrower sense. This has been discussed extensively in the literature (Borsboom & Mellenbergh, 2002; Messick, 1995, Schwall, Hedge, & Borman, 2012, etc.,) but goes beyond the scope of this investigation. We will focus on a different aspect: the numerical structure in which this semantic reference between items and truth scores can be expressed. This mapping of item response scores and truth value scores is the second crucial part of validity.

The Structure of Measurement and Rating Scales

A (measurement) scale is defined as the triplet of an empirical relational structure, a numerical relational structure, and a function depicting this empirical relational structure on the numerical relational structure (see Luce, Suppes, & Krantz, 2006). If, for example, extraversion is measured with the NEO-Five Factor Inventory (NEO-FFI, Costa & McCrae, 1989), it is assumed that the self-rating of the items used to measure extraversion do preserve a certain degree of information about the underlying empirical construct. The easiest way to investigate the degree of information preservation is to relate empirical differences between persons to numerical differences in persons' ratings. Therefore, if two different persons with two different empirical degrees of extraversion rate themselves regarding extraversion, this empirical relation should be preserved in the numerical relation (i.e., the numbers on the

rating scale). As however the true empirical degree of extraversion is a latent construct and can therefore never be observed directly, this study adopted physical scales as objects of measurement and defined validity in terms of successful mapping of item scales onto the real-score scales. Thus, the physical scale is treated as the real attribute (with negligible measurement error), and meaningful relations on this attribute between subjects shall be reflected on the item scales. As physical scales are interval (intervals between objects are meaningful) or even ratio (ratios between objects are meaningful) scales in nature, the information about the real-scores can only be adequately represented by interval scaled rating scale data. However, it directly follows from the definition of an interval scale, that meaningful differences between subjects on a continuous physical attribute can never be fully reflected on a rating scale with a small finite number of categories (Clason & Dormody, 1994). Therefore, we can only investigate whether rating scales preserve interval scaled information to a satisfying degree of approximation.

Measurement with Rating Scales

The level of measurement of (self-) rating scales has been the subject of intense controversies at least since Stevens (1946) presented his influential operationalist interpretation of measurement scales in the social sciences (Michell, 1997). Numerous authors have thus far questioned the suggestion that rating scales[1] provide information on a metric level (e.g., Clason & Dormody, 1994; Goldstein & Hersen, 2000; Michell, 1997) for a variety of mostly theoretical reasons. Clason and Dormody, for example, argued that a latent variable measured by a rating scale could at best be regarded as a coarse ordinal representation of the underlying variable. These measurement properties of rating scales are directly related to the cognitive processes assumed to underlie a person's response on a rating scale. The cognitive process model for answering questionnaire items by Tourangeau, Rips, and Rasinski (2000) postulates that respondents form an internal judgment of how they rank on the dimension described by the item before mapping their judgment on the provided response scale. In an ideal measurement scenario, this internal judgment is made solely on the real continuous dimension of the empirical relational structure (including its properties, such as origin and intervals) and item responses result from a direct mapping on the coarse rating scale, so that they choose the scale

[1] In the present article, the term „rating scale" refers to response scales for questionnaire items, which is used synonymously.

value closest to their judgment. Assuming that a person chooses the rating closest to their own judgment, this rating would only adequately reflect a adequate mapping from an empirical relational structure if it is not biased.

Reasons for Bias in the Rating Process

A framework incorporating such potential biases in a model of human judgement is the Truth and Bias Model of Judgment (T & B model) by West and Kenny (2011). The model states that the judgment, such as the rating on a response scale, is the product of two systematic components, the truth and the bias, where the latter may comprise more than one bias variable. Both truth and bias consist of a single value and a force, which pulls the judgment towards that value. Moreover, the strength with which the bias variables influence the judgement can be affected by moderator variables, since not everyone's judgment is equally susceptible to the same influences.

The T & B model provides a fitting framework for responses on (self-) rating scales, as the assumed truth value can be thought of as the score on a real continuous dimension and the bias provides an explanation for the often-stated assumption that properties of the continuum are not adequately reflected by rating scales (e.g., Clason & Dormody, 1994; Goldstein & Hersen, 2000). Next to moderator variables like gender or personality traits, also information that is implicitly given in the item wording is susceptible to influence the rating process, like the adopted frame of reference for the rating.

The personal frame of reference is crucial for a person's response on a rating scale and ultimately leads to the question of validity: if a person is asked to rate his or her age, which construct of age does the person use as a reference, the chronological age or the subjective age? If the latter is used, the frame of reference is prone to differ between respondents: it is unlikely that people will consider themselves "very old", even if they know that they belong to the elderly in their society, when asked to adopt a personal frame of reference. A personal frame of reference might rather motivate people to compare themselves to significant others in their personal environment rather than to the whole population. In rather homogeneous personal environments, this would lead to people selecting the middle categories rather than to selecting the extreme categories of rating scales and therefore to the loss of metric information.

The effect of the frame of reference for the validity of personality inventories has been discussed extensively by Lievens, de Corte, and Schollaert (2008), who found introducing a fixed frame of reference increases the validity of a questionnaire by reducing inter-person variability and intra-person inconsistency, supporting results obtained in various other investigations (e.g., Schmit, Ryan, Stierwalt, & Powell, 1995). Moreover, Schleicher, Day, Mayes, and Riggio (2002) even showed that frame of reference training can improve reliability and validity of ratings in assessment centers. Therefore, in this investigation Study 1 comprises rating scales incorporating different frames of references to not only benefit from a fixed frame of reference but also to investigate which frame of reference shows to be most beneficial for relating a rating scale to a physical scale as closely as possible.

Study 1

Rationale

To investigate the influence of the perspective induced by a frame of reference, Study 1 used an online questionnaire with five-point Likert-type scales (5LTSs) as response format and three variations in item phrasing to induce a general, a personal, and an outsiders' frame of reference. The effects of this frame of reference and the relationship between the real-scores and the ratings on the 5LTS were then investigated with separate analyses for men and women to take possible non-linear effects of gender into account.

The different frames of reference for judging an attribute, could be expected to bias the ratings in different ways, as respondents seem to differentiate well between different perspectives on themselves (see Carlson et al., 2011). Therefore, the numerical response structure should differ between the phrasings. Moreover, gender might moderate the bias in the ratings in a non-linear way, which is supported by studies on the effect of gender on the self-perception of body-related variables (see Paeratakul et al., 2002). The numerical response structure for male and female subjects could therefore be expected to differ in a non-linear manner (i.e., more complex than by a stretch and/or a shift). Combined, the following hypotheses were investigated:

Hypotheses

(i) The continuous s in the German population are not adequately partitioned into estimated latent intervals representing the categories of the self-rating scales for all types of item phrasings.

(ii) Within each attribute, intervals of the real-scores representing the categories of the self-rating scales are expected to differ between the three types of item phrasings.

(iii) Within each attribute, intervals of the real-scores representing the categories of the self-rating scales are expected to differ between men and women.

Sample

A total of n = 2091 German speaking participants (1634 female) took part in the study. An additional 10% had started the online questionnaire but did not finish and therefore did not enter the analysis. The age ranged between 18 and 95 years (quartile 1 = 24; median = 28; quartile 3 = 38) and 58 % had received 12 or more years of school education. The participants were contacted via the social network Facebook and e-mails. They received no gratification for their participation.

Method

Materials

An online questionnaire in German was created with the software SoSci Survey Version 2.6.00 (Leiner, 2014) and administered on the platform www.soscisurvey.de, asking subjects to indicate their age, height, shoe size, weight, and monthly income. First, the real-scores had to be stated and subsequently, all subjects were asked to provide self-ratings of these five attributes on three differently phrased items per attribute using 5LTSs. The phrasing of the items differed with respect to the frame of reference: for the attribute “age”, the three items were phrased “I am old” (*general frame of reference*), “I consider myself old” (*personal frame of reference*) and “Others consider me old” (*outsiders’ frame of reference*). The same variation was applied to all five attributes in exactly this manner. The exact formulations for all items and all frames of reference are listed in *Appendix A*. The participants had to be at least 18 years of age.

The data collection that included Study 1 comprised another sample of subjects that responded to a 5LTS with bar plots of the population distribution for all attributes to guide their rating. Presenting and discussing the additional data goes beyond the scope of the present investigation but the data is available on https://osf.io/4pcdb/.

Procedure

The participants filled out the questionnaire online by logging in from their personal computers or smartphones. After indicating their demographic information on the first page, participants faced five pages showing three statements regarding each attribute (one page per attribute, i.e., age, height, shoe size, weight, and income), phrased as described above for the conditions of the within-subject factor. The completion of the questionnaire took approximately three minutes.

Analysis

The statistical computations were conducted using the open statistical software R (R Core Team, 2018). Figures were created in base R or with the package *ggplot2* (Wickham, 2009). The relationships of real-scores with the 5LTS were investigated via generalized linear regressions for ordinal responses (proportional odds). The models partition the domain of the real-scores into intervals, each reflecting the region of highest response probability for a certain category of the 5LTS. Proportional odds regressions were estimated with the *ordinal* package (Christensen, 2015). The real-scores were used to predict responses in each of the different scales. Additionally, gender was considered by calculating different models for males and females. Likelihood ratio tests were conducted to test if the thresholds used to partition the domain of the real-scores can be considered equidistant in the proportional odds regressions.

All parameters of the regression analyses can be found in *Appendix B* and all results of the likelihood ratio tests between proportional odds regression models with fixed intervals and the models with freely varying thresholds, presented in the following section, are listed in *Appendix C.* When comparing the estimated middle intervals to the central moments of the real-scores, the German population means were used.

Results

Sample Distribution

Despite the far greater number of female participants, the distributions for men and women were similar in shape for the attribute age but with different peaks for height, shoe size, dress size, weight, and income. *Table 1* provides an overview of the descriptive statistics. The sample means did not differ strongly from the German population means.

Table 1: Descriptive statistics sample Study 1

Attribute	Mean	Median	SD	Min	Max	Population mean
Women						
Age (years)	34.96	28	15.88	15	85	49.56
Height (cm)	168	168	5.81	152	186	165.89
Shoe size (European)	38.41	38	1.56	35	43	38.60
Dress size (European)	38.12	37.5	3.50	32	50	-
Weight (kg)	63.08	60	11.27	46	110	70.17
Monthly Income (Euro)	1840	1500	1578.86	0	10000	1166.41
Men						
Age (years)	38.82	35	15.24	16	81	49.56
Height (cm)	181	180	6.92	160	202	178.72
Shoe size (European)	43.30	43	1.55	38	46	42.27
Dress size (European)	49.93	50	4.39	34	60	-
Weight (kg)	81.90	80	11.40	43	115	85.30
Monthly Income (Euro)	3169	2800	2587.99	0	15000	1891.61

SD = Standard deviation; Min = Minimum value; Max = Maximum value; Population mean = Estimated mean of the German population, according to the German General Social Survey, 2014, and the Deutscher Fußreport (German Foot Report), 2009; Monthly Income was reported pre-tax.

General Frame of Reference

Figures 1 and *2* illustrate the relationship between self-ratings on items with a general frame of reference and the respective real-score values as covariates. It can be seen that the thresholds of almost all items are far from equidistant, while the intervals differ more strongly for women compared to men. Likelihood ratio tests support this observation. The means of the real-scores of the German population did not consistently coincide with the middle category.

Personal Frame of Reference

Almost all attributes showed strongly varying thresholds for the self-ratings in items with a personal frame of reference, implying that the mean differences in the real-score values were not preserved by the estimated threshold parameters of the proportional odds regression model. The variation in the intervals between the thresholds was more pronounced for female subjects, as can be observed in *Figures 1* and *2* and was again underlined by the likelihood ratio tests. For both genders, the attributes' means only sometimes fell into the central category. Note that some of the thresholds were estimated in impossible regions of the physical scale (e.g., a negative value for age). The proportional odds models assume a single slope between the thresholds, which can result in extreme thresholds if these options are rarely selected (i.e. have few data points) or if the relationship with the independent variable is weak.

Outsiders' Frame of Reference

This phrasing also showed varying thresholds, as *Figures 1* and *2* illustrate. However, for both men and women, the divergence from equidistant thresholds was descriptively less severe compared to the other two phrasings and likelihood ratio tests show more non-significant results than significant ones. Also, the population means of the attributes were located in the central category in most conditions, yet only once more than for the personal frame of reference.

Discussion

The results of Study 1 show that inducing a frame of reference by phrasing the items so that the responder takes a general perspective, a personal perspective, or the view of others impacted the way in which the real-scores were

reflected by the ratings on a 5LTS. In addition, it was shown that the bias induced by the item phrasing was moderated in a non-linear way by the gender of the respondent.

Effect of Phrasing

The intervals of the models in Study 1 only partly deviated from assumed equidistance, therefore only partially supporting hypothesis (i). The three types of phrasing used here had an effect on the estimated thresholds, yet the influence differed between the attributes and genders. This means that hypotheses (ii) and (iii) were supported by the results. In particular, the models estimated independently for men and women showed a clear difference in the pattern: while for male subjects, the phrasing had an effect mainly when asked for their income, women's ratings resulted in equidistant models for age, shoe size, and weight only when the outsiders' perspective was induced. Moreover, the difference between the phrasings seemed to be strongest for the self-rating of weight.

The finding that the perception of what others think differs from the personal image of oneself has long been known (see Bem, 1972) and the differences between men and women in this regard are discussed below. Yet, the present results imply that the real-scores for the attributes and items that were tested here are more adequately represented by the 5LTS if the perspective of others is adopted.

On the other hand, the central category of the general perspective comprised the estimated means of the German population (*Figures 1* and *2*) more often than the outsiders' perspective. However, the real-scores of psychological constructs, typically measured with questionnaires, cannot be known. Yet it seems reasonable to assume that asking for the general, the personal, and the outsiders' perspective might induce different internal judgments of how to rank oneself on the dimension described by the item, for example, an aspect of extraversion or emotional stability. This notion is also supported by Carlson and Furr (2011), who found that people are able to accurately differentiate between their view of themselves and how others see them.

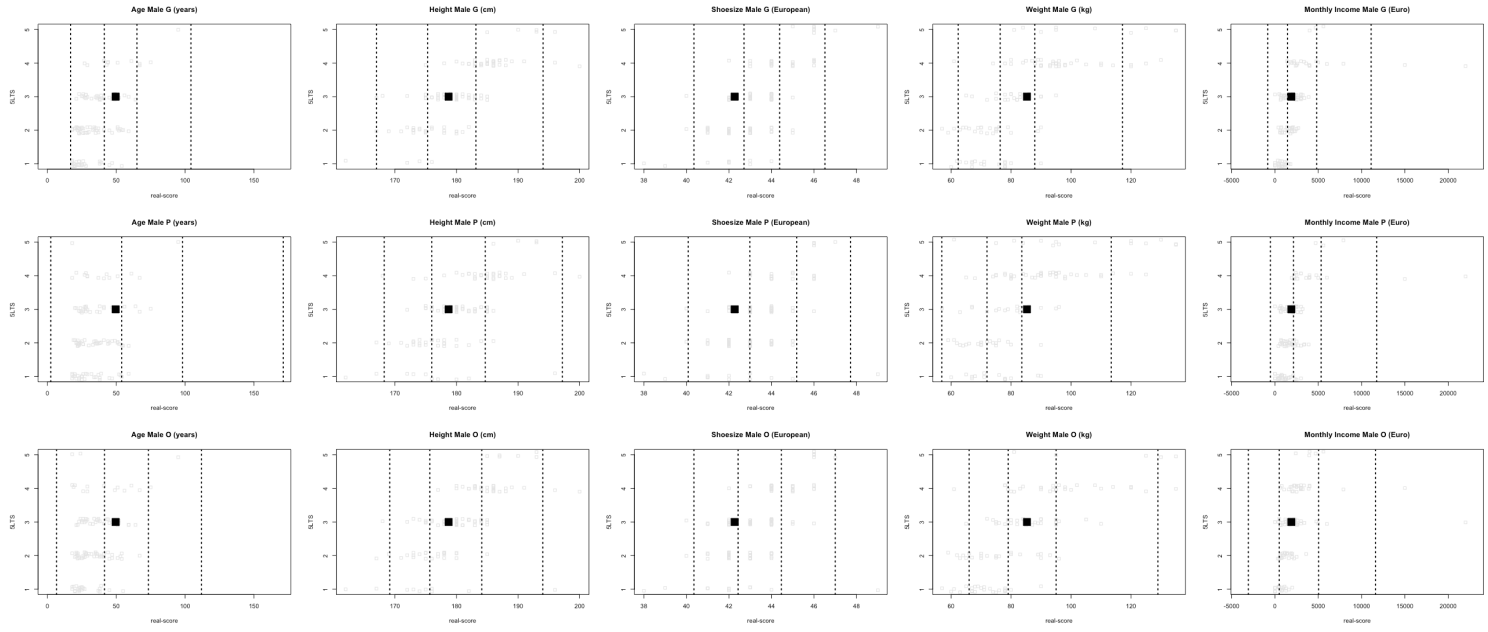

Figure 1: The thresholds of the proportional odds models of the 5LTSs for the five attributes and three frames of references are depicted for the male participants. 5LTS = five-point Likert-type scale; real-score = numerical value of the attribute; Black square = Population mean of the real-scores for male subjects; Vertical lines = predicted thresholds between two categories; G/O/P = general/personal/outsiders' perspective. Monthly Income was reported pre-tax.

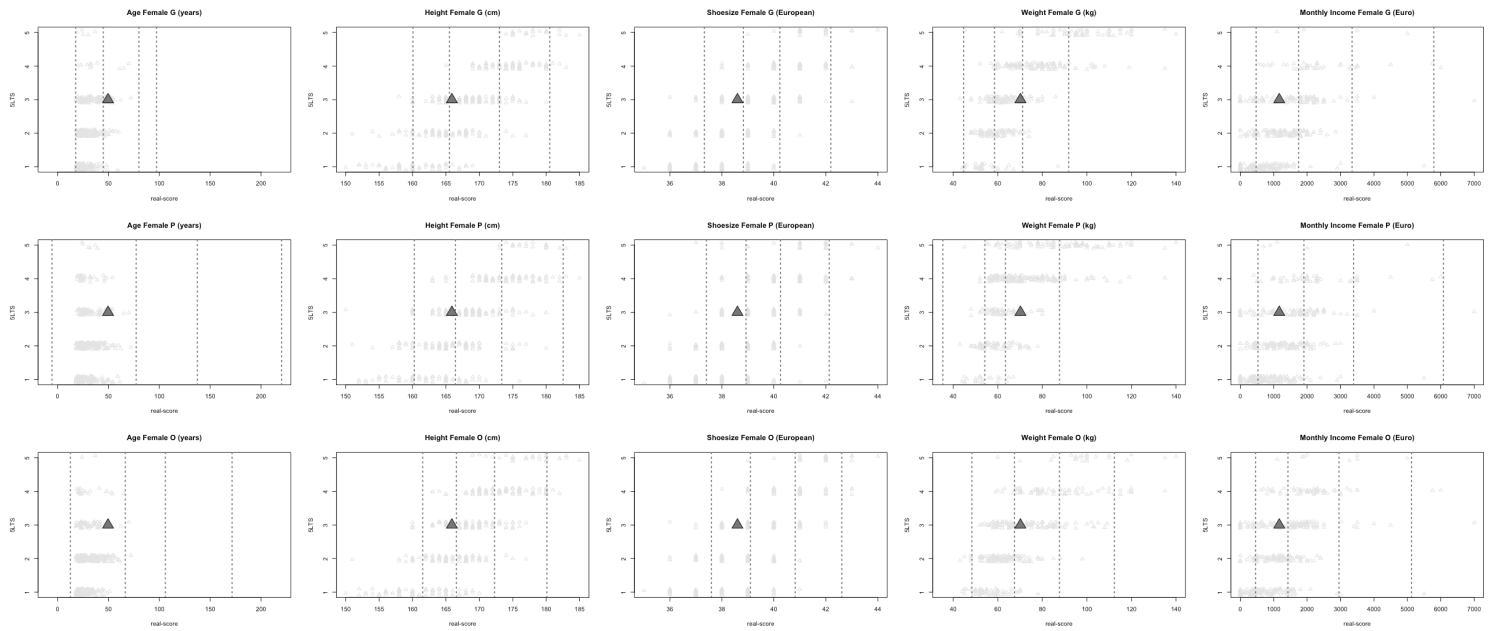

Figure 2: The thresholds of the proportional odds models of the 5LTSs for the five attributes and three frames of references are depicted for the female participants. 5LTS = five-point Likert-type scale; real-score = numerical value of the attribute; Grey triangle = Population mean of the real-scores for female subjects; Vertical lines = predicted thresholds between two categories; G/O/P = general/personal/outsiders' perspective. Monthly Income was reported pre-tax.

But what does that mean for the validity of rating scales in general? Obviously, three different questions were asked, and it appears only logical that they led to different patterns of answers. However, it can be assumed that all three questions are based on a truth value, as represented (even though restricted to a certain degree of granularity, given by the number of response options on the scale) by the real-score. The difference between these questions should therefore either represent different biases, as formulated in the T & B model, or, possibly, there is no such thing as a uniform latent construct independent of perspective or frame of reference. A potential solution that connects these two ideas is that the construct should be considered a combination of the truth value, here represented by the (numerical) real physical value, and the bias, here represented by the frame of reference. This view would be in accordance with the elaborations provided by West and Kenny (2011) and leave room for a concept of validity that could hold for psychological questionnaires.

Effect of Gender

The effect of the subjects' gender seemed to interact with the frame of reference induced by the phrasing, since the threshold pattern differed strongly and non-linearly between men and women in Study 1 (this notion is based on the descriptive distribution of the thresholds, as a statistical model test was not possible). In the framework of the T & B model, gender could therefore be considered a moderator variable, since it affected the influence of the bias induced by the phrasing. Thus, the frame of reference may be given by internal or external reference points (see Skaalvik & Rankin, 1995) and additionally moderated by the personal context of the respondent (including his or her gender).

Study 2

Rationale

One important particularity of the items used in Study 1 is the positive wording and that all LTSs included a central category. Regarding the first point, one could argue that referring only to one side of the dimension ("I am old") led respondents to adopt a categorical rating approach rather than a dimensional approach. For male subjects, the means of population the real-scores coincided with the centers of the scales more often than not, which provides the important information that the commonly applied positive wording of items does not necessarily lead to a shift in

the center of the dimension. However, this correspondence did almost never occur for female subjects, further highlighting the gender-related differences in self-perception (see Furnham et al., 2002) and suggesting that the way verbal anchors are interpreted may also interact with the respondent's gender. The second point, namely the central category, is also a subject of heated debate in psychometrics (e.g., Moors, 2008; Borgers, Sikkel, & Hox, 2004). This is, because the use of a central category in response scales has been reported to exhibit various effects on the respondents, such as the middle category endorsement (see Kulas & Stachowski, 2009).

To investigate if the positive phrasing of the items could be responsible for the inadequate representation of the real-life values on the scales, an additional study was conducted using neutral formulations for all items. In addition, a Likert-type scale with four response categories (4LTS) was used in addition to a 5LTS, to investigate possible effects of the central category.

Hypotheses

(i) As in Study 1, the continuous real-scores in the German population are not adequately partitioned into latent intervals representing the categories of the self-rating scales for all types of item phrasings.

(ii) The means of the population real-scores are expected to deviate from the middle category/center of the 5LTSs.

Sample

A total of n = 1769 participants (1405 female) participated in the study. Again, around 10% had started but not completed the questionnaire and did not enter the analysis. The age range was 18 – 77 years (quartile 1 = 24; median = 29; quartile 3 = 37) and 66 % had received 12 or more years of school education. As in Study 1, the participants were contacted via the social network Facebook or e-mails and received no remuneration for their participation.[2]

Method

Materials

An online questionnaire was created with the software SoSci Survey Version 2.6.00 (Leiner, 2014) and provided on the platform www.soscisurvey.de. The materials were similar to Study 1, except that the items were

[2] An additional sample of 586 participants filled out a dichotomous response scale with the same phrasings. The data of these additional subjects are available in the open science framework repository https://osf.io/4pcdb/.

worded neutrally with the response options on the LTSs represented by unmarked boxes, ending with “low” and “high” (German: “niedrig” and “hoch”). Also, only the general and the outsiders’ perspective were used for the items. The general perspective is the most standard way to formulate questionnaire items (see Goldstein & Hersen, 2000) and the outsiders’ perspective showed the least ambiguous results in Study 1. Moreover, shoe size was not included because of its coarse scale and restricted variance. The exact formulations for all items and all frames of reference are listed in *Appendix A.* About half of the participants (594) faced a 4LTS, the other half a 5LTS.

Procedure

The procedure was identical to the procedure in Study 1.

Analysis

The statistical computations followed exactly the analyses described in Study 1. All parameters of the regression analyses are depicted in *Appendix B.* The results of the likelihood ratio tests between proportional odds regression models with fixed intervals and models with freely varying thresholds are listed in *Appendix C.*

Results

Sample Distribution

One subject reported a monthly income of 50,000 Euro (30,000 Euro more than the second highest income and 26.6 *sd* above average) in combination with a self-rating of 3 on the 5LTS, which was almost certainly a mistake. It was decided to remove the outlier from all figures to retain an adequate scaling but to keep the outlier in the analyses, since this combination is improbable yet not impossible. The analysis can easily be redone without this subject, using the dataset and the analysis scripts available at https://osf.io/4pcdb/. *Table 2* provides an overview of the descriptive statistics.

Table 2: Descriptive statistics sample Study 2

Attribute	Mean	Median	SD	Min	Max
Women					
Age (years)	31.80	29	10.70	18	77
Height (cm)	168.10	168	6.46	150	192
Weight (kg)	69.67	65	17.26	38	170
Monthly Income (Euro)	1353	1200	1179.54	0	20000
Men					
Age (years)	32.80	29.5	11.34	18	75
Height (cm)	181.20	181	7.53	160	210
Weight (kg)	83.32	80.50	16.83	50	160
Monthly Income (Euro)	2120	1690	3205.64	0	50000

SD = Standard deviation; Min = Minimum value; Max = Maximum value; Monthly Income was reported pre-tax.

General Frame of Reference

Just like in Study 1, the thresholds of most items showed deviation from equidistance with the intervals differing more strongly for women compared to men. As depicted in *Figures 3* and *4*, the finding holds for the 4LTS and the 5LTS. The likelihood ratio tests (*Appendix C*) however, showed close to no significant results, thus indicating no deviation from equidistance. The relation of the means of the population real-scores and the intervals depended on the attribute: while age and income were shifted towards the left in all cases, height and weight showed means that generally coincided with the central category of the 5LTS and were close to the estimated threshold between categories two and three of the 4LTS. An exception are the ratings of women regarding their own weight, which tended to be too high compared to the real-score.

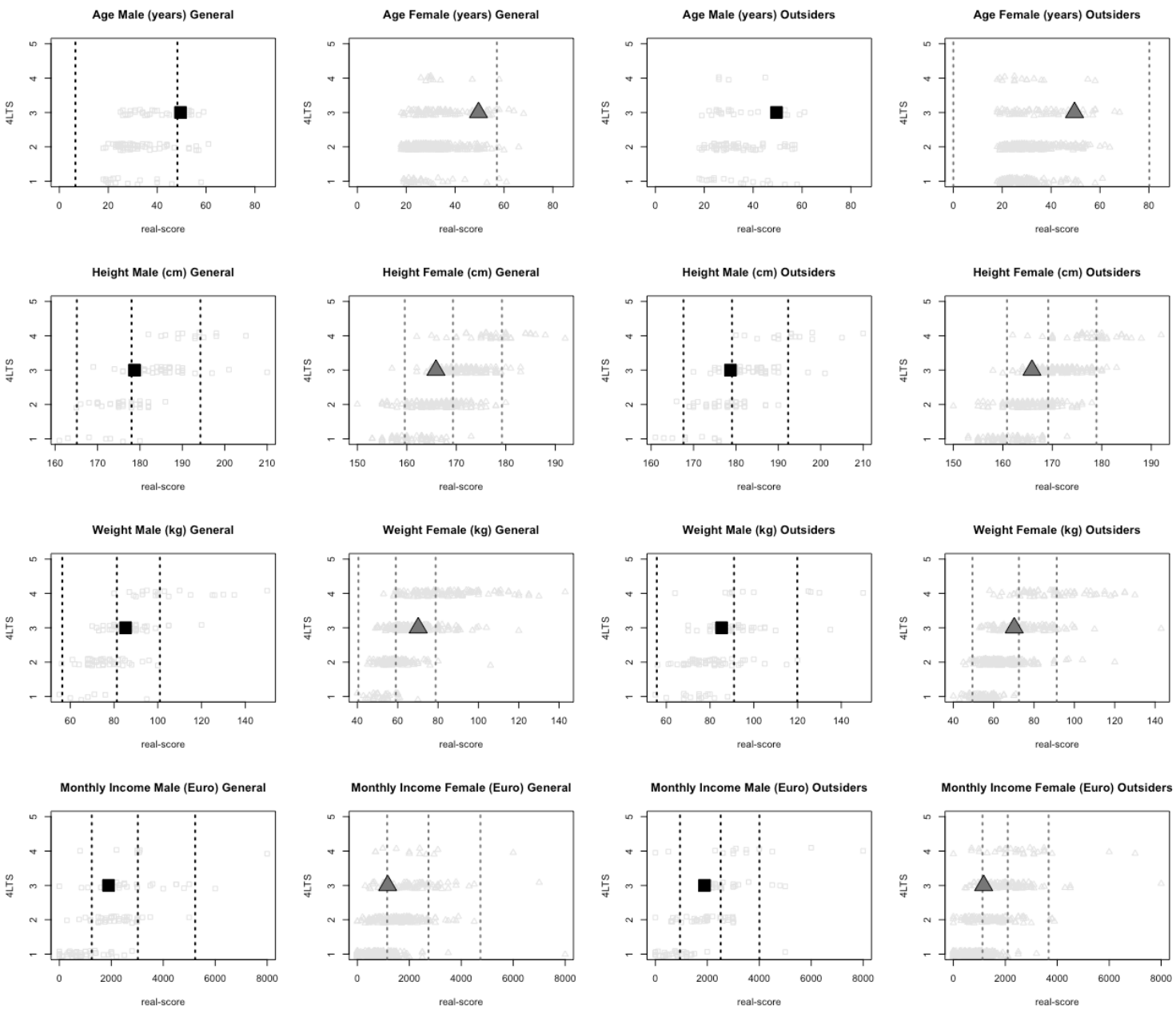

Figure 3: The thresholds of the proportional odds models of the 4LTSs for the four attributes and two frames of references are depicted for the male and female participants. 4LTS = four-point Likert-type scale; real-score = numerical value of the attribute; Black square = Population mean of the real-scores for male subjects; Grey triangel = Population mean of the reals-scores for the female subjects; Vertical lines = predicted thresholds between two categories; General/Outsider = general/outsiders perspective. Monthly Income was reported pre-tax.

Outsiders' Frame of Reference

Figures 3 and *4* show varying thresholds for all items on both the 4LTS and the 5LTS for the outsiders' perspective. For both men and women, the thresholds seemed to vary a little more compared to the general phrasing, which is supported by a greater number of significant likelihood ratio tests, as can be seen in *Appendix C*. Just as for

the general perspective, the means of the population real-scores were shifted to the left for age and income but generally in the central category of the 5LTS or close to the central threshold of the 4LTS.

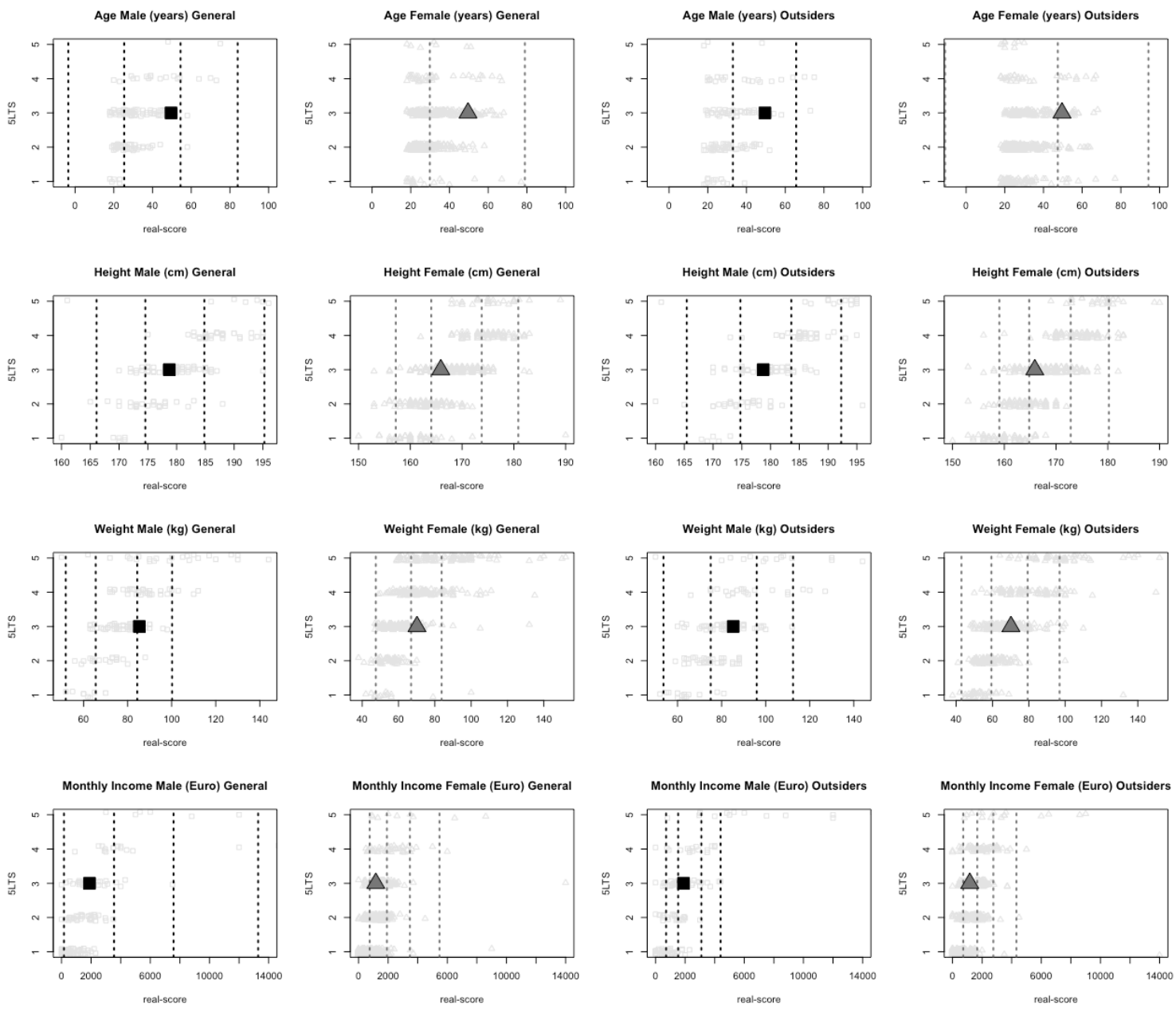

Figure 4: The thresholds of the proportional odds models of the 5LTSs for the four attributes and two frames of references are depicted for the male and female participants. 5LTS = five-point Likert-type scale; real-score = numerical value of the attribute; Black square = Population mean of the real-scores for male subjects; Grey triangel = Population mean of the reals-scores for the female subjects; Vertical lines = predicted thresholds between two categories; General/Outsider = general/outsiders perspective. Monthly Income was reported pre-tax.

Discussion

Study 2 was conducted to further investigate two possible confounds of Study 1: the use of positive wording and the use of response scales with central categories. The results show that the use of neutral wording does not change the overall pattern of results found in Study 1 but seems to have a desirable effect on the ratings of height and weight. Also, the use of the 4LTS, led to less divergence from equidistance compared to the 5LTS. The hypotheses were only partially supported: (i) The 5LTS model did not represent the mean intervals of the population real-score values adequately, but the 4LTS showed more equally distributed thresholds, and (ii) the means of age and income of the German population did not consistently coincide with the central categories of the 5LTS but they mostly did for height and weight.

It is interesting that 4LTS showed less divergence from equidistance than the 5LTS. It has often been argued that central categories may serve as a "dumping ground" for various types of responses to an item, such as unwillingness to share personal information (Kulas & Stachowski, 2009), yet an effect on the level of measurement has not been discussed thus far. However, providing a central category may lead to various effects in the response process (see Moors, 2008; Borgers et al., 2004), so the choice of an even or uneven number of response categories should be based on additional criteria. This is underlined by the finding that the 4LTS only worked somewhat adequately for two of the four attributes and that the lower number of significant likelihood ratio tests for both scales, compared to Study 1, may well be due to the smaller sample sizes (i.e. less power).

The neutral phrasing of the items was used because the positive verbal anchor of the standard phrasings could have been held responsible for the discrepancy between the middle categories of the response scales and the centers of the real-life values. This notion is not supported by the present findings, as the relation of the thresholds and the central moments of the 5LTS in Study 2 is remarkably similar to that in Study 1: age and income are shifted towards the left with height and weight mostly showing a correspondence of means and central categories with the exception of the self-perception of weight in women. This finding holds for the German population means as well as for the sample means (*Table 1*). Even though verbal anchors have been shown to be able to affect the distributional parameters of response samples (e.g., French-Lazovik & Gibson, 1984), such an effect could not be observed in the

present investigation. Thus, the positive phrasing does not seem to be responsible for the bias observed in the representation of the real-scores on the rating scales. So, this may serve as good news for most psychological questionnaires, which are phrased with a positive verbal anchor (see Goldstein & Hersen, 2000; Weijters & Baumgartner, 2012).

Finally, the outsiders' frame of reference did not result in more desirable results compared to the general perspective in Study 2, even though its central category comprises the German population mean more often. This puts the slight advantage of this phrasing in Study 1 into perspective, even more so as the general frame of reference seemed to actually work better with neutrally worded items.

General Discussion

The present results provide empirical support for the notion that parametric information of physical attributes like body height or weight is not preserved accurately by rating scales (see Jamieson, 2004): intervals between real-scores are not reflected when the values are being mapped onto a (self-) rating scale. Yet, the findings do not imply that the use of parametric methods is generally unwarranted for rating scales.

Notably, the present studies provide only empirical results that cannot resolve theoretical debates about the level of measurement and is limited to several aspects: since the scaling of the response formats is coarser than the scaling of all real-scores (e.g., height in cm vs. five response options on a LTS), it not possible for the rating scales to contain this information to the same degree of granularity (Clason & Dormody, 1994). Yet, as assumed by the classical test theory (see Lord et al., 1968), if the residual has a mean of zero and each person selects the rating scale value closest to the real-life value (on the same dimension), the centers, differences, and ratios of the real-life values should on average be reflected by those of the rating scales. This reflection may be distorted, since the intra-personal variances could differ, but then the difference should not show a systematic pattern on the between-person level. Yet, patterns of distortions show rather clearly in this investigation, for example through comparably small central categories.

Statistical inference is bound to the moments of samples and populations, meaning that it may for example be suggested that the means of two populations differ but not that every person of population one has a higher value than every person in population two. In the same way, intervals between ratings might not represent the intervals in the

real-scores but, assuming an unsystematic distortion due to a coarse rating scale (see Clason & Dormody, 1994), the mean intervals should. A systematic distortion, such as a bias, on the other hand, could indicate that the real-score is not purely measured by the item, which is in line with several investigations of self-perception (e.g., Furnham et al., 2002; Sinclair & Cheung, 2016). The item phrasings were varied in Study 1 to induce a difference in frame of reference, which is known to influence the validity of questionnaires (e.g., Bing et al., 2004; Schmit et al., 1995) and showed a strong impact on the way respondents in the present study rated themselves on 5LTSs.

The impact of removing the positive verbal anchors and applying neutral item wording did not lead to strong improvements in general but provided the most promising representation of the real-scores when combined with the 4LTS. However, note that the *p*-values of the likelihood ratio tests should be interpreted with care, since a total of 56 tests were conducted and the power in Study 2 is lower than in Study 1. Also, the fact that – unlike in Study 1 – the general perspective led to a slightly more accurate representation compared to the outsiders' perspective in Study 2 underlines the differential influence of the frame of reference. In combination with the various effects of item wording and verbal anchor points (see, e.g., French-Lazovik & Gibson, 1984), the picture of results implies that a pure measurement of the objective real-score may not be obtained using questionnaire items, at least for the attributes measured in the present series of studies.

However, not measuring the objective real-score does not necessarily undermine the value of self-ratings. Messick (1993), for example, argues that the classical definition of validity (dividing it into content, criterion, and construct validity) fails to take several aspects of item scores into account, such as the implications and the meaning of the score. As it has already been shown that perceived income (Sinclair & Cheung, 2016) and perceived body image (Mable, Balance, & Galgan, 1986) are more closely connected with stress than the objective measures, the self-ratings seem to indeed carry important implications and meaning. Taken together: the present findings do not imply that self-ratings are not of value but merely indicate that they may not serve as an unbiased measure for the objective construct even if this construct exists in the physical world. Given that a self-rating aims at reflecting the real- score (or true score in the classical testing theory), the item (or its phrasing and response options) seems to induce a bias, which is connected to the frame of reference for the same construct. This notion ultimately leads to the

question about the existence of a non-biased construct or if the measured construct is always part bias part truth. Following West and Kenny (2011) as well as Kunda (1990), human judgment is always guided by reality (i.e., a true value or real-score) but constrained by underlying motivations. The present study suggests that the frame of reference adds an additional bias and may be inseparable from the construct itself.

The fact that the central moments of the German population of the real-life values rarely coincide with the central categories of the scales could also be due to the fact that the raters do not compare themselves with the whole German population but probably with a smaller population close to their personal surroundings, independently from the induced frame of reference of the item wording. As depicted in *Tables 1* and *2*, the samples of all studies showed no strong deviations from the estimated German population mean except for age and income: the samples tended to be younger than the population mean. For an overview, *Appendix D* depicts comparisons of the central categories with the estimated sample means for all studies.

Limitations

The results obtained here are derived from a very small set of attributes and items and therefore need not hold for psychological questionnaires in general: typical questionnaires involve more than one item per construct, all of which have little in common with the items used in the present investigation. Nevertheless, the technique of self-ratings remains the same and the profound caveat raised here should be kept in mind. Also, an influence of the item presentation cannot be excluded: since all phrasings in Study 2 were presented next to each other, the respondents may have felt urged to vary their ratings.

Outlook

More interesting analyses could be undertaken with the present data, but the limited space of one article did not leave room to investigate all possible questions. Therefore, we uploaded all data sets (those presented in this article and several more) to the online repository. The interested reader is encouraged to download the data from https://osf.io/4pcdb/ and use it to gain further insight into the relationship between self-ratings and real-scores. To facilitate this, an overview and description of all data sets is given in *Appendix E*.

The present investigation sheds light on many particularities of ratings with response scales. The debate about the quality of measurement with rating scales, however, will go on, in particular with respect to multi-item scales and varying measurement models. We hope that the present findings contribute to the critical discussion about measurement issues in psychological science and can provide some vantage points for examining these topics with increasing detail.

References

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, *6*, 1-62.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: a replication and extension. *Journal of Applied Psychology*, *89*(1), 150.

Borgers, N., Sikkel, D., & Hox, J. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality and Quantity*, *38*(1), 17-33.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology*, *65*(3), 546.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505-514.

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them?. *Journal of personality and social psychology*, *101*(4), 831.

Chang, V. W., & Christakis, N. A. (2003). Self-perception of weight appropriateness in the United States. *American journal of preventive medicine*, *24*(4), 332-339.

Christensen, R. H. B. (2015). *ordinal—Regression Models for Ordinal Data. R package version 2015.1-21*.

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, *35*, 4.

Costa, P. T., & McCrae, R. R. (1989). NEO five-factor inventory (NEO-FFI). *Odessa, FL: Psychological Assessment Resources*.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement*, *8*(1), 49-57.

Furnham, A., Badmin, N., & Sneade, I. (2002). Body image dissatisfaction: Gender differences in eating attitudes, self-esteem, and reasons for exercise. *The Journal of psychology*, *136*(6), 581-596.

Goldstein, G., & Hersen, M. (2000). *Handbook of psychological assessment*. New York: Elsevier.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, *38*(12), 1217–1218.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*(3), 489-493.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Leiner, D. (2014). SoSci Survey [Computer program]. Version 2.6. 00.

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology*, *93*(2), 268.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology, 40*, 1–55.

Luce, R. D., Suppes, P., & Krantz, D. H. (2006). *Foundations of measurement: representation, axiomatization, and invariance* (Bd. 3). Courier Corporation.

Mable, H. M., Balance, W. D., & Galgan, R. J. (1986). Body-image distortion and dissatisfaction in university students. *Perceptual and Motor Skills*, *63*(2), 907-911.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, *1993*(2), i-18.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Michell, J. 1997, Quantitative science and the definition of measurement in psychology,

British Journal of Psychology, 88, 355-383.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, *42*(6), 779-794.

Paeratakul, S., White, M., Williamson, D., Ryan, D., & Bray, G. (2002). Sex, Race/Ethnicity, Socioeconomic Status, and BMI in Relation to Self-Perception of Overweight. *Obesity Research, 10*(5), 345–350.

Paulhus, D. L., Lysy, D. C., & Yik, M. S. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests?. *Journal of personality*, *66*(4), 525-554.

R Core Team (2018). R A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Richter, M. & Schaefer, K. (2009). *Der deutsche Fußreport 2009. Auswertung und Ergebnisse der jüngsten, bundesweiten Fuß- und Beinmessaktion*.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*(4), 735.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, *80*(5), 607.

Schwall, A. R., Hedge, J. W., & Borman, W. C. (2012). Defining age and using age-relevant constructs. *The Oxford handbook of work and aging*, 169-186.

Sinclair, R. R., & Cheung, J. H. (2016). Money Matters: Recommendations for Financial Stress Research in Occupational Health Psychology. *Stress and Health*, *32*(3), 181-193.

Skaalvik, E. M., & Rankin, R. J. (1995). A test of the internal/external frame of reference model at different levels of math and verbal self-perception. *American Educational Research Journal*, *32*(1), 161-184.

Stevens, S. S. (1946). *On the theory of scales of measurement*. Bobbs-Merrill, College Division.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge: Cambridge University Press.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological review*, *118*(2), 357.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737-747.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer Science & Business Media.

Appendix A: Item phrasings

Table A1
Items Study 1

English	German
General frame of reference	
I am old	Ich bin alt
I am tall	Ich bin groß
I have big feet	Ich habe große Füße
I weigh much	Ich wiege viel
I earn much	Ich verdiene viel
Personal frame of reference	
I consider myself old	Ich halte mich für alt
I consider myself tall	Ich halte mich für groß
I consider my feet big	Ich halte meine Füße für groß
I find that I weight much	Ich finde, dass ich viel wiege
I find that I earn much	Ich finde, dass ich viel verdiene
Outsiders‘ frame of reference	
Others consider myself old	Andere halten mich für alt
Others consider myself tall	Andere halten mich für groß
Others consider my feet big	Andere halten meine Füße für groß
Others find that I weight much	Andere finden, dass ich viel wiege
Others find that I earn much	Andere finden, dass ich viel verdiene

Table A1: English translations and German original phrasings of the items used in Study 1 for the general, the personal, and the outsiders' frames of reference.

Table A2
Items Study 2

English	German
General frame of reference	
How do you rate your age?	Wie schätzen Sie Ihr Alter ein?
How do you rate your income?	Wie schätzen Sie Ihr Einkommen ein?
How do you rate your height?	Wie schätzen Sie Ihre Körpergröße ein?
How do you rate your weight?	Wie schätzen Sie Ihr Körpergewicht ein?
Outsiders‘ frame of reference	
How do others rate your age?	Wie schätzen andere Ihr Alter ein?
How do others rate your income?	Wie schätzen andere Ihr Einkommen ein?
How do others rate your height?	Wie schätzen andere Ihre Körpergröße ein?
How do others rate your weight?	Wie schätzen andere Ihr Körpergewicht ein?

Table A2: English translations and German original phrasings of the items used in Study 2 for the personal, and the outsiders’ frames of reference.

Response formats of the response scales

Scales with positive verbal anchors

The five-point Likert-type scale comprised five boxes, marked with “völlig unzutreffend”, “unzutreffend”, “weder noch”, “zutreffend”, and “völlig zutreffend” (English: “completely incorrect”, “incorrect”, “neither”, “correct”, and “completely correct”).

Scales with neutral verbal anchors

The neutral four-point scale comprised four boxes, the ends marked with “niedrig”, and “hoch” (English: “low”, and “high”).

The neutral five-point scale comprised four boxes, the ends marked with “niedrig”, and “hoch” (English: “low”, and “high”).

Appendix B: Regression Coefficients

Table B1

Proportional Odds Regression for 5LTS (Study 1)

Male Subjects

	0|1	1|2	2|3	3|4	RS	SE	R^2
Age (years)							
G	16.790	41.310	65.029	104.282	.083	.015*	.031
P	2.407	53.936	98.181	171.332	.030	.013*	.006
O	6.570	41.438	73.364	111.861	.051	.013*	.008
Height (cm)							
G	166.999	175.317	183.159	194.062	.389	.051*	.175
P	168.257	175.986	184.668	197.189	.265	.040*	.142
O	169.161	175.668	184.113	194.018	.310	.043*	.163
Shoe Size (Europe)							
G	40.356	42.708	44.384	46.515	1.245	.170*	.064
P	40.085	42.983	45.187	47.716	.850	.144*	.049
O	40.355	42.431	44.463	46.997	1.047	.157*	.093
Weight (kg)							
G	62.374	76.369	87.907	117.167	.119	.016*	.227
P	56.922	71.980	83.600	113.399	.092	.014*	.172
O	66.019	79.047	95.021	128.960	.108	.015*	.193
Monthly Income (€)							
G	-832.239	1454.135	4823.395	11098.182	.001	<.001*	.089
P	-524.189	2143.922	5332.213	11743.110	.001	<.001*	.066
O	-3069.715	494.295	5039.970	11605.045	<.001	<.001*	.058

Female Subjects

	0|1	1|2	2|3	3|4	RS	SE	R^2
Age (years)							
G	17.759	44.825	79.922	97.282	.070	.010*	.082
P	-5.523	77.209	137.297	220.188	.022	.009*	.01
O	12.482	66.471	105.848	171.495	.038	.009*	.037
Height (cm)							
G	160.069	165.514	173.011	180.535	.407	.026*	.22

P	160.266	166.422	173.349	182.538	.305	.022*	.139
O	161.530	166.558	172.273	180.108	.377	.025*	.17
Shoe Size (Europe)							
G	37.326	38.829	40.231	42.193	1.427	.094*	.123
P	37.408	38.929	40.257	42.131	1.182	.083*	.075
O	37.598	39.101	40.821	42.614	1.272	.087*	.127
Weight (kg)							
G	44.680	58.559	71.135	91.853	.148	.010*	.172
P	35.363	54.204	63.444	87.707	.103	.009*	.102
O	48.388	67.481	87.683	112.327	.108	.008*	.178
Monthly Income (€)							
G	474.480	1747.114	3345.6564	5791.383	.001	<.001*	.127
P	528.991	1904.747	3390.474	6080.537	.001	<.001*	.103
O	456.765	1427.107	2954.054	5124.974	.001	<.001*	.106

Table B1: 5LTS = five-point Likert-type scale; x|y = Threshold between categories x and y; RS = Regression coefficient for the real-score; SE = Standard error of the regression coefficient for the real-life value; R^2 = McFadden's Pseudo R^2; G/P/O = general/personal/outsiders' perspective; * = $p < .05$; Monthly Income refers to the pre-tax income.

Table B2

Proportional Odds Regression for 4LTS (Study 2)

Male Subjects

	0|1	1|2	2|3	RS	SE	R^2
Age (years)						
G	6.564	48.278	-	.071	.019*	.074
O	-165.936	178.839	430.475	.008	.017*	.001
Height (cm)						
G	165.124	178.03	194.259	.259	.038*	.299
O	167.64	179.082	192.325	.271	.038*	.304
Weight (kg)						
G	56.334	81.299	100.98	.145	.022*	.287
O	55.666	90.982	119.879	.081	.014*	.146
Monthly Income (€)						
G	1247.66	3021.889	5227.168	.001	<.001*	.191
O	948.996	2515.426	4006.662	.001	<.001*	.232

Female Subjects

	0|1	1|2	2|3	RS	SE	R^2
Age (years)						
G	-20.221	57.073	110.585	.051	.010*	.033
O	.002	80.165	134.889	.355	.009*	.014
Height (cm)						
G	159.582	169.335	179.226	.341	.023*	.297
O	160.845	169.17	178.923	.355	.023*	.308
Weight (kg)						
G	40.532	59.072	78.796	.130	.010*	.229
O	49.471	72.497	91.25	.130	.009*	.250
Monthly Income (€)						
G	1146.612	2738.474	4737.78	.001	<.001*	.297
O	1121.467	2094.599	3668.124	.002	<.001*	.308

Table B2: 4LTS = four-point Likert-type scale; x|y = Threshold between categories x and y; RS = Regression coefficient for the real-score; SE = Standard error of the regression coefficient for the real-life value; R^2 = McFadden's Pseudo R^2; G/O = general/outsiders' perspective; * = $p < .05$; Monthly Income refers to the pre-tax income.

Table B3

Proportional Odds Regression for 5LTS (Study 2)

Male Subjects

	0|1	1|2	2|3	3|4	RS	SE	R^2
Age (years)							
G					.101	.017	.129
	-3.497	25.351	54.468	84.004			
O					.052	.014	.041
	-13.184	32.989	65.662	109.644			
Height (cm)							
G					.327	.04	.295
	166.089	174.568	184.823	195.265			
O					.33	.039	.287
	165.427	174.753	183.6	192.28			
Weight (kg)							
G					.13	.017	.230
	51.905	65.512	84.338	100.192			
O					.102	.014	.183
	53.609	75.050	95.940	112.423			
Monthly Income (€)							
G					<.001	<.001	.074
	165.322	3549.642	7572.082	13291.056			
O					.001	<.001	.260
	714.827	1531.253	3100.39	4399.112			

Female Subjects

	0|1	1|2	2|3	3|4	RS	SE	R^2
Age (years)							
G					.052	.009*	.036
	-30.946	29.826	78.948	115.717			
O					.039	.008*	.019
	-10.811	47.395	94.240	134.620			
Height (cm)							
G					.345	.022*	.280
	157.201	164.04	173.792	180.858			
O					.41	.025*	.334
	159.043	164.81	172.834	180.212			
Weight (kg)							
G					.115	.009*	.204
	31.065	47.499	66.949	83.737			
O					.131	.009*	.249
	42.928	59.411	79.383	96.943			
Monthly Income (€)							
G					.001	<.001*	.280
	754.273	1918.913	3472.5	5474.172			
O					.001	<.001*	.334
	724.771	1673.188	2760.564	4320.722			

Table B3: 5LTS = five-point Likert-type scale; x|y = Threshold between categories x and y; RS = Regression coefficient for the real-score; SE = Standard error of the regression coefficient for the real-life value; R^2 = McFadden's Pseudo R^2; G/O = general/outsiders' perspective; * = $p < .05$; Monthly Income refers to the pre-tax income.

Appendix C: Likelihood-ratio tests

Table C1

Likelihood Ratio Tests of Equidistant Thresholds for Male Subjects (Study 1)

	Phrasing	χ^2	*p*
Age			
	General	1.77	.41
	Personal	1.81	.41
	Outsiders‘	0.33	.85
Height			
	General	2.57	.28
	Personal	3.88	.14
	Outsiders‘	2.85	.24
Shoe Size			
	General	2.39	.30
	Personal	1.62	.44
	Outsiders‘	0.94	.62
Weight			
	General	15.83	**< .001**
	Personal	15.69	**< .001**
	Outsiders‘	14.55	**< .001**
Income			
	General	10.45	**.005**
	Personal	7.16	**.03**
	Outsiders‘	4.82	.09

Table C1: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders' = *Phrasing O*; χ^2 = χ^2-value; *p* = Probability of comitting a Type-1-error.

Table C2

Likelihood Ratio Tests of Equidistant Thresholds for Female Subjects (Study 1)

	Phrasing	χ^2	*p*
Age			
	General	6.93	**.03**
	Personal	5.77	.06
	Outsiders‘	5.73	.06
Height			
	General	9.46	**.009**
	Personal	9.50	**.009**
	Outsiders‘	12.46	**.002**
Shoe Size			
	General	6.98	**.03**
	Personal	5.91	.05
	Outsiders‘	2.04	.36
Weight			
	General	17.60	**< .001**
	Personal	50.37	**< .001**
	Outsiders‘	3.476	.18
Income			
	General	11.85	**.003**
	Personal	12.55	**.002**
	Outsiders‘	25.34	**< .001**

Table C2: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders’ = *Phrasing O*; $\chi^2 = \chi^2$-value; *p* = Probability of comitting a Type-1-error.

Table C3

Likelihood Ratio Tests of Equidistant Thresholds for Male Subjects for the 4LTS (Study 2)

	Phrasing	χ^2	*p*
Age			
	General	-	-
	Outsiders‘	1.57	.21
Height			
	General	1.47	.23
	Outsiders‘	.57	.45
Weight			
	General	1.47	.22
	Outsiders‘	.95	.33
Income			
	General	.79	.37
	Outsiders‘	.05	.83

Table C3: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders’ = *Phrasing O*; $\chi^2 = \chi^2$-value; *p* = Probability of comitting a Type-1-error; The log-likelihood (and therefore the model fit) for age with general phrasing was identical, so that no test could be conducted.

Table C4

Likelihood Ratio Tests of Equidistant Thresholds for Female Subjects for the 4LTS (Study 2)

	Phrasing	χ^2	*p*
Age			
	General	8.40	**<.01**
	Outsiders‘	7.51	**<.01**
Height			
	General	.02	.88
	Outsiders‘	2.79	.09
Weight			
	General	.33	.57
	Outsiders‘	4.08	**<.05**
Income			
	General	2.61	.11
	Outsiders‘	14.25	**<.001**

Table C4: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders’ = *Phrasing O*; $\chi^2 = \chi^2$-value; *p* = Probability of comitting a Type-1-error; The general models for age could not be compared to the equidistant model, because category 4 was not selected by any participant.

Table C5

Likelihood Ratio Tests of Equidistant Thresholds for Male Subjects for the 5LTS (Study 2)

	Phrasing	χ^2	*p*
Age			
	General	.01	1
	Outsiders‘	3.32	.19
Height			
	General	.98	.61
	Outsiders‘	.10	.94
Weight			
	General	1.71	.42
	Outsiders‘	1.29	.52
Income			
	General	2.91	.23
	Outsiders‘	7.39	**<.05**

Table C5: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders’ = *Phrasing O*; $\chi^2 = \chi^2$-value; *p* = Probability of comitting a Type-1-error.

Table C6

Likelihood Ratio Tests of Equidistant Thresholds for Female Subjects for the 5LTS (Study 2)

	Phrasing	χ^2	*p*
Age			
	General	8.67	**<.05**
	Outsiders‘	6.42	**<.05**
Height			
	General	13.02	**<.01**
	Outsiders‘	9.962	**<.01**
Weight			
	General	2.11	.35
	Outsiders‘	3.04	.22
Income			
	General	10.92	**<.01**
	Outsiders‘	9.257	**<.01**

Table C6: General = *Phrasing G,* Personal = *Phrasing P*, Outsiders’ = *Phrasing O*; $\chi^2 = \chi^2$-value; *p* = Probability of comitting a Type-1-error.

Appendix D: Comparison of central categories with estimated population means

Attribute	Study1 5LTS G	Study1 5LTS P	Study1 5LTS O	Study2 5LTS G	Study2 5LTS O	
Age (years)	Yes	No	No	Yes	Yes	49.56
Height (cm)	Yes	No	No	Yes	Yes	165.89
Shoe size (European)	No	No	No	-	-	38.60
Dress size (European)	-				-	-
Weight (kg)	Yes	No	Yes	No	Yes	70.17
Monthly Income (Euro)	No	No	No	No	No	1166.41
Age (years)	Yes	No	Yes	Yes	Yes	49.56
Height (cm)	Yes	Yes	Yes	Yes	Yes	178.72
Shoe size (European)	No	No	No	-	-	42.27
Dress size (European)	-	-	-	-	-	-
Weight (kg)	Yes	No	Yes	No	Yes	85.30
Monthly Income (Euro)	Yes	No	Yes	No	Yes	1891.61

Table D1: Estimated population mean = Estimated mean of the German population, according to the German General Social Survey, 2014, and the Deutscher Fußreport (German Foot Report), 2009; 5LTS = five-point Likert-type scale; G/P/O = general/personal/outsiders' perspective; Monthly Income refers to the pre-tax income; Yes/No = Central category comprises / does not comprise the estimated German population mean.

Appendix E: Datasets stored on repository https://osf.io/4pcdb/

Study 1: Three different frames of reference

- Five-point Likert-type scale
    - General, Personal, and Outsiders' frame of reference
    - With and without barplots indicating the population distribution

Study 2: Three response formats with two different frames of reference and neutral formulations

- Four-point Likert-type scale
- Five-point Likert-type scale
    - All response formats with General and Outsiders' frame of reference