Survey Research Methods (2026)
Vol. 20, No. 1, pp. 33-41
https://doi.org/10.18148/srm/2026.v20i1.8346
European Survey Research Association

Is there a Single Best Way how to Word a Typical English Agree-Disagree Scale in the German Language: Results from a Survey Experiment

Lukas Schick[1] · Dorothée Behr · Cornelia Neuert · Clemens Lechner
[1]GESIS Leibniz Institute for the Social Sciences

In a survey experiment, we analyzed how different versions of a response scale affect the distributional characteristics and quality of the resulting survey data. Toward that end, we compared four different German-language versions of a five-point agree-disagree (AD) response scale, randomly assigned to four groups of online access panelists for a total of 15 items. The response scales were taken from different studies and varied in polarity or scale option intensity. Comparisons of frequency distributions as well as of response quality (response styles, response differentiation, response times) did not show any systematic differences between the response scales. Although there were no systematic differences in the overall sample, the few significant effects we found appeared to be largely due to the responses of participants with lower levels of education. Further research is warranted using non-access panel respondents and their perception of differently worded AD response scales, experimentally modified response scales, and other languages beyond the English-German pair.

Keywords: agree/disagree scales; ordinal scales; response scales; scale labeling; scale translation

1 Introduction

Ordinal response scales are crucial components of most surveys. Such response scales map respondents' characteristics such as the strength of an attitude, the firmness of a belief or the importance of a value onto numbers, which then form the basis of quantitative data analyses. When designing response scales from scratch, or alternatively choosing existing response scales, researchers need to take important measurement decisions, amongst others regarding the polarity of a scale (bipolar vs. unipolar scale) and the intensity and distance between scale points (Dillman et al., 2014; Menold & Bogner, 2015). In cross-national research, to ensure comparability, translated response scales should ideally reproduce the measurement properties inherent in the source language instrument (Behr, 2023; Harkness et al., 2010). This, however, may not always be achieved.

Up until now, there is little research on the effects of differently worded or translated response scales that shift measurement properties to different degrees. Filling this gap, in this article, we assess the effects of four differently worded German-language agree-disagree (AD) scales on response distributions and on various dimensions of response quality.

1.1 AD Response Scales

The AD response scale is a prominent and frequently used ordinal response scale both in national and cross-national surveys (e.g., General Social Survey, GSS; International Social Survey Program, ISSP) as well as in standardized instruments with global reach (e.g., Big Five Inventory-2, BFI-2; Soto & John, 2017). Different item contents can be

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2026.v20i1.8346) contains supplementary material, which is available to authorized users.

Corresponding author: Lukas Schick, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany (Email: lukas.schick@gesis.org)

combined under the AD response format, which allows for an easy and efficient implementation of a set of diverse items (Dykema et al., 2022). Moreover, AD response formats may be part of established batteries or trend items, and hence the default format in a survey repeated at regular intervals (Dillman et al., 2014).

AD response scales have been criticized because of the extra burden placed on respondents who first have to process a given statement and then transform their judgment onto an AD response format. Questions with item-specific (IS) response options that directly focus on the construct of interest (e.g., 'How would you rate your health—excellent, very good, good, fair or bad?') have been put forward as an alternative that eases the response task for the respondent and improves response quality (Saris et al., 2010). While acknowledging the methodological debate on AD vs. IS response scales (Dykema et al., 2022; Höhne & Lenzner, 2018; Liu et al., 2015; Timbrook et al., 2021), in this article we focus on AD formats due to their continued wide-spread use. Other researchers keep on investigating AD formats, too, such as when comparing endpoint vs. completely labeled AD scales (Höhne et al., 2021), the ideal number of AD response options (Revilla et al., 2014), or the effects on AD/IS scales on different devices (Höhne et al., 2018).

1.2 AD Response Scales in Cross-National Studies

Valid findings in cross-national research depend on the comparability of questionnaire translations. The translation of AD formats is particularly challenging. Harkness (2003) and Harkness, Pennell, and Schoua-Glusberg (2004) showed different real-life approaches to translate AD scales into French, German, and other languages. Their illustrations show that questionnaire translations may change, for example, the intensity of modifiers (e.g., *strongly* shifting to *completely*), the scale polarity (e.g., turning a bipolar scale into a unipolar one), or the meaning of the midpoint. Such changes can result from structural and lexical differences between languages and resulting choices by translation teams, but they could also occur inadvertently because translation teams are not familiar with measurement properties of response scales, or because of intentional decisions by translation teams when preferring to use home-grown and familiar scales (Harkness, 2003) or those that better suit cultural conversational norms of the target country (Behr & Shishido, 2016; Shishido et al., 2009). In two long-standing cross-national survey programs, the same underlying English-language bipolar AD scale (*agree strongly—agree—neither agree nor disagree—disagree strongly*) is translated differently into German, using a bipolar wording in the German ESS version ('zustimmen'—'ablehnen'; 'agree'—'reject') and a unipolar wording ('zustimmen'—'nicht zustimmen'; 'agree'—'do not agree') in the German ISSP version; in the ISSP version, the modifiers at the end points have shifted to extreme modifiers ('voll und ganz'—'überhaupt nicht', being the equivalent of 'completely'), while the midpoint essentially stays the same across both versions and resembles a bipolar midpoint ('weder noch'; 'neither nor'). Do these differences, which shift—at least theoretically—important measurement properties, matter in practice?

There is little research on the effects of deviations in response scale translations (i.e., when certain measurement properties from the source instrument are not reproduced in a translation) or, more specifically, on the effects of differently translated response scales of the same underlying source scale. In one of the few publications on the topic, Villar (2009) investigated the absence/presence of a modifier in the second and forth label of the classical AD format in questionnaire translations of the ISSP. Adding a modifier in translations resulted in higher extreme response styles but did not affect acquiescence. Au et al. (2011), Mohler et al. (1998), and Rohrmann (2015) approached equidistance and comparability of scale labels using rating tasks (rather than producing a translation that mirrors the source in semantic terms), but their results have not carried over into translations used, for example, in the ESS or the ISSP.

Triggered by the aforementioned differences in frequently used German-language AD scales, scholars have tackled different German-language AD formats in research (fully-labeled vs. end-labeled; bipolar vs. unipolar, 5-point vs. 7-point, etc.). For example, a preference for positive categories (positivity bias) was found for a fully-labeled AD bipolar scale compared to a fully-labeled AD unipolar scale (Höhne et al., 2022; Menold, 2023); a moderate midpoint in a unipolar AD scale received more answers than its neutral or indifferent counterpart in the bipolar format (Höhne et al., 2022). Mismatching midpoints in unipolar scales (those that would rather fit to bipolar scales) equally provoked a positivity bias, when compared to a unipolar scale with a fitting midpoint (Menold, 2021; 2023). This research provides evidence that different verbalizations in AD formats can matter. It needs to be noted, however, that Höhne and colleagues as well as Menold did not replicate the exact ESS or ISSP AD wordings in their research, a gap which our present research addresses.

1.3 Research Question

Ultimately, we do not know how the different existing ESS and ISSP German translations of the same underlying English-language AD scale work, that is, versions that differ from each other in important scale characteristics. Previ-

ous research comparing different AD formats suggests that there may be differences in response distributions or response quality. Thus, our research questions are:

Do different existing German-language versions of AD scales affect respondents' response behavior and survey estimates? Are some versions less affected by undesirable response behaviors?

To answer these questions, we compare response distributions and common indicators of response quality i.e. *response differentiation* to examine straightlining between different scale versions; *response styles* to investigate whether different labeling of scale points leads to a bias in the selection of certain scale points; and *response time* as an indicator of the accessibility of different scale versions (e.g., Gummer & Kunz, 2021; Hylligus et al., 2014; Krosnick & Alwin, 1988).

In complement to the ESS and ISSP response scales, we added two further AD scales to this study, equally in use in German measurement settings, to make this study on potential effects more comprehensive. This allows us to compare several possible variants of the scale in terms of polarity, extreme endpoint labels, and the use of intensity modifiers.

As there is evidence from prior research that some respondents may respond differently to surveys in general (e.g., those with lower education or higher age show higher acquiescence on average, as do men; e.g., Lechner et al., 2019), and react differentially to linguistic differences in the survey stimuli (e.g., Rammstedt et al., 2023a), we also compare whether different subgroups responded differently to the AD translations.

In the following, we describe our data and experimental design. We then present the main findings and conclude with a discussion.

2 Data and Methods

2.1 Sample and Survey

This study uses data from a web survey experiment with 1010 respondents (approx. 250 per response scale) from Germany, which we conducted in December 2022. Respondents were recruited from a non-probability access panel provider (bilendi AG) and were based on quotas for gender, age (18–29, 30–49, 50–79), and education (with university entrance qualification, without university entrance qualification). In total, 49.9 % of respondents were female, 50.0 % had a university entrance qualification, and they were evenly distributed among the three age groups. Completing the entire questionnaire took an average of 8.5 min, with our response scale experiment positioned in the first half of the questionnaire.

2.2 Experimental Design

We compared four different 'real-life' German AD scales taken from the European Social Survey (ESS, 2020), the International Social Survey Programme (ISSP, 2021), the German Internet Panel (GIP; Blom et al., 2020), and one used for administering the Big Five Inventory 2 (BFI-2; German version by Danner et al., 2016)[1]; a short version of the established BFI-framework to measure personality, which is included in many surveys (Rammstedt et al., 2023b). All scales have five response options and are fully labeled. Table 1 shows the scale versions used per experimental condition and the corresponding labels in German (the language of this experiment) and in English. The German-language ESS and GIP scales are bipolar (using 'zustimmen'—'agree' and 'ablehnen'—'reject' as endpoints) while the German-language ISSP and BFI scales are unipolar (ranging from 'zustimmen'—'agree' to 'nicht zustimmen'—'do not agree'). In all scales, the third (middle) label indicates ambivalence or indifference towards the object. In addition, the GIP and BFI contain modifiers on the second and fourth label ('eher'—'a little'). Moreover, the ISSP, GIP, and BFI versions include an extreme endpoint ('voll und ganz', 'überhaupt nicht'), whereas the ESS versions follows a close translation of 'strongly' ('stark') on both sides of the scale.

We randomly assigned respondents to one of the four experimental conditions (see Table 1). Bivariate analyses of gender [χ^2 (3) = 3.14, p = 0.370], age [χ^2 (6) = 1.54, p = 0.956], and education [χ^2 (3) = 5.33, p = 0.149] showed an equal distribution of respondents to the four experimental conditions (see Table A1 in the Appendix for the sociodemographic sample composition by experimental conditions). Respondents then answered the same 15 items on various topics using the scale of the experimental condition to which they were assigned.

We aimed to select items that differed in terms of topic and sensitivity. We asked two behavioral (e.g., "In the past, information on the internet affected my health behaviour in a positive way") and three attitudinal questions on health information using the internet (e.g., "The internet is useful to help people decide if their symptoms are serious enough to go to the doctor") as well as three items on the assessment of doctors in general (e.g., "Doctors care more about their earnings than about their patients."), the latter of which we regarded as more emotional (ISSP 2021, Health

[1] The latter scale wording is also used in the German Longitudinal Election Study (GLES).

Table 1

Experimental scale versions (English source and German translations)

Condition 1: ESS (bipolar, less extreme end point labels, no modifiers for 2nd and 4th option)	Condition 2: ISSP (unipolar, extreme end point labels, no modifiers for 2nd and 4th option)	Condition 3: GIP (bipolar, extreme end point labels, modifiers for 2nd and 4th option)	Condition 4: BFI (unipolar, extreme end point labels, modifiers for 2nd and 4th option)
German			
Stimme stark zu	Stimme voll und ganz zu	Stimme voll und ganz zu	Stimme voll und ganz zu
Stimme zu	Stimme zu	Stimme eher zu	Stimme eher zu
Weder noch	Weder noch	Weder noch	Teils, teils
Lehne ab	Stimme nicht zu	Lehne eher ab	Stimme eher nicht zu
Lehne stark ab	Stimme überhaupt nicht zu	Lehne voll und ganz ab	Stimme überhaupt nicht zu
English			
Agree strongly	Strongly agree	Agree strongly	Agree strongly
Agree	Agree	Agree a little	Agree a little
Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neutral; no opinion
Disagree	Disagree	Disagree a little	Disagree a little
Disagree strongly	Strongly disagree	Completely disagree	Disagree strongly

The English version of the GIP-Scale is our own translation. The other English language versions are existing source scales. A literal and explanatory back translation (Son, 2018) can be found in the appendix (see Table A8)

and Health Care II). Three items that may elicit socially desirable responding were on attitudes towards homosexuals (e.g., "If a close family member was a gay man or a lesbian, I would feel ashamed"; ESS 2020, Round 10) and four items measured intellectual curiosity (BFI-2; Danner et al., 2016, e.g., "I am someone who is curious about many different things.").

The items were asked in an item-by-item format and the response scale was displayed horizontally below each item. Items of the same topic were asked on the same web page. An overview of all items used can be found in the Appendix (Table A2). In the analyses, (strong) agreement was coded as 1 and 2, the middle category as 3, and (strong) disagreement as 4 and 5.

2.3 Analysis

To analyze whether different German-language versions of AD scales affect response behavior and survey estimates, we compare frequency distributions of the responses (chi-square tests, adjusted residuals) and test whether the mean values differ significantly between the scale versions (ANOVA). Additionally, we examine the following different indicators of response quality:

Response Styles: Response styles are understood as the tendency to select certain response options in preference to other response options regardless of content. If the response styles differ depending on the scale version, this can lead to bias when comparing the data (Baumgartner & Steenkamp, 2001). In this study, the tendencies to select the end scale points 1 and 5 (*extreme response style*), the middle scale point 3 (*midpoint response style*) as well as the positive scale points 1 and 2 (*acquiescence response style*) and the negative scale points 4 and 5 (*disacquiescence response style*) are examined. Response styles are calculated by summing the number of corresponding response options for each respondent across all items. This is used to compare the mean of the number of corresponding response options by scale version.

Response Differentiation: Response differentiation indicates whether respondents are more likely to select different response options or more likely to select the same response option when answering the 15 items. It was calculated according to Linville et al. (1986) and ranges from 0 to 1. A value of 0 would mean that respondents always selected the same response option to answer all 15 items, while a higher value towards 1 would represent more variation in response selection (Krosnick and Alwin, 1988; Linville et al., 1986).

Response Times: Understanding and answering a question requires cognitive effort on the part of the respondent (e.g., Höhne et al., 2017; Couper & Kreuter, 2013; Tourangeau et al., 2000). To investigate whether there were differences in cognitive processing between the different scale ver-

sions, we compared the response time it took respondents to answer the items using Universal Client Side Paradata (Kaczmirek & Neubarth, 2014). To compare mean response times, we added the page-wise timestamps (in seconds) for the web pages that contained the items of our experiment. Outliers were excluded according to Höhne and Schlosser (2018), who recommend excluding outliers according to the method proposed by Hoaglin et al. (2000). That is, all respondents who had response times below/above the median plus/minus the upper and lower quartile range multiplied by 3 were excluded [Lower/Upper threshhold = $Q_{.50} \pm (3 \times (Q_{.50} - Q_{.25}))$]. The thresholds were calculated separately for each experimental group.

As outlined above, previous studies have demonstrated that socio-demographic characteristics can significantly impact response behavior. Moreover, since our study data stems from a non-probability sample and the generalizability of the findings has constraints, we conducted subgroup comparisons to examine whether the effects observed in the whole data set occur consistently in specific subgroups. This is seen as a valuable approach to evaluate the validity of the homogeneity assumption (Kohler et al., 2019).

We thus compared respondents with higher education (i.e., with university entrance qualification) and lower education (i.e., without university entrance qualification) as well as men and women and contrasted younger (18–39 years) with older respondents (40–75 years). The following analyses were carried out for these groups in the same way as for the overall sample: comparisons of mean values by item and topic as well as indicators of response quality.

3 Results

3.1 Total Sample

To investigate whether response behavior and survey estimates differ depending on the German version of the AD scale version, we compared frequency distributions and means of the 15 items. For four out of 15 items, we observed significant differences in frequency distributions (item 2: χ^2 (12) = 21.75, $p < 0.05$; item 7: χ^2 (12) = 29.91, $p < 0.01$; item 8: χ^2 (12) = 32.15, $p < 0.01$; item 9: χ^2 (12) = 28.06, $p < 0.01$). The items belong to different topics, except for items 7 and 8, which were both behavioral questions. A closer look at the distribution by using the adjusted residuals did not reveal a consistent pattern. For item 2 and item 7, it is the third scale point of the ESS that deviates to a greater extent; for item 9, it is the fifth scale point of the BFI; and for item 8, it is the fourth scale point of the BFI (see Table A3 in the Appendix).

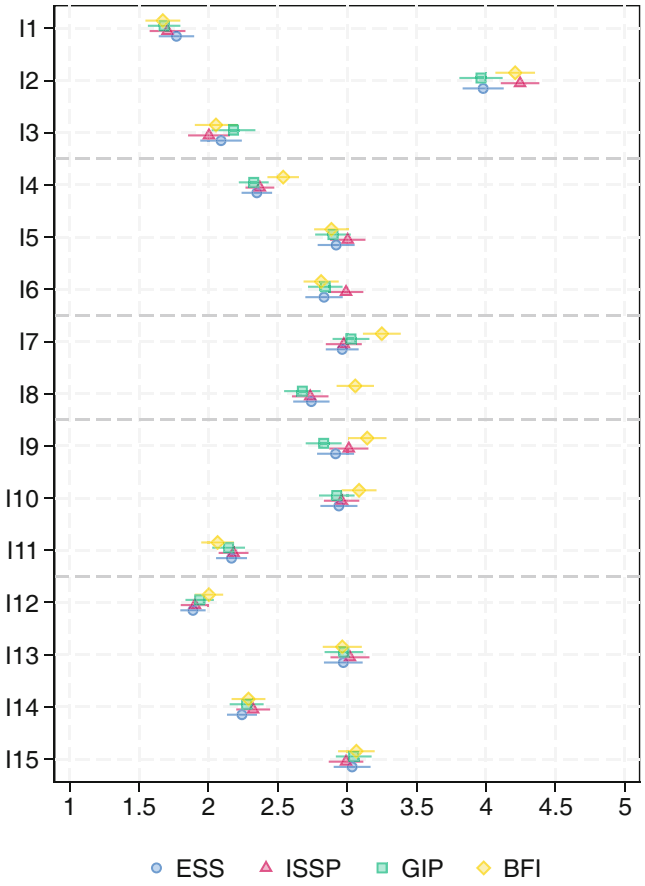

Fig. 1

Item means by scale version & item topic I' = Items. Item topics: Items 1–3 = Social Desirability, Items 4–6 = Emotions, Items 7–8 = Behavior, Items 9–11 = Attitude, Items 12–15 = Intellectual Curiosity. Error bars indicate 95% confidence interval

Fig. 1 below shows the mean values and 95% confidence intervals of the items by scale version and item topic. Upon closer inspection, no systematic differences by scale version (unipolar vs. bipolar, with modifiers vs. without modifiers) or item topic were observed.

Considering the item means, we found statistically significant differences for five items (items 2, 4, 7, 8 and 9; see Table A4 in the Appendix). For item 2, there is a statistically significant difference between the ISSP and the GIP scales, for items 4 and 9 between the GIP and BFI scales. For item 7, the ESS and ISSP scales differ from the BFI scale; and for item 8 the ESS, ISSP as well as the GIP scales differ from the BFI scale. Thus, four out of five significant differences occurred when comparing to the BFI response scale.

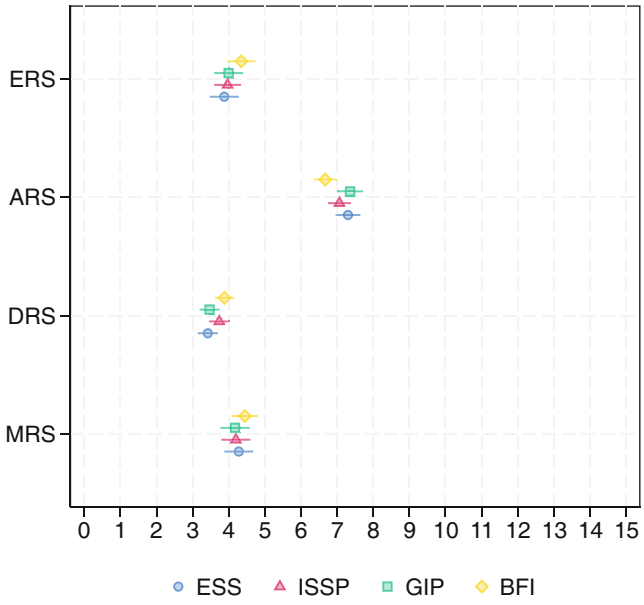

Fig. 2

Response styles means by scale versions ERS Extreme response style, scale points 1 or 5, ARS Acquiescence response style, scale points 1 or 2, DRS Disacquiescence response style, scale points 4 or 5, MRS Midpoint response stlye, scale point 3. The response styles are calculated by summing up the number of responses to the respective scale points across all 15 items. This results in a possible value range of 0 to 15 for each response style. Error bars indicate 95% confidence interval

Next, we compared response styles across scale versions (see Fig. 2). The scale versions did not differ in the extent of extreme, midpoint, and disagree responses. With regard to acquiescent responding, we found significant differences between the GIP and the BFI scale. Respondents answering the 15 items with the GIP scale (a bipolar scale) selected the response options 1 and 2 more often than respondents answering with the BFI scale (a unipolar scale).

Response differentiation was similarly distributed across the four scales (ESS: M = 0.64, SE = 0.01; ISSP: M = 0.65, SE = 0.01; GIP: M = 0.63, SE = 0.01; BFI: M = 0.67, SE = 0.01) and the overall difference was not statistically significant, F(1006) = 2.20, p = 0.08. That is, respondents differentiated their answers to a very similar extent.

There were also no significant differences in response times between the scale versions (F(905) = 1.61, p = 0.18). The mean values of the processing times were 95.51 s (ESS, SE = 3.15), 104.01 s (ISSP, SE = 3.21), 96.95 s (GIP, SE = 3.22) and 96.18 s (BFI, SE = 2.85).

3.2 Subgroups

Table A6 in the appendix shows the results of the single-factor ANOVA with the scale versions as the dependent variable, broken down for education, age, and gender subgroups. For the sake of comparability, only the significant differences between the scale versions within the subgroups are listed in the table

For respondents with lower education, there was a statistically significant difference in the mean values for item 2 between those who were assigned the BFI scale and those who were assigned the ESS scale or the GIP scale. In addition, there was a statistically significant difference between the mean values of the ISSP and GIP scale versions. Significant differences were also observed for Item 9 between the BFI scale and both the ESS and GIP scales. For items 7 and 8, the BFI scale differed significantly from all other scales in terms of mean values. For respondents with higher education, no statistically significant differences between the scales were found. For males and females, as well as for the young and older age groups, there were statistically significant differences observable for one or two items. Interestingly, these differences occurred mainly between the BFI scale and one of the other scales: for males (item 7), for females (items 4 and 5), as well as for the younger age group (item 7) and the older age group (item 9).

A similar pattern can also be seen when analyzing the response styles and quality indicators within the subgroups (see Appendix Table A7; again, only statistically significant differences are shown). Significant differences across response scales only appeared in the group with lower education. There is a significant higher level of disacquiescence for the BFI compared to the ESS and a significant lower level of acquiescence for the BFI compared do the ESS and GIP.

Overall, it is noticeable that significant differences between the different response scale versions appear almost exclusively in the group with low education, while in the other groups with higher education and in the age and gender groups the differences remain largely insignificant. Almost all significant differences between the four response scales in the overall sample are also found in the subgroup of low educated respondents, with the exception of item 4. In addition, the significance levels in this group are higher than in the overall data set.

4 Conclusion

The aim of the study was to determine whether differently worded (translated) German versions of fully labeled 5-point AD scales, currently used in survey programs or psychological tests, affect respondents' response behavior.

As discussed in the introduction, some previous studies observed effects on response behavior depending on wording and number of response categories (e.g., Menold, 2023; Höhne et al., 2022). This study explicitly focused on using existing German-language (translated) fully verbalized 5-point AD scales and comparing them with regard to frequency distributions and several response quality indicators. Based on the data of all respondents, the results showed only a few statistically significant differences in the frequency distributions of the different response scale versions, observed means, as well as several response style indicators across the scale versions. Furthermore, we found no statistically significant differences between the scale versions with regard to response differentiation and response time, the latter reflecting the cognitive effort used for answering by the respondents. However, zooming in on subgroups, we find that the few significant effects, as mentioned, were mainly triggered by low-educated respondents. At the same time, the low-educated group and the subgroup analysis show more clearly that the few significant differences primarily relate to the BFI scale. This could be due to a different understanding of the BFI scale within the educational groups, for instance due to the midpoint of the BFI scale. The BFI scale uses a different midpoint compared to all other scales ('teils/teils'—'partly this, partly that', in the source: "Neutral; no opinion.").

While being authentic, the tested response scales differed in more than just one element (e.g., polarity and modifier). Hence, the observed differences cannot clearly be linked to one component of the scales. Future research should take this up, for instance, by varying the label of the midpoint or polarity, while keeping everything else constant.

Apart from the "behaviour" of the BFI in this context, there were no consistent differences between specific scale versions and topics. For the ESS and the ISSP, our results are comforting since different German language translations of the same underlying English AD scale seem to be both fitting or at least not deviating from each other. The fact that there were almost no systematic differences between the four scale versions, although they did differ to some extent in the polarity and use of intensity modifiers, may be attributed to the fact that all scale versions had balanced positive and negative scale points and an identically (or similarly in the case of the BFI scale) worded midpoint. Our findings suggest that this balance of scale options and the existence of the mid-point led to respondents interpreting the scale options in a similar relationship to each other, despite different verbalization of the 5-point scale. An experimental design that, in addition to examining the response distributions, also includes the respondents' understanding, e.g., by asking for the ranking of different positive response options, could provide more detailed information about this. Furthermore, it is important to consider that the data in the study were collected from a non-probability access panel and therefore respondents may be familiar with surveys and labeled scales. This could result in these respondents perceiving the scales differently compared to respondents based on probability sampling. We encourage further research on whether our findings hold for other questions, within a survey of the general population, and also for other languages.

Last but not least, least, while in this paper we can assess the comparability of different German AD scales, we cannot draw any conclusions about comparability to the English source version or to other language versions; after all, we did not test this. We encourage further research within a language and across languages to advance knowledge on (challenges with) cross-national response scales (e.g., Axelsson & Dahlberg, 2024; He, Chi, and van de Vijver, 2021).

References

Au, W., Rohrmann, B., Taylor, P., Ho, J.M., & Yeung, S. (2011). Developing equivalent Chinese and English scale-point labels for rating scales used in survey research. *Asian Journal of Social Psychology*, *14*(2), 91–111. https://doi.org/10.1111/j.1467-839X.2010.01333.x.

Axelsson, S., & Dahlberg, S. (2024). Measuring happiness and life satisfaction amongst Swedish citizens: an inquiry into semantic equivalence in comparative survey research. *Journal of Happiness Studies*, *25*(8), 113. https://doi.org/10.1007/s10902-024-00827-7.

Baumgartner, H., & Steenkamp, J.-B.E.M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840.

Behr, D. (2023). *What to consider and look out for in questionnaire translation*. GESIS Survey Guidelines. https://doi.org/10.15465/GESIS-SG_EN_043.

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. Smith & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 269–287). SAGE. https://doi.org/10.4135/9781473957893.n19.

Blom, A.G., Fikel, M., Friedel, S., Höhne, J.K., Krieger, U., Rettig, T., & Wenz, A. (2020). *German Internet Panel, Welle 40 (März 2019) (Version 1.0.0) [Dataset]*. GESIS Data Archive. https://doi.org/10.4232/1.13463. SFB 884 'Political Economy Of Reforms', Universität Mannheim

Couper, M.P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in*

Society), *176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x.

Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. (2016). Die deutsche Version des Big Five Inventory 2 (BFI-2). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/ZIS247.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th edn.). Wiley.

Dykema, J., Schaeffer, N. C., Garbarski, D., Assad, N., & Blixt, S. (2022). Towards a reconsideration of the use of agree-disagree questions in measuring subjective evaluations. *Research in Social and Administrative Pharmacy*, *18*(2), 2335–2344. https://doi.org/10.1016/j.sapharm.2021.06.014.

European Social Survey. (2020). *ESS round 10 source questionnaire*. London: ESS ERIC Headquarters c/o City, University of London.

Gummer, T., & Kunz, T. (2021). Using only numeric labels instead of verbal labels: stripping rating scales to their bare minimum in web surveys. *Social Science Computer Review*, *39*(5), 1003–1029. https://doi.org/10.1177/0894439320951765.

Harkness, J. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). John Wiley.

Harkness, J., Pennell, B., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Wiley. https://doi.org/10.1002/0471654728.ch22.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115–140). Wiley. https://doi.org/10.1002/9780470609927.ch7.

He, J., Chi, R., & Van De Vijver, F. J. R. (2021). People use scales differently: dealing with survey response styles in cross-cultural research[1]. *Journal of Intercultural Communication & Interactions Research*, *1*(1), 83–100. https://doi.org/10.3726/jicir.2021.1.0006.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research* (pp. 219–237). Wiley. https://doi.org/10.1002/9781118763520.ch10.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (2000). *Understanding robust and exploratory data analysis*. Wiley.

Höhne, J. K., & Lenzner, T. (2018). New insights on the cognitive processing of agree/disagree and item-specific questions. *Journal of Survey Statistics and Methodology*, *6*(3), 401–417. https://doi.org/10.1093/jssam/smx028.

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata surveyfocus. *Social Science Computer Review*, *36*(3), 369–378. https://doi.org/10.1177/0894439317710450.

Höhne, J. K., Schlosser, S., & Krebs, D. (2017). Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field Methods*, *29*(4), 365–382. https://doi.org/10.1177/1525822X17710640.

Höhne, J. K., Revilla, M., & Lenzner, T. (2018). Comparing the performance of agree/disagree and item-specific questions across PCs and Smartphones. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *14*(3), 109–118. https://doi.org/10.1027/1614-2241/a000151.

Höhne, J. K., Krebs, D., & Kühnel, S. M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality. *Social Science Research*, *97*, 102544. https://doi.org/10.1016/j.ssresearch.2021.102544.

Höhne, J. K., Krebs, D., & Kühnel, S. M. (2022). Measuring income (in)equality: comparing survey questions with unipolar and bipolar scales in a probability-based online panel. *Social Science Computer Review*, *40*(1), 108–123. https://doi.org/10.1177/0894439320902461.

International Social Survey Programme (2021). *International social survey programme: health and health care II—final source questionnaire*

Kaczmirek, L., & Neubarth, W. (2014). *Universal Clientside Paradata (UCSP), version 2*. http://kaczmirek.de/ucsp/ucsp_ver2.html

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*(4), 526–538. https://doi.org/10.1086/269128.

Lechner, C. M., Partsch, M. V., Danner, D., & Rammstedt, B. (2019). Individual, situational, and cultural cor-

relates of acquiescent responding: Towards a unified conceptual framework. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 426–446. https://doi.org/10.1111/bmsp.12164.

Linville, P.W., Salovey, P., & Fischer, G.W. (1986). Stereotyping and perceived distributions of social characteristics: an application to ingroup-outgroup perception. In J.F. Dovidio & S.L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 165–208). Academic Press.

Liu, M., Lee, S., & Conrad, F.G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly*, *79*(4), 952–975. https://doi.org/10.1093/poq/nfv034.

MacDonald, P.L., & Gardner, R.C. (2000). Type I error rate comparisons of post hoc procedures for I j chi-square tables. *Educational and Psychological Measurement*, *60*(5), 735–754. https://doi.org/10.1177/00131640021970871.

Menold, N. (2021). Response bias and reliability in verbal agreement rating scales: does polarity and verbalization of the middle category matter? *Social Science Computer Review*, *39*(1), 130–147. https://doi.org/10.1177/0894439319847672.

Menold, N. (2023). Verbalization of rating scales taking account of their polarity. *Field Methods*, *35*(4), 378–391. https://doi.org/10.1177/1525822X231151314.

Menold, N., & Bogner, K. (2015). *Gestaltung von Ratingskalen in Fragebögen*. GESIS Survey Guidelines. https://doi.org/10.15465/GESIS-SG_015.

Mohler, P.P., Smith, T.W., & Harkness, J. (1998). Respondents' ratings of expressions from response scales: a two-country, two-language investigation on equivalence and translation. In J. Harkness (Ed.), *Cross-cultural survey equivalence* (Vol. 3, pp. 159–184). Zentrum für Umfragen, Methoden und Analysen -ZUMA.

Rammstedt, B., Roemer, L., & Lechner, C.M. (2023a). Do simpler item wording and response scales reduce acquiescence in personality inventories? A survey experiment. *Personality and Individual Differences*, *214*, 112324. https://doi.org/10.1016/j.paid.2023.112324.

Rammstedt, B., Roemer, L., Mutschler, J., & Lechner, C. (2023b). The big five personality dimensions in large-scale surveys: an overview of 25 German data sets for personality research. *Personality Science*, *4*(1), e10769. https://doi.org/10.5964/ps.10769.

Revilla, M.A., Saris, W.E., & Krosnick, J.A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73–97. https://doi.org/10.1177/0049124113509605.

Rohrmann, B. (2015). *Designing verbalized rating scales: Sociolinguistic concepts and psychometric findings from three cross-cultural projects (Roman Research Road)*. http://www.rohrmannresearch.net/pdfs/vqs-projects.pdf

Saris, W., Revilla, M., Krosnick, J.A., & Shaeffer, E.M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61–79. https://doi.org/10.18148/srm/2010.v4i1.2682.

Shishido, K., Iwai, N., & Yasuda, T. (2009). Designing response categories of agreement scales for cross-national surveys in east asia: the approach of the Japanese general social surveys. *International Journal of Japanese Sociology*, *18*(1), 97–111. https://doi.org/10.1111/j.1475-6781.2009.01111.x.

Son, J. (2018). Back translation as a documentation tool. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, *10*(2), 89–100. https://doi.org/10.12807/ti.110202.2018.

Soto, C.J., & John, O.P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096.

Timbrook, J., Smyth, J.D., & Olson, K. (2021). Are self-description scales better than agree/disagree scales? *International Journal of Market Research*, *63*(2), 201–215. https://doi.org/10.1177/1470785320971592.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The psychology of survey response*. Vol. 1. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322.

Villar, A. (2009). Agreement answer scale design for multilingual surveys: effects of translation-related changes in verbal labels on response styles and response distributions. In *Survey research and methodology (SRAM) program: dissertations and theses, 3*. https://digitalcommons.unl.edu/sramdiss/3.