Item Construction Rules Revisited: Learnings from Measurement of Latent Variables with Gold-Standard Items

Kathryn Eichhorn[1], Markus Bühner[2], Florian Pargent[2], Janika Saretzki[2,3,4], Larissa Sust[2], Jonas Hauck[5], and Sven Hilbert[5]

[1]Institute of Psychology, University of the Bundeswehr Munich

[2]Department of Psychology, LMU Munich

[3]Department of Psychology, University of Graz

[4]Charlotte Fresenius University of Psychology

[5]Faculty of Human Sciences, University of Regensburg

Author Note

Kathryn Eichhorn https://orcid.org/0000-0003-3676-9420

Markus Bühner https://orcid.org/0000-0002-0597-8708

Florian Pargent https://orcid.org/0000-0002-2388-553X

Janika Saretzki https://orcid.org/0000-0002-6536-8266

Larissa Sust https://orcid.org/0000-0002-3389-1626

Jonas Hauck https://orcid.org/0009-0002-8872-0530

Sven Hilbert https://orcid.org/0000-0001-5808-8357

This is version 1 (last modification 2025-08-05) of our preprint published on PsyArXiv. All materials (reproducible manuscript, analysis code, datasets, codebooks, questionnaires) are available in the project's repository on the Open Science Framework (OSF) at https://osf.io/p7492/. A Quarto Manuscript website is hosted at https://florianpargent.github.io/gold-standard-

items/. The data has previously been analyzed in the dissertation thesis of the first author at https://doi.org/10.5282/edoc.23628. As part of the dissertation, a preregistration was published at https://osf.io/cz3uv/?view_only=9cac02db231b48629fea9ae53c3038b9 that differs from the analyses reported in the current manuscript. The authors declare that they have no financial conflicts of interest in relation to the content of the article, and have no non-financial conflicts of interest.

Kathryn Eichhorn and Markus Bühner contributed equally to this manuscript. Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/): *Kathryn Eichhorn***:** conceptualization, investigation, resources, data curation, and writing – original draft. *Markus Bühner***:** formal analysis, supervision, and writing – original draft. *Florian Pargent***:** formal analysis, software, validation, and writing – review & editing. *Janika Saretzki***:** writing – review & editing. *Larissa Sust***:** writing – review & editing. *Jonas Hauck***:** writing – review & editing. *Sven Hilbert***:** conceptualization, supervision, and writing – review & editing

Correspondence concerning this article should be addressed to Markus Bühner, Department of Psychology, LMU Munich, Leopoldstr. 13, D-80802 Munich 80802, Germany, Email: buehner@lmu.de

Abstract

This article investigates whether gold-standard items are helpful for questionnaire construction. Based on an initial item pool (138 items) derived by a deductive item construction (N = 124) and prototype approach (N = 24), followed by cognitive interviews (N = 8) and a pilot study (N = 390) of a preliminary item set (61 items), three 12-item scales were constructed to measure the physical traits of body height, body weight, and age. We collected data on these scales using response formats with either two (N = 921) or six (N = 933) categories. We also collected numeric self-reports of body weight, body height, and age as gold-standard items. Confirmatory factor analyses revealed that the gold-standard items did not consistently exhibit the highest loading on their corresponding latent variable. Furthermore, when controlling for the self-reported physical body height, body weight, and age as gold-standard items, as well as gender, we did not always find an interpretable, systematic residual variance. Finally, the pattern of correlations between the latent variables did not reflect the correlations between the self-reported gold-standard items, suggesting that the item scales and the gold-standard items do not have the same validity. While these results are consistent with previous studies, our analyses also showed that items with two response categories were at least as valid as those with six categories, contradicting past recommendations. When constructing a questionnaire, we would argue that the items intended to measure the latent variable most directly should have the highest loading on that variable. If this is not the case, the content validity is questionable at best. The implications are that intensive cognitive pretesting is necessary. Questionnaires with different response formats should be compared empirically and hypotheses about which items best represent the latent variable should be tested.

Item Construction Rules Revisited: Learnings from Measurement of Latent Variables with Gold-Standard Items

Psychological researchers typically develop questionnaires and use latent variable modeling to test whether they measure the intended psychological construct (Holt et al., 2010). According to the most applied test models (e.g., factor analytic models) especially item discrimination parameters are interpreted as how well items correlate with the latent variable (see Lord et al., 1968). Highly correlated items with the latent variable indicate that these items are more important to measure the latent variable compared to items with lower correlations. Thus, the item discrimination parameter is typically used to assess the quality of an item measuring the latent variable.

To assure that we obtain high item discrimination parameters it is essential to start with a precise definition of the latent variable. The definition is the starting point for deriving items with content validity. The item construction is believed to be theoretically sound if items are selected from a larger population of possible indicators of the latent variable in a representative manner to achieve high content validity and ensure that we measure the indented construct. Let us assume that we try to measure the latent variable *interpersonal warmth* with the following adjectives: *affectionate*, *friendly*, *talkative*, *unprejudiced*. We ask how well each adjective describes yourself from *not at all* (0) to *fully and entirely* (4) with five response categories. If these items measure warmth, they should at least all significantly correlate with the adjective warmth measured with the same response format. Even more, the warmth item should have the highest discrimination parameter conducting a factor analysis with empirical data, otherwise a different construct is measured, e.g., friendliness, in the case that the adjective *friendly* has the highest loading on the latent variable. This can be seen as a robustness check that the latent variable has the intended meaning. In the following sections, we refer to such items as *gold-standard items*.

In this study, we construct scales for the "physical traits" *body height*, *body weight*, and *age* and compare them with a gold-standard item, that is, the numerical self-report of the respective physical property. We derive three assumptions that should hold if a valid measurement of the mentioned latent variables is possible. We base these assumptions on the relevant literature on previous studies, item wording, and response categories, which we will outline in the next section.

Theoretical Background

Evidence Using Gold-Standard Items

To demonstrate the validity of psychological measurements and the practical utility of psychological measurement models, Bortolotti et al. (2013) compared psychological measurement with physical measurement. They constructed a scale with 27 items intended to assess body height. Using a response format with two response categories, participants were asked to respond to items such as "Do I think, I would do well in a basketball team?" (Bortolotti et al., 2013, p. 2350). At the end of the scale, participants were asked to report their actual height in centimeters, the self-reported physical gold-standard item. The parameters estimated from the latent variable body height were then compared with the physical, gold-standard measurements. The results revealed a strong correlation (.86, not corrected for disattenuation) between the estimated person parameters derived from a two-parameter logistic model and the self-reported physical height. Still, the gold-standard item did not exhibit the highest loading on the scale (Bortolotti et al., 2013). Van der Linden (2016) also employed a height measurement based on items with two response categories (e.g., "I bump my head quite often.", p. 27) to illustrate the usefulness of logistic item response models. According to the author, this approach is unusual, as body height appears to be a physical variable that can only be measured using a yardstick. However, differences in height influence behavior, making it quite plausible to view such behavioral indicators as proxies for body height and to derive a psychological measurement from

them. The two examples suggest that, first, gold-standard items do not always have the highest loading on the latent variable. Second, measuring a "physical" latent variable might be a psychological representation of body height rather than a measure of physical body height.

Item Wording

One important aspect of item construction is item wording. In this context, Pargent et al. (2019) and Goretzko et al. (2020) showed that the quality of item wording did, in fact, not impact model fit based on commonly used fit indices of confirmatory factor analyses. Sound item wording may not be crucial for achieving good model fit, but it clearly is for understanding the item correctly: For example, if a person does not correctly recognize an item with negative polarity, the person's value on the scale is immediately over- or underestimated by a few points. In some cases, this can lead to considerable distortions in the test score. If a person does not understand the items correctly, their response cannot be valid, regardless of any model fit indicators (for an extreme example, see Maul, 2017). Thus, it would be problematic to evaluate the quality of a questionnaire by solely looking at the model fit of the latent variable model, the discrimination parameters and/or reliability estimates. In this context, McNeish et al. (2018) stated that good fit indices (e.g., Root Mean Square Error of Approximation, RMSEA; Standardized Root Mean Square Residual, SRMR; and the Comparative Fit Index, CFI) must be considered together with measurement quality (reliability) and the size of standardized factor loadings. They showed examples where all fit indices were excellent, but the standardized item loadings were .40, and another model where the fit indices were inadequate, but the factor loadings were .90. Thus, fit indices can be misleading without considering the quality of measurement.

Response Scale

Hilbert et al. (2016) presented initial evidence that different response scales might not measure the same latent variable. They found that the correlation (disattenuated for measurement

error) between latent variables measured by identical items ranged only from .76 to .88 when using items with two versus five response categories and a visual analogue scale. The question arises: Which of these questionnaires measures the true construct? What is well known is how the number of response categories affects reliability estimates. A vast amount of literature compares different numbers of response categories for rating scale items. We quote some exemplary studies to demonstrate how complex and difficult it is to choose a reasonable number of response categories and highlight what we can conclude from their results. Preston and Colman (2000) reported that two response categories exhibited the lowest retest-reliability and a relatively poor validity and discriminating power. They recommend applying rating scales with at least seven response categories. In contrast, Lee and Paek (2014, p. 663) concluded that "caution should be made if a scale has only two response categories, but that limitation may be overcome by manipulating other scale features, namely, scale length or item discrimination". Weijters et al. (2010) recommend rating scales with five fully labeled response categories for samples drawn from the general population. In their recommendations, the authors considered an extreme response style, responses to reversed items, and an acquiescence response style. However, Masuda et al. (2017) found that low-motivated persons tend to use the middle category, implying that a middle category is not advisable. Similarly, Bradley et al. (2015, p. 8) concluded from their study that "for constructing measures from survey responses, the inclusion of a neutral middle category distorts the data to the point where it is not possible to construct meaningful measures." All in all, the reviewed literature suggests that to achieve high reliability, seven response categories seem the be useful, but there is also evidence to omit the middle category due to the inconsistent use of it. Thus, we infer that there is evidence to use rating scales with six response categories without a middle category.

Implications

We quote from a review on the application of Item Response Theory (IRT) models and factor analyses (Holt et al., 2010, p. 288): "In our opinion, researchers could take far better advantage of their theoretical knowledge and/or expectations by incorporating their a priori knowledge of the items and scales in the analyses. This should be reflected 1) by a more frequent application of confirmatory techniques, especially in the construction of new scales and 2) by adding interpretability of factors and content of items to the criteria used for model evaluation." Building on this summary statement, the aforementioned research and recommendations, we assess three physical "traits" to replicate and extend previous findings. First, we constructed scales for body height, body weight, and age. Second, we used items with two response categories to replicate the results of Bortolotti et al. (2013) for body height. Third, we applied an additional format containing six response categories without a middle category which low-motivated persons tend to overuse (e.g., see Masuda et al., 2017). We aimed to investigate whether 1) the results of Bortolotti et al. (2013) hold true for body height and similar physical traits, and whether 2) there really is a psychological component in self-report questionnaires of physical traits as assumed by Van der Linden (2016). Furthermore, we assumed that items with six response categories have higher reliability and thus, higher loadings and better validity than those with two response categories (Lee & Paek, 2014; Preston & Colman, 2000; Weijters et al., 2010). The concrete assumptions underlying our research are the following:[1]

- *Assumption 1:* From a theoretical point of view, the self-reported physical body height, body weight, and age as the self-reported gold-standard items should have the highest factor loading (which is equivalent to a high correlation between the latent variable and the item applying a

[1]No analyses reported in the current manuscript were preregistered. However, the data presented in this manuscript has previously been analyzed in the German dissertation thesis of the first author (Eichhorn, 2019). As part of the dissertation, a preregistration was published at https://osf.io/cz3uv/?view_only=9cac02db231b48629fea9ae53c3038b9 that differs from the analyses reported in the current manuscript. The results of the preregistered analyses are reported in Eichhorn (2019).

one-factor model) compared to other questionnaire items measuring the respective physical trait.

- *Assumption 2:* The measurement of physical traits with questionnaire items should have a psychological component because the self-reported physical measurement and gender (as there are gender differences in perception, see McCreary, 2002) cannot fully explain the item responses. Systematic correlated residuals should remain.
- *Assumption 3a, b:* The correlational pattern between the latent variables for the physical traits should correspond to the correlation of the self-reported gold-standard items (3a) and the factor loadings and reliability estimates should be higher for scales with 6-point response categories compared to 2-point response categories (3b).

Methods

All materials (reproducible manuscript, analysis code, datasets, codebooks, questionnaires) for this manuscript are openly available on the Open Science Framework (OSF) at https://osf.io/p7492/ (Eichhorn et al., 2025). A Quarto Manuscripts website presenting the manuscript and electronic supplementary materials is hosted at https://florianpargent.github.io/gold-standard-items/.

Item Construction

First, we constructed items for the physical characteristics of body height, body weight, and age. Age was defined as chronological age or age in years, corresponding to the time elapsed since a person's birth (Montepare & Lachman, 1989; Schwall, 2012). Height was defined as the height of an upright person from the sole of the foot to the top of the head in centimeters, and body weight as the physical mass of a person in kilograms (Martin, 1929). The initial items were developed by 124 psychology students (deductive item construction, Burisch, 1984) and 24 persons (prototype approach, see Broughton, 1984) who deviated from the German population mean at least one standard deviation in the relevant characteristics (separately for women and

men). The 24 persons were asked to think of prototypical behaviors for each construct to assess the whole spectrum of the latent variable. This process resulted in an initial set of 138 items (response categories were not tested). In a second step, we examined how these items were interpreted and whether they were connected to additional concepts unrelated to the constructs using cognitive interviews. We conducted interviews with a length of two to three hours and applied various cognitive procedures, including probing, paraphrasing, concurrent-think-aloud, and retrospective-think-aloud (Prüfer & Rexroth, 1996, 2005). The literature on cognitive interviews recommends five to 30 interviews but indicates that the most serious problems can already be identified with a small number of interviews (Willis, 2005). The interviews were conducted with 8 people from the target group (4 women and 4 men) aged between 21 and 77 (M = 44.50, SD = 20.78). The level of education ranged from high school diploma to university degree. After the cognitive interviews, misleading items were reformulated or eliminated (Faulbaum et al., 2009), resulting in 61 remaining items. In a third step, these 61 items were pilot tested online on a new sample, which initially consisted of 456 people. However, we excluded 66 subjects because they had not completed the questionnaire, resulting in a sample of 390 participants aged between 18 and 77 years (M = 31.91, SD = 12.82), including 302 women (77.44 %) and 88 men (22.56 %). The level of education was distributed as follows: 0 % no school leaving certificate, 3.33 % secondary school leaving certificate/elementary school or equivalent, 17.18 % secondary school or equivalent, 35.90 % vocational baccalaureate or high school diploma, 40.77 % college degree or university degree, 2.82 % doctorate or habilitation. Out of all respondents, 95.38 % stated German as their native language. Based on these survey responses, we excluded items if a) their main item loading was not on the intended factor, b) their main and secondary loadings were almost equal, or c) they had loading below or equal to .30. Additionally, if (d) items had very similar content, items with lower loadings were excluded. More details on the process of item construction can be found in Eichhorn (2019). We

reproduced the exploratory factor analyses, item total correlations and reliability estimates that were computed in the final item selection stage and report them in our electronic supplementary materials. Based on these criteria 12 final items were selected for each physical trait (see Table A1).[2] For each of the three questionnaires, two versions were created – one with a 6-point scale combining verbal endpoints ("does not apply" to "applies", only endpoints were labeled) with full numeric anchoring (numbers from 1 to 6) and one with a 2-point scale combining verbal labels ("does not apply" and "applies") with numeric anchors (numbers 0 and 1).

Statistical Analysis

All analyses were conducted in R version 4.5.1 (2025-06-13). A full list of software versions can be found in Appendix C. The data were analyzed using confirmatory factor analyses using the package *lavaan* (Rosseel, 2012). We used the weighted least squares, mean and variance adjusted (WLSMV) estimator, treating the questionnaire items as ordinal and the self-reported physical gold-standard items body height, body weight, and age as continuous variables. First, we specified (1) unidimensional models without any error correlations, and (2) models with one latent variable and correlated errors. In both models the gold-standard items and the other 12 items were specified as indicators of a single latent variable. Second, to test whether there is a latent variable while controlling for the gold-standard item and gender, (3) models were specified with a single latent variable where the questionnaire items were additionally predicted by the gold-standard item and gender. Third, to compare the correlations between the latent variables with the correlations between the gold-standard items, (4) full three factor models without correlated errors were estimated, in which only the questionnaire items load on their respective latent variable and correlations are estimated between the latent variables, between the gold-standard items, and between the latent variables and the gold-standard items. We also conducted

[2]Compared to the pilot test, one height and one weight item were slightly reworded to simplify the wording and reduce similarities with other items in the scale.

reliability analyses using the package *MBESS* (Kelley, 2023) and report McDonald's Omega as a reliability estimate of internal consistency for each scale.

Sample

The total sample of our main study consisted of 1854 participants aged between 18 and 97 years ($M = 32.70$, $SD = 16.01$), including 1036 women (56%) and 818 men (44%). The overall sample consisted of two independent subsamples, which were divided into the 2-point ($N = 921$) and 6-point ($N = 933$) response format conditions.

Participants were recruited in trains, colleges, universities, adult education centers, fitness studios and various public places in Germany and Austria in November and December 2017. Participation was voluntary and without reimbursement. Each participant was given a piece of paper informing them about the voluntary and anonymous nature of their participation, the purpose of the study and the data to be collected. When they gave consent, the participant was handed a paper and pencil questionnaire, alternating between the 2-point and 6-point response scale versions. When the participant handed back the questionnaire, the facilitator immediately checked for missing values, in which case the participant was politely asked to complete the missing questions. With this procedure, we achieved complete responses for all questionnaires. No power analysis was performed to determine the sample size. Instead, the goal was to collect at least 500 subjects per response format condition over a period of six weeks. No data was analysed before data collection ended.

Table 1 shows the distribution of gender, level of education, native language, and age within the two conditions. As expected based on the randomization procedure, the two subsamples are very similar. Due to the left-skewed age distribution and outlier values, we report both the mean and the median.

Table 1

Sample characteristics within response category conditions

		Two categories (N=921)	Six categories (N=933)
Age	Mean (SD)	32.8 (15.9)	32.6 (16.1)
	Median (IQR)	25.0 (23.0)	25.0 (21.0)
	Range	18.0 - 89.0	18.0 - 97.0
Gender	Women	512 (55.6%)	524 (56.2%)
	Men	409 (44.4%)	409 (43.8%)
Education	No school leaving certificate	1 (0.1%)	2 (0.2%)
	Secondary school leaving certificate/elementary school or equivalent	35 (3.8%)	43 (4.6%)
	Secondary school or equivalent	129 (14.0%)	119 (12.8%)
	Vocational baccalaureate or high school diploma	446 (48.4%)	437 (46.8%)
	College degree or university degree	280 (30.4%)	311 (33.3%)
	Doctorate or habilitation	30 (3.3%)	21 (2.3%)
Mother tongue	German	837 (90.9%)	860 (92.2%)
	Other	84 (9.1%)	73 (7.8%)

Note. SD = standard deviation; IQR = interquartile range.

Results

Assumption 1: The self-reported physical body height, weight, and age should have the highest loadings on their latent variables.

Table 2 displays the measurement models for the three scales for body height, body weight, and age for two and six response categories each.

Table 2

Standardized item loadings (correlations with the latent variable) for two and six response categories

Measure	Height		Weight		Age	
Categories	Two	Six	Two	Six	Two	Six
			Items			
1	.780	.611	.564	.564	.430	.493
2	.816	.829	.456	.503	.668	**.756**
3	.786	.740	**.853**	**.822**	.671	.535
4	**.883**	.851	**.931**	**.941**	**.764**	**.738**
5	.778	.742	**.949**	**.902**	**.786**	**.774**
6	**.877**	.817	**.886**	**.897**	.129	.230
7	.843	.808	**-.778**	**-.749**	**.921**	**.837**
8	.623	.721	-.649	-.619	.586	.571
9	**-.887**	-.835	-.420	-.421	**.795**	.700
10	-.843	-.719	.613	.683	**-.810**	-.681
11	**-.877**	-.818	.245	.261	.453	.489
12	**.876**	.815	.696	**.751**	.698	**.739**
Physical item	.873	.863	.698	.690	.759	.719
			Model fit			
χ^2	423.713	1,113.389	394.701	1,003.750	69.371	1,647.923
df	65	65	65	65	65	65
p	< .001	< .001	< .001	< .001	< .001	< .001
CFI	.971	.936	.950	.948	.880	.830
RMSEA	.077	.132	.074	.124	.102	.162
SRMR	.091	.069	.086	.069	.104	.105

Note. CFI = Comparative Fit Index (scaled); RMSEA = Root Mean Squared Error of Approximation (scaled); SRMR = Standardized Root Mean Residual. Scaled χ^2, scaled df (degrees of freedom), and scaled p values are reported. Correlations with the latent variable higher as the self-reported physical item are printed in bold. The standardized item loadings for the self-reported physical items are presented in a grey shaded box.

Overall, both response formats produced similar loading patterns, but those with two response categories obtained slightly higher loadings compared to six response categories for most items. Questionnaires with two response categories showed slightly better fit indices than those with six categories. All model fits were within the expected range for psychological questionnaires (Goretzko et al., 2024). In particular, the fit indices for the questionnaires on body height and weight with two response categories showed the best fit.

For the body height scales, the self-reported physical height loaded higher or similarly high on the latent variables compared to the questionnaire items. However, this pattern slightly differed for the weight and age scales, where several items loaded substantially higher on the latent variable than the physical items: For weight, the items 4 (“I am obese.”) and 5 (“I weigh a lot.”) had especially high loadings on the latent variable for the questionnaires with two and six response categories. For age, item 7 (“I have already lived most of my life.”) had especially high loadings on the latent variable for the questionnaires with two and six response categories. The item loadings for the weight and age scales suggest that the gold-standard items were not essential for measuring the latent variable.

Even after allowing for correlated errors as suggested by standard modification indices, the loading patterns did not improve (see Table A1). Loading patterns with and without modifications were correlated at .98 ($p < .001$) across all questionnaires (body height, body weight, age) and response variants (two and six categories).

Assumption 2: The measurement should have a psychological component, and the item responses cannot be fully explained by the self-reported physical measurement and gender. A systemic residual should remain when corrected for self-reported physical height and gender.

Table 3 displays item loadings on the latent variable when controlling for the self-reported physical item and self-reported gender.

Table 3

Standardized item loadings on the latent variable for two and six categories controlling for the self-reported physical item and self-reported gender

	Height			Weight			Age		
Item	Latent	Physical	Gender	Latent	Physical	Gender	Latent	Physical	Gender
				Two categories					
1	0.223	**0.804**	-0.138	0.241	0.452	-0.043	0.336	0.287	-0.058
2	0.368	**0.882**	-0.274	0.441	0.313	-0.185	**0.795**	0.278	-0.071
3	0.405	**0.922**	-0.475	0.478	**0.764**	-0.357	0.015	**0.655**	0.077
4	0.292	**0.895**	-0.114	**0.682**	**0.740**	-0.278	**0.637**	0.481	-0.024
5	0.346	**0.742**	-0.067	0.574	**0.877**	-0.346	**0.754**	0.454	-0.056
6	0.312	**0.955**	-0.279	**0.652**	**0.796**	-0.417	0.271	-0.005	-0.004
7	0.299	**0.922**	-0.231	-0.385	**-0.767**	0.265	0.270	**0.797**	-0.043
8	0.207	**0.972**	**-0.778**	-0.246	**-0.627**	0.155	0.377	0.405	-0.246
9	-0.319	**-0.883**	0.088	-0.020	-0.505	0.095	0.024	**0.777**	0.094
10	-0.361	**-0.917**	0.270	0.263	0.468	0.070	-0.265	**-0.687**	0.017
11	-0.303	**-1.061**	0.399	0.174	0.143	0.086	0.292	0.333	-0.016
12	0.318	**1.011**	-0.373	**0.634**	0.561	-0.400	0.577	0.448	0.036
				Six categories					
1	0.223	**0.624**	-0.116	0.362	0.432	-0.122	0.458	0.258	0.053
2	0.365	**0.869**	-0.197	0.413	0.387	-0.259	**0.755**	0.327	0.041
3	0.408	**0.701**	-0.147	0.475	**0.759**	-0.315	0.029	0.583	0.059
4	0.391	**0.863**	-0.169	**0.643**	**0.788**	-0.445	**0.700**	0.367	0.051
5	0.306	**0.695**	-0.016	0.536	**0.846**	-0.403	**0.698**	0.409	0.016
6	0.363	**0.834**	-0.174	**0.632**	**0.756**	-0.477	0.273	0.087	0.014
7	0.369	**0.830**	-0.184	-0.396	**-0.755**	0.374	0.302	**0.762**	0.031
8	0.355	**0.899**	-0.484	-0.277	**-0.627**	0.194	0.426	0.395	-0.121
9	-0.382	**-0.801**	0.099	-0.114	-0.464	0.043	0.036	**0.754**	0.053
10	-0.291	**-0.790**	0.248	0.331	**0.625**	-0.045	-0.201	**-0.664**	-0.048
11	-0.323	**-0.949**	0.340	0.187	0.210	-0.065	0.443	0.270	-0.059
12	0.321	**0.909**	-0.273	0.577	0.591	-0.404	**0.613**	0.447	0.112

Note. Standardized regression weights greater than .60 in absolute values are printed in bold. Latent = standardized item loading on the respective latent variable (height, weight or age); Physical = standardized regression weight, predicting the item response by the respective self-reported physical measures (height, weight or age); Gender = standardized regression weight, predicting the item response by the self-reported gender.

For body height, we obtained low item loadings on the latent variable but high regression weights with the gold-standard item of physical body height. In particular, the third item ("Chairs and tables are usually too low for me.") exhibited the highest loading on the latent variable for both response formats. In contrast, item 11 ("I have to stand at the front of group photos so that I can be seen clearly.") and item 12 ("When I hug other people in greeting, I have to bend downwards.") were best predicted by the self-reported physical body height for both response formats.[3] This pattern does not convincingly confirm the existence of a psychological latent height variable after controlling for the self-reported physical item and gender.

For body weight, the items with high loadings on the latent variable also exhibited high associations with the self-reported physical body weight. Item 4 ("I am obese.") had the highest loading on the latent variable for both response formats. Item 5 ("I weigh a lot.") was explained best by the self-reported physical body weight. Item 9 ("I have a lot of space between the armrests in airplane seats.") had a low loading on the latent variable but was moderately predicted by the self-reported physical body weight for both response formats. Notably, item 11 ("If I have the choice between the elevator and the stairs, I take the elevator.") was not related to either the self-reported physical item or the latent variable for both response formats, indicating that it may represent a more health-related perspective. Taken together, the resulting latent weight variable is hard to interpret.

[3]Note that although the models for height (controlling for the physical item and gender) estimated some absolute standardized loadings greater than 1, all models converged and the diagnostics used by the *lavaan* package (Rosseel, 2012) did not report any problems.

For age, we observed the biggest discrepancy between items with high loadings on the latent variable and items best predicted by self-reported physical age. Item 3 (“I have a lot of life experience.”) and item 9 (“I have years of work experience.”) did not load on the latent variable after controlling for physical age and gender but were predicted by self-reported physical age for both response formats. In contrast, the three items most prototypical of the latent variable were item 2 (“Over time, my mental capacity has decreased.”), item 5 (“Over time, my memory has deteriorated.”), and item 4 (“Over time, my ability to react has decreased.”) for both response formats. Thus, the latent variable might be interpreted as mental age, but not all items have a substantial loading on this general factor.

When looking at self-reported gender, there are moderate standardized regression weights for some items of body height (e.g., item 8 “Pants are often so short for me.”) and body weight (e.g., item 6 “I need to lose weight.”). These items may cause problems when applied to both men and women (differential item functioning; see Hilbert et al., 2022, for a comparable case). Notably, there are only non-significant low standardized regression weights of self-reported gender predicting age items.

Assumption 3a, b: The correlational pattern between the latent variables for the construct should correspond to the correlations of the self-reported gold-standard items and the validity should be higher for the 6-point response categories compared to the 2-point response categories.

As depicted in Table 4, the correlations between the latent variables do not match those between the self-reported physical items. While height and age showed relatively similar low correlations, the physical correlations of weight and age were overestimated by the correlations between the latent variables and the physical correlations between height and weight were underestimated. Most notably, correlations did not differ substantially between the questionnaires with two and with six response categories. Descriptively, the physical correlations seemed to be

reproduced by the latent correlations slightly better for the two response categories. However, reliability estimates (Omega) were lower for two response categories compared to six response categories.

Table 4

Correlations between the latent variables body height, body weight and age, reliability estimates, and correlations between the corresponding self-reported physical items for two and six response categories

Measure	Height		Weight		Age	
Categories	Two	Six	Two	Six	Two	Six
Height	.878	.920	**.252*** **	**.182*** **	**-.053**	**.003**
Weight	.658***	.589***	.785	.886	**.297*** **	**.371*** **
Age	-.089**	-.116***	.185***	.204***	.794	.855

Note. Correlations between the latent variables are shown above the diagonal (printed in bold) and correlations between the self-reported physical measures are shown below the diagonal. The correlations were taken out of a latent variable model (one for two and one for six response categories) including all self-reported physical items and all latent variables with the allocated items. The models showed the following fit: scaled chi-square (six response categories) = 6,203.19, $df = 696$, $p < .001$, scaled CFI = .85, scaled RMSEA = .09, SRMR = .11; scaled chi-square (two response categories) = 3,440.02, $df = 696$, $p < .001$, scaled CFI = .86, scaled RMSEA = .07, SRMR = .12. In the diagonal (grey shaded boxes) the reliability estimates Omega for the scales can be found. * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Discussion

Summary

The goal of this study was to find out whether gold-standard items have the highest loadings on the corresponding latent variable. This was investigated using items with two and six response categories which measure these physical traits. The results reveal that this is only an exception, and two response categories perform not necessarily worse than six response categories regarding validity and reliability. Gold-standard items might help to ensure the interpretation of the latent variable. The choice of the response format should be based on empirical testing and extensive cognitive interviews. It seems not sufficient to take advice from the literature on how to choose the number of response categories to optimize reliability. The benefit of gold-standard items will be discussed in the following sections.

Gold-Standard Items

Gold-standard items did not work as theoretically expected in this study. However, we still believe they are important. Only the gold-standard item for body height – for six, not for two response categories – had the highest loading on the latent variable. Bortolotti et al. (2013), who previously applied the same procedure with two response categories to investigate a gold-standard item for body height, found a correlation of .86 with body height. This correlation pretty much matches our correlation with the latent variable. The correlation was slightly higher for our 2-response category scale (.87) compared to our 6-response category scale (.86), but there were also items which had descriptively equal or higher correlations (biserial correlations) with the sum score than the self-reported physical item (Pearson correlation). This might be because we reported a latent correlation corrected for disattenuation, and our scale had 12 instead of 27 items. Nevertheless, our results depict a conceptual replication of Bortolotti et al. (2013) with a different questionnaire.

The results differed for the other latent variables of body weight and age, which exhibited lower correlations between the latent variable and their gold-standard items. This seems

problematic as these items were designed to be the "gold" measure of the latent variable and should ensure a consistent interpretation. Thus, we should carefully reflect what this finding means for item construction for psychological scales of non-physical traits. If this basic requirement is not met, even for traits that seem rather straightforward to measure, what about more complex constructs? To avoid a misleading interpretation of a latent variable, it might be very helpful to formulate a priori hypotheses about the importance of the constructed items as we did here to test the construct validity of a test. If we, for example, want to measure *interpersonal warmth* as a big five facet and construct items that describe this facet, an item that most directly measures warmth should have its highest loading on the latent variable, otherwise the interpretation of the latent variable is not justified.

We must be aware, that general model fit does not prove how well items are formulated (see Pargent et al., 2019). Model fit also does not reveal if the applied items are helpful to measure the latent variable. Model fit simply compares a model implied variance/covariance matrix with an empirical variance/covariance matrix. In this study, the a priori most suitable items did not have always the highest discrimination parameters or loadings. Thus, simply selecting items according to their loadings and letting the factor analysis decide which item is good or not is probably insufficient. There must be an item content loading fit.

The reliability estimate of the questionnaire informs us only about the precision of measurement and not whether our measurement is valid. This does not necessarily contradict attempts to choose items based on reliability (e.g., see Zijlmans et al., 2019) since items must have high loadings to achieve high reliability estimates for the scale.

Thus, we suggest using gold-standard items as described above. It should be noted, however, that it is critical not to implement the gold-standard item in the final scale since we would then create a part-whole relationship. This no desirable situation since all test models (classical and IRT models) assume that the probability of a positive response of a person to an item should not depend on this person's response to any other item (Debelak & Koller, 2020).

Physical Traits as Psychological Variables

Previous hints suggest that measuring physical attributes through psychological questionnaires is not always connected to a psychological latent variable guiding behavior (Hilbert et al., 2022; Van der Linden, 2016). Similarly, in our study, it remained unclear if a psychological latent variable beyond body height and weight was measured and how such a latent variable could be interpreted. Only for age there was a latent variable, which can be interpreted as cognitive or psychological age (Barak & Schiffman, 1981). Furthermore, it seems helpful to include additional items, that are not intended to be measured but may further explain the item responses of interest, such as gender or items from the nomological net to reveal possible dependencies and to decide whether these dependencies are in line with the definition of the latent variables.

Number of Categories

As expected, reliability estimates were higher for scales with six response categories compared to two categories (Preston & Colman, 2000). The results with regard to model fit and validity, indicated as correlations between latent variables compared to the correlation between gold-standard items, are mixed. Some fit indices suggested slightly better fit for items with two response categories and some for six categories. The same holds true for the differences between the correlations of the gold-standard items and the correlations of the latent variables. Except for reliability, the study revealed no clear picture to choose between two or six response categories. Hilbert et al. (2016, 2022) showed that the same items with a different number of response categories do not measure the same latent variable. This result highlights that the choice of the number of response categories constructing a questionnaire should be based on an empirical study comparing several options accompanied by an intensive cognitive pretesting phase and not only based on the literature which focuses on maximal reliability.

Recommendations

The results presented here should prompt reflection and reconsideration of standard procedures of item construction in psychology. First, it may be helpful to more strongly consider content validity when developing items. Items that either represent a kind of gold-standard with specific hypotheses about the ranking of the loadings or construct-divergent items that (in)validate the items intended to measure a latent variable should be included in the construction process. Revisiting the example of *interpersonal warmth* as a personality facet, such additional validation items could, for example, measure warmth itself or openness. An openness scale or marker items that show high correlations with these warmth items should not be included in the final scale but only used for validation purposes. These demands are not entirely new, but should be newly emphasized based on the results of our analyses, calling for more stringent item construction approaches. Second, the choice of an optimal response scale may not be as clear-cut as previously suggested (e.g., Lee & Paek, 2014). We found that items with two response categories were not per se unfavorable, as often suggested in the literature. While our binary scales obtained lower reliability estimates as expected (see Preston & Colman, 2000; Revilla et al., 2014), other results regarding validity and model fit were mixed. Since reliability is easier to control for by adding more suitable items, construct validity seems to be the stronger argument when choosing the number of response categories. In sum, we believe that decisions on response scales should be empirically founded, especially in high stakes situations (e.g., clinical diagnosis or personnel selection), comparing alternative numbers of response categories in pilot testing. In this context, it is important to recognize that the number of response categories can influence what latent variables are measured (see Hilbert et al., 2016, 2022). Thus, comparing the construct validity of questionnaires differing in the number of response categories is essential. Nevertheless, it remains necessary to subject response scales to intensive cognitive pretesting to determine a suitable number of categories. On the one hand, cognitive interviews can reveal

when single categories are systematically overlooked by respondents lacking information or cognitive capacity to differentiate categories. On the other hand, cognitive pretests help assess participants' cognitive load, which should always be kept in mind. For example, two-point scales, which are commonly criticized, have previously been shown to require less effort from the respondents (Hilbert et al., 2016).

Conclusion

Based on the results of the present study, we advocate the inclusion of gold-standard items, i.e. items that are supposed to measure the intended construct most directly, to check the interpretation of latent variables in the construction of psychological questionnaires. In addition, for questionnaires that are used to make particularly important decisions, we recommend empirically testing the choice of response format, considering both validity and reliability to make a well-founded decision. The pre-selection of response formats should include cognitive pretests of the items.

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., Dervieux, C., & Woodhull, G. (2025). *Quarto* (Version 1.7) [Computer software]. https://doi.org/10.5281/zenodo.5960048

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://doi.org/10.32614/CRAN.package.papaja

Barak, B., & Schiffman, L. G. (1981). Cognitive age: A nonchronological age variable. *Advances in Consumer Research*, *8*(1), 602–606.

Barth, M. (2025). *tinylabels: Lightweight variable labels*. https://doi.org/10.32614/CRAN.package.tinylabels

Bortolotti, S. L. V., Tezza, R., De Andrade, D. F., Bornia, A. C., & De Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*(4), 2341–2360. https://doi.org/10.1007/s11135-012-9684-5

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. M. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(1), 1–12. https://doi.org/10.29115/SP-2015-0001

Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology*, *47*(6), 1334–1346. https://doi.org/10.1037/0022-3514.47.6.1334

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*(3), 214–227. https://doi.org/10.1037/0003-066X.39.3.214

Conigrave, J. (2023). *Corx: Create and format correlation matrices*. https://CRAN.R-project.org/package=corx

Debelak, R., & Koller, I. (2020). Testing the Local Independence Assumption of the Rasch Model With Q_3 -Based Nonparametric Model Tests. *Applied Psychological Measurement*, *44*(2), 103–117. https://doi.org/10.1177/0146621619835501

Eichhorn, K. (2019). *Item und skala: Empirische untersuchungen zur gültigkeit psychologischer messungen anhand physikalischer merkmale* [Ludwig-Maximilians-Universität München]. http://nbn-resolving.de/urn:nbn:de:bvb:19-236286

Eichhorn, K., Bühner, M., Pargent, F., Saretzki, J., Sust, L., Hauck, J., & Hilbert, S. (2025). *Measuring latent variables with gold standard items*. OSF. https://doi.org/10.17605/OSF.IO/P7492

Faulbaum, F., Prüfer, P., & Rexroth, M. (2009). *Was ist eine gute Frage?* VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91441-1

Gohel, D., & Skintzos, P. (2024). *Flextable: Functions for tabular reporting*. https://CRAN.R-project.org/package=flextable

Goretzko, D., Pargent, F., Sust, L. N. N., & Bühner, M. (2020). Not Very Powerful: The Influence of Negations and Vague Quantifiers on the Psychometric Properties of Questionnaires. *European Journal of Psychological Assessment*, *36*(4), 563–572. https://doi.org/10.1027/1015-5759/a000539

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. https://www.jstatsoft.org/v40/i03/

Hilbert, S., Küchenhoff, H., Sarubin, N., Toyo Nakagawa, T., & Bühner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *23*(1), 3–24. https://doi.org/10.4473/TPM23.1.1

Hilbert, S., Pargent, F., Kraus, E., Naumann, F., Eichhorn, K., Ungar, P., & Bühner, M. (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

Holt, J. C. ten, Duijn, M. A. J. van, & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272–297.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2025). *semTools: Useful tools for structural equation modeling*. https://CRAN.R-project.org/package=semTools

Kelley, K. (2023). *MBESS: The MBESS r package*. https://CRAN.R-project.org/package=MBESS

Lee, J., & Paek, I. (2014). In Search of the Optimal Number of Response Categories in a Rating Scale. *Journal of Psychoeducational Assessment*, *32*(7), 663–673. https://doi.org/10.1177/0734282914522200

Lord, F. M., Novick, M. R., Birnbaum, A., & Lord, F. M. (1968). *Statistical theories of mental test scores* (2. print). Addison-Wesley.

Martin, R. (1929). *Anthropometrie: Anleitung Zu Selbständigen Anthropologischen Erhebungen* (2nd ed). Springer Berlin / Heidelberg.

Masuda, S., Sakagami, T., Kawabata, H., Kijima, N., & Hoshino, T. (2017). Respondents with low motivation tend to choose middle category: Survey questions on happiness in Japan. *Behaviormetrika*, *44*(2), 593–605. https://doi.org/10.1007/s41237-017-0026-8

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McCreary, D. (2002). Gender and Age Differences in the Relationship Between Body Mass Index and Perceived Weight: Exploring the Paradox. *International Journal of Men's Health*, *1*(1), 31–42. https://doi.org/10.3149/jmh.0101.31

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Montepare, J. M., & Lachman, M. E. (1989). "You're only as old as you feel": Self-perceptions of age, fears of aging, and life satisfaction from adolescence to old age. *Psychology and Aging*, *4*(1), 73–78. https://doi.org/10.1037/0882-7974.4.1.73

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. https://tibble.tidyverse.org/

Pargent, F., Hilbert, S., Eichhorn, K., & Bühner, M. (2019). Can't Make it Better nor Worse: An Empirical Study About the Effectiveness of General Rules of Item Construction on Psychometric Properties. *European Journal of Psychological Assessment*, *35*(6), 891–899. https://doi.org/10.1027/1015-5759/a000471

Posit team. (2025). *RStudio: Integrated development environment for r*. Posit Software, PBC. http://www.posit.co/

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Prüfer, P., & Rexroth, M. (1996). *Verfahren zur evaluation von survey-fragen: Ein überblick* (Vol. 1996/05, p. 34). Zentrum für Umfragen, Methoden und Analysen -ZUMA-.

Prüfer, P., & Rexroth, M. (2005). *Kognitive interviews* (Vol. 15, p. 26). Zentrum für Umfragen, Methoden und Analysen -ZUMA-.

R Core Team. (2025a). *Foreign: Read data stored by 'minitab', 's', 'SAS', 'SPSS', 'stata', 'systat', 'weka', 'dBase', …* https://CRAN.R-project.org/package=foreign

R Core Team. (2025b). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schneider, W. J. (2025). *apaquarto* [Computer software]. https://github.com/wjschne/apaquarto

Schwall, A. R. (2012). *Defining Age and Using Age-Relevant Constructs*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195385052.013.0080

Van der Linden, W. J. (2016). Unidimensional logistic response models. In W. J. Van der Linden (Ed.), *Handbook of item response theory, volume one: models* (pp. 13–30). Chapman & Hall/CRC. https://doi.org/10.1201/9781315374512

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. https://forcats.tidyverse.org/

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. https://stringr.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools*. https://CRAN.R-project.org/package=purrr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. https://readr.tidyverse.org

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. https://tidyr.tidyverse.org

William Revelle. (2025). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Willis, G. (2005). *Cognitive Interviewing*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983655

Zijlmans, E. A. O., Tijmstra, J., Van Der Ark, L. A., & Sijtsma, K. (2019). Item-Score Reliability as a Selection Tool in Test Construction. *Frontiers in Psychology*, *9*, 2298. https://doi.org/10.3389/fpsyg.2018.02298

Appendix A

CFA models with correlated errors

Table A1

Latent variable loadings for two and six response categories with correlated errors

Measure	Height		Weight		Age	
Categories	Two	Six	Two	Six	Two	Six
			Items			
1	.782	.616	.572	.571	.440	.509
2	.823	.839	.462	.510	.575	.705
3	.788	.745	**.817**	**.800**	.604	.468
4	.854	.824	**.938**	**.958**	**.787**	**.766**
5	.785	.749	**.969**	**.913**	.727	.727
6	.880	.822	**.832**	**.861**	.131	.240
7	.850	.815	**-.741**	**-.719**	**.909**	**.789**
8	.627	.725	-.664	-.627	.600	.591
9	-.859	-.805	-.433	-.427	.756	.660
10	-.799	-.673	.626	.692	**-.790**	-.607
11	-.844	-.792	.248	.264	.465	.505
12	.881	.822	.582	.678	.715	**.767**
Physical item	.884	.871	.717	.700	.782	.732
			Model fit			
χ^2	339.111	844.509	24.293	686.681	489.419	1,167.002
df	63	63	63	63	62	62
p	< .001	< .001	< .001	< .001	< .001	< .001
CFI	.977	.952	.973	.965	.918	.882
RMSEA	.069	.115	.055	.103	.087	.138
SRMR	.080	.060	.074	.061	.088	.091

Note. CFI = Comparative Fit Index (scaled). RMSEA = Root Mean Squared Error of Approximation (scaled). SRMR = Standardized Root Mean Residual. Scaled chi-square, scaled df (degrees of freedom), and scaled p values are reported. Correlations with the latent variable higher as the self-reported physical item are printed in bold. The standardized item loadings for the self-reported physical items are presented in a grey shaded box.The following correlated errors were specified according to high modification indices and good interpretability (parenthesis error correlations for two and six response categories): “I have to look up when I talk to other people.” (H10) and “I have to stand at the front of group photos so that I can be seen clearly.” (H11) (0.57, 0.44); “In the supermarket, I can reach the things on the top shelf without any problems.” (H4) and “I need a chair if I want to get things from the top shelf.” (H9) (-0.51, –0.40); “I need to lose weight.” (W6) and “I should eat less.” (W12) (0.71, 0.55); “I have a wide waistband when it comes to pants.” (W3) and “My pants have a small waistband.” (W7) (-0.55, –0.33); “Over time, my mental capacity has decreased.” (A3) and “I have years of work experience.” (A9) (0.53, 0.49); “Over time, my mental capacity has decreased.” (A2) and “Over time, my memory has deteriorated.” (A5) (0.62, 0.40); “I have already lived most of my life.” (A7) and “My whole life still lies ahead of me.” (A10) (-0.41, –0.48).

Appendix B

Item translations

Table B2

Translated items used in this study

Item	Item wording (Original wording in German)
Height 1	I keep my head down when I walk through doors. (Ich ziehe den Kopf ein wenn ich durch Türen gehe.)
Height 2	In the middle of a crowd, I can see over most other people. (Inmitten einer Menschenmenge kann ich über die meisten anderen Menschen hinwegblicken.)
Height 3	Chairs and tables are usually too low for me. (Stühle und Tische sind meistens zu niedrig für mich.)
Height 4	In the supermarket, I can reach the things on the top shelf without any problems. (Im Supermarkt komme ich ohne Probleme an die Sachen im obersten Regal.)
Height 5	When I drive a car that was previously driven by someone else, I usually have to push the seat backwards. (Wenn ich ein Auto fahre, das vorher jemand anderes gefahren hat, muss ich üblicherweise den Sitz nach hinten schieben.)
Height 6	When I stretch out straight in hotel beds, my feet overlap. (Wenn ich mich in Hotelbetten gerade ausstrecke, dann stehen meine Füße über.)
Height 7	On airplanes, I bump my knees against the seat in front of me. (Im Flugzeug stoße ich mit den Knien am Vordersitz an.)
Height 8	Pants are often too short for me. (Hosen sind mir häufig zu kurz.)
Height 9	I need a chair if I want to get things from the top shelf. (Ich brauche einen Stuhl, wenn ich Sachen aus dem obersten Regal holen möchte.)
Height 10	I have to look up when I talk to other people. (Wenn ich mich mit anderen Menschen unterhalte muss ich nach oben schauen.)

Item	Item wording (Original wording in German)
Height 11	I have to stand at the front of group photos so that I can be seen clearly. (Ich muss mich bei Gruppenfotos nach vorne stellen, damit ich gut zu sehen bin.)
Height 12	When I hug other people in greeting, I have to bend downwards. (Wenn ich andere Menschen zur Begrüßung umarme, muss ich mich nach unten beugen.)
Weight 1	I am often afraid that chairs will give way under me. (Ich habe oft Angst, dass Stühle unter mir nachgeben.)
Weight 2	Other people may think that I sit on the sofa in the evening and eat chocolate and potato chips. (Möglicherweise denken andere Menschen von mir, dass ich abends auf dem Sofa sitze und Schokolade und Chips esse.)
Weight 3	I have a wide waistband when it comes to pants. (Bei Hosen habe ich eine große Bundweite.)
Weight 4	I am obese. (Ich bin dick.)
Weight 5	I weigh a lot. (Ich wiege viel.)
Weight 6	I need to lose weight. (Ich müsste abnehmen.)
Weight 7	My pants have a small waistband. (Meine Hosen haben eine kleine Bundweite.)
Weight 8	I am good at squeezing through narrow gaps. (Ich kann mich gut durch enge Spalten quetschen.)
Weight 9	I have a lot of space between the armrests in airplane seats. (Zwischen den Armlehnen in Flugzeugsitzen habe ich viel Platz.)
Weight 10	I take up a lot of space in the elevator. (Im Aufzug nehme ich viel Platz ein.)
Weight 11	If I have the choice between the elevator and the stairs, I take the elevator. (Wenn ich die Wahl habe zwischen Aufzug und Treppe, dann nehme ich den Aufzug.)

Item	Item wording (Original wording in German)
Weight 12	I should eat less. (Ich sollte weniger essen.)
Age 1	I have physical ailments. (Ich habe körperliche Gebrechen.)
Age 2	Over time, my mental capacity has decreased. (Im Laufe der Zeit hat meine geistige Leistungsfähigkeit abgenommen.)
Age 3	I have a lot of life experience. (Ich habe viel Lebenserfahrung.)
Age 4	Over time, my ability to react has decreased. (Im Laufe der Zeit hat meine Reaktionsfähigkeit abgenommen.)
Age 5	Over time, my memory has deteriorated. (Im Laufe der Zeit hat meine Merkfähigkeit abgenommen.)
Age 6	I worry a lot about getting older. (Ich mache mir viele Gedanken über das Älterwerden.)
Age 7	I have already lived most of my life. (Den größten Teil meines Lebens habe ich bereits gelebt.)
Age 8	I find it increasingly difficult to follow technical developments. (Ich habe zunehmend Schwierigkeiten technischen Weiterentwicklungen zu folgen.)
Age 9	I have years of work experience. (Ich habe jahrelange Berufserfahrung.)
Age 10	My whole life still lies ahead of me. (Mein ganzes Leben liegt noch vor mir.)
Age 11	I used to be more willing to take risks. (Früher war ich risikofreudiger.)
Age 12	I carry out everyday activities more slowly than before. (Ich verrichte Alltagstätigkeiten langsamer als früher.)

Note: Original wording in German shown in parenthesis.

Appendix C

Software versions

Quarto (Allaire et al., 2025) version 1.7.32 and the extension apaquarto (Schneider, 2025) were used together with RStudio (Posit team, 2025) to build the manuscript and the manuscript website.

R (Version 4.5.1; R Core Team, 2025b) and the R-packages *corx* (Version 1.0.7.2; Conigrave, 2023), *dplyr* (Version 1.1.4; Wickham et al., 2023), *flextable* (Version 0.9.9; Gohel & Skintzos, 2024), *forcats* (Version 1.0.0; Wickham, 2023a), *foreign* (Version 0.8.90; R Core Team, 2025a), *ggplot2* (Version 3.5.2; Wickham, 2016), *lavaan* (Version 0.6.19; Rosseel, 2012), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *MBESS* (Version 4.9.3; Kelley, 2023), *papaja* (Version 0.1.3; Aust & Barth, 2024), *psych* (Version 2.5.6; William Revelle, 2025), *purrr* (Version 1.1.0; Wickham & Henry, 2025), *readr* (Version 2.1.5; Wickham, Hester, et al., 2024), *semTools* (Version 0.5.7; Jorgensen et al., 2025), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.3.0; Müller & Wickham, 2023), *tidyr* (Version 1.3.1; Wickham, Vaughan, et al., 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019) and *tinylabels* (Version 0.2.5; Barth, 2025) were used for data analysis.

Some dependencies are not included in this list but can be found in the `renv.lock` file in our online repository.