Psychological Test and Assessment Modeling, Volume 63, 2021 (1), 119–145

Structural Validity of Overclaiming Scores: analysing PISA 2012 data

Marek Muszyński[1], *Artur Pokropek*[2] *& Tomasz Żółtak*[2]

Abstract

Overclaiming technique is a promising tool that can account for self-assessment imprecision, improve cross-country comparability and screen for fakers in high-stakes contexts. Despite the rising popularity of the overclaiming technique - evidenced by an increasing number of papers, citations and versions of the method - very little is known about the internal structure of scales designed to measure overclaiming tendencies. It is especially worrisome, as internal structure is one of the main sources of construct validity and its coherence to the assumed theory vouches for scores' interpretability and validity. We aim to fill in this research gap and use the obtained results to comment on the validity of using overclaiming technique's scores in research practice, where a two-factorial structure of the tool is assumed. To this end, we analyse the PISA 2012 overclaiming scale's internal structure by applying confirmatory multilevel factor analysis. Our results suggest that items in the PISA overclaiming scale cannot be simply interpreted as reals (construct variance) and foils (bias variance) as both types of items measure both types of variance. We also show that the simple ontic status of an item is not enough to guarantee intended measurement characteristics, namely that real items only measure genuine knowledge and that foil items solely capture a tendency to overclaim. The obtained results are used not only to reflect on overclaiming technique's internal structure but also to give advice on constructing such scales in the future.

Keywords: overclaiming technique, internal structure, factor analysis, construct validity, PISA.

[1] *Correspondence concerning this article should be addressed to:* Dr. Marek Muszyński, Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland. E-mail: marek.muszynski@ifispan.edu.pl

[2] Institute of Philosophy and Sociology, Polish Academy of Sciences

Noncognitive constructs such as personality traits, attitudes, self-assessments, or reported behaviour are of great interest in every area of the social sciences (Paulhus & Vazire, 2007; Ziegler, 2015). They are predominantly measured using self-report standardised questionnaires that usually contain a predefined set of response options. Research results show that self-reports constitute a large part of the research methods used in many subdisciplines of the social sciences (Brückner, 2009; Brutus, Gill, & Duniewicz, 2010; Woszczynski & Whitman, 2004). This self-reports popularity is due to their cost efficiency, ease of administration, and flexibility to assess a broad range of constructs. Moreover, they are believed to provide valid, interpretative, standardized, and comparable information across subjects (Lucas & Baird, 2004).

However, the use of self-report does not come without complications. Assumptions that respondents use and interpret the given response categories in the same way (comparability assumption) and give unbiased and honest responses are not always held (Paulhus & Vazire, 2007; Wetzel, Böhnke & Brown, 2016). The main reason for this are response biases, defined as a systematic tendency to answer questions on other basis than their content (Paulhus, 1991). Response biases introduce a systematic source of error variance to the measurement, thus reducing its validity and comparability (Messick, 1989; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Wetzel et al., 2016; Ziegler, 2015).

Many methods have been devised to control for response biases so far, however, none of them achieved the status of an unquestionable standard. Among the methods proposed, some are of unverified validity. This situation is especially vexing in case of low-stakes, non-intrusive measurement situations in which many methods developed to account for response biases in high-stakes settings simply cannot be implemented (Krumpal, 2013; Paulhus, 2002; Ziegler, 2015; Ziegler, MacCann, & Roberts, 2012). However, new solutions to capture response bias also in low-stakes measurement situations, including international large-scale assessments (ILSAs), have been recently proposed. Such techniques have to be easy to use and score, pose low cognitive burden on participants, be cost- and time-efficient and offer high flexibility regarding its use in diverse modes, populations, and contexts. Most importantly, they have to be valid indicators of response biases, offering reliable means to control for spurious variance and to raise measurement validity (Ferrando, 2005).

One of the techniques designed to deal with response bias in self-reports is the overclaiming technique (Paulhus, Harms, Bruce, & Lysy, 2003). The technique is based on an idea proposed by Phillips and Clancy (1972), who gauged commercial products' knowledge (e.g. books, movies, etc.), and placed non-existent product names among the list of factual products to control for response bias (self-enhancement tendencies). In a question asking about recognizing famous politicians an item labelled “Barack Obama” would be an example of one measuring political knowledge, while an item labelled “Peter Lawn” would be an illustration of one gauging bias tendencies. The terms “reals” and “foils” are often used to address these kinds of items, respectively (Paulhus et al., 2003). Phillips and Clancy also proposed the terms “overclaiming” and “overclaimer” (1972). Overclaiming is thus defined as “the degree to which individuals claim knowledge, about factually non-existent items” (Müller & Moshagen, 2018). The idea behind this technique is straightforward - if participants claim to know non-existent items or to possess non-existent skills it is considered as a clear indication of response bias (e.g. socially desirable responding, self-enhancing tendencies, etc.). Obviously, apart from knowledge, participants can also assess their skills, abilities, behaviours, possessions, etc., hence the overclaiming technique is potentially a very versatile tool, applicable in many research situations.

Despite the simple and appealing idea, the studies on overclaiming technique yield a mixed pattern of method's utility. Some studies showed that this method is a valid suppressor of bias variance in high-stakes contexts (e.g. Bing, Kluemper, Davison, Taylor, & Novicevic, 2011; Dunlop, Bourdage, de Vries, McNeill, Jorritsma, Orchard, Austen, Baines, & Choe, 2020; Mesmer-Magnus, Viswesvaran, Deshpande, & Joseph, 2006), but not in low-stakes contexts (e.g. Feeney & Goffin, 2015; Ludeke & Makransky, 2016). Moreover, also other results cast doubt on overclaiming technique's utility (Kam, Risavy, & Perunovic, 2015; Musch, Ostapczuk, & Klaiber, 2012) and, most importantly, on its interpretation as a measure of positivity bias (social desirability responding, self-enhancement tendencies, etc.; Dunlop, Bourdage, de Vries, Hilbig, Zettler, & Ludeke, 2017; Franzen & Mader, 2019; Goecke, Weiss, Steger, Schroeders, & Wilhelm, 2020; Mesmer-Magnus et al., 2006; Müller & Moshagen, 2018; 2019a; 2019b; Steger, Schroeders, & Wilhelm, 2020). Alternative explanations, linking overclaiming to memory biases (Dunlop et al., 2017) or careless responding (Barber, Barnes, & Carlson, 2013; Ludeke & Makransky, 2016), were proposed recently.

It seems that the overclaiming technique can still be regarded as an important tool to control for response biases in self-report measures, as evidenced by Kyllonen and Bertling (2013), who showed that the method helped to improve the self-reports' criterion validity and cross-country comparability in the PISA 2012 data. Problems with its validity and interpretability may result from still insufficient validity studies and possible imperfections of the technique's versions created so far (Goecke et al., 2020). Therefore, more attention should be devoted to analysing its validity to identify potential problems and propose adequate remedies in future developments of the tool. One of the areas where validity evidence is especially lacking is research on overclaiming technique internal structure, namely the relations between items and proposed latent factors they create (AERA, APA, & NCME, 2014, pp. 16, 26-27, 220). The evidence on coherence between the proposed (theoretical) and actual (empirical) measurement model, scale's dimensionality and its internal consistency (reliability) is one of the key sources of construct validity, especially when subscores are to be used and interpreted (AERA, APA, & NCME, 2014; Messick, 1989; 1995; Rios & Wells, 2014). Hence, this article aims to complement this research lacuna and present a comprehensive internal structure study of the overclaiming questionnaire from the PISA 2012 database (OECD, 2014). Confirmatory factor analysis (CFA) will be used to this aim, along with internal consistency analysis (Rios & Wells, 2014). Moreover, as analyses will entail data from all the countries participating in the PISA 2012 cycle we additionally present overclaiming technique's internal structure on the cross-country level.

Previous Research

Due to lack of previous research, we can address the problem of overclaiming technique's internal structure only by relying on theoretical assumptions and using results of tentative research attempts, as well as referring to indirect evidence on technique's subscores correlations and internal consistency analysis. In the first place, the theoretical scale composition is straightforward: the foils and the reals form two separate factors that have different substantial interpretations. However, the empirical data at hand to support this assumption is rather scarce and only tentative evidence exists. Pokropek (2014) analysed overclaiming technique as a tool to increase self-reports criterion-related validity and determined that the scale was not unidimensional and that foils and reals could form separate factors. Similar determinations were made also by Goecke and collaborators (2020), who analysed psychological mechanisms that could be responsible for overclaiming, and by Yang, Barnard-Brak and Lan (2019), who analysed latent classes among students who responded to an overclaiming questionnaire. Both of these studies modelled overclaiming technique scores using two-dimensional structure, creating separate factors for foils and reals. However, among these studies only one contrasted the two-factor solution with alternative models. Another initial evidence comes from the study that used CFA for categorical data and compared the most plausible models, demonstrating that overclaiming technique items grouped to factors by difficulty, not according to their reals-foils (ontic) status (Muszyński, 2020). Nevertheless, these studies can be treated only as an incentive for a more thorough research, not a piece of firm evidence, as they have serious limitations. First of all, they analysed only data from one country (Poland), hence the replicability of its results on the whole PISA sample is uncertain. Moreover, they treated overclaiming scale's internal structure only as a marginal issue in their research, hence the factor solutions were only tentatively assessed without comparing all of the important model candidates, nor modelling the between-level structure. Finally, both studies treated their results only as an exploration, without offering a fully-fledged explanation of the results.

The information on overclaiming technique's internal consistency is even scarcer. In one of the few studies that reported such measures, the inter-item correlations were not very high (Pearson's correlations in the range of 0.28 to 0.50; Franzen & Mader, 2019). On the other hand, Joseph, Berry and Deshpande (2009) reported a quite high Cronbach's alpha of 0.78 for a list of foil items, whereas Randall and Fernandes (1991) reported a very similar value ($\alpha = 0.70$) for a measure consisting of both foils and reals. It is to note, that these studies are not directly comparable due to different characteristics that affect alpha values (e.g. number of items) or that could otherwise influence tool properties (e.g. proportion of reals to foils, item content).

Research questions and Hypothesis

The research question is to investigate the internal structure of the overclaiming technique and in particular bring evidence on whether participants use the same, similar or dissimilar mechanisms when answering to reals versus foils. The first assumption (same processes) would mean that only one factor or, alternatively, two, strictly correlated factors should emerge. On the other hand, using dissimilar processes to answer reals and foils should lead to a creation of two (or more) relatively unrelated factors. It is also interesting whether all items would load on their respective, pre-specified factors, indicating that all foils were indeed considered as such by participants (cf. Ferrando, 2005; Franzen & Mader, 2019; Leite & Cooper, 2010 for situations where foils were treated as reals and *vice versa*).

If overclaiming technique is driven mainly by positivity bias (self-enhancement) all items should be correlated with each other and, even if two factors emerge (e.g. one for reals, one for foils), they should share a large proportion of their variance, as participants can distort their answers to both reals and foils in order to yield a more favourable image of themselves (Hülür, Wilhelm, & Schipolowski, 2011). Such a pattern should also emerge if overclaiming technique is driven by participants claiming foils familiarity due to overgeneralisation of known terms, e.g. as a result of overconfidence or very disinhibited (creative) semantic network (Atir, Rosenzweig, & Dunning, 2015; Paulhus, 2012). Another possible explanation of such a result is stylistic responding, e.g. acquiescence or careless responding (Ludeke & Makransky, 2016). However, if overclaiming is driven by failure in metacognitive monitoring and control, e.g. participants claim to know foils because they are genuinely convinced that they know and understand a given topic, then two relatively unrelated factors should emerge: one for foils, one for reals, based on different cognitive mechanisms predicted to be engaged in responding to reals and foils (Paulhus & Dubois, 2014). In such an occasion the emerging factors should be negatively correlated to some extent as highly competent participants should not claim foils familiarity due to their supreme abilities to search one's memory and/or to their ability to inhibit the alluring foils (Stanovich & Cunningham, 1992; Stanovich & West, 1989).

Moreover, easy and hard items can form separate factors. Such a pattern can be explained on the basis that different mechanisms are responsible for answering easy and hard items. This explanation requires the emerging "easy" and "hard" factors to be only minimally correlated. This solution might suggest that distinct processes are responsible for evaluating known concepts while others for admitting lack of knowledge on concepts that are only vaguely recognised or not known at all. It could be possible that very hard items, denoting very specialised or uncommon knowledge, could be treated in the same way as foils: items that also are unfamiliar concepts. An analogous evidence from cognitive tests can be evoked here and tentatively extrapolated to overclaiming items. The assumption that different item difficulty can lead to both quantitatively but also qualitatively distinct processes used in item responding was formulated by Campbell (1963) and was fairly well documented – by eye-tracking, verbal protocols, and self-ratings – in the domain of intelligence tests (Chuderski, Jastrzębski, Kroczek, Kucwaj, & Ociepka, 2020; Jarosz & Wiley, 2012). In the latter domain participants are known to use different strategies that differ according to respondent's working memory capacity, but also item difficulty. Moreover, there is also evidence that correct *versus* incorrect responses are also related to qualitatively different processes, distinguished also on the metacognitive level (Danek & Wiley, 2017). Therefore, we will also test a model with two dimensions, one for easy items, another for hard items.

We do not have strong theoretical premises for the between-country level structure, hence we do not possess any readily accessible explanations on how to interpret emerging relations on this level of analysis. Here, we propose an exploratory analysis focusing on whether the within-country structure could be replicated also on the between-country level.

Albeit it is difficult to pose any specific hypothesis due to scarcity of evidence, we aim to test the assumed, but insufficiently documented two-factor overclaiming technique structure, where reals and foils form separate factors. We also assume, following Pokropek (2014) and Muszyński (2020) who performed initial analyses on the same dataset, that the PISA 2012 overclaiming questionnaire can yield other multi-dimensional structures, e.g. with items grouped to two factors on the basis of their difficulty. Given that instruments yielding a two-factor oblique structure are often better modelled by bifactor solutions we also test these models in our analysis, forming one general factor for all items, and specific factors for respective subsets. Validating overclaiming scale's internal structure may bring interesting information not only on overclaiming mechanisms, but also provide practical advice on scaling, scoring and interpreting existing overclaiming scales and constructing new ones in the future.

Method

Participants

PISA uses two-stage stratified samples of students enrolled in lower-secondary or upper-secondary institutions and aged between 15 years and 3 months and 16 years and 2 months in the year of the testing in order to represent the full population of this cohort in every participating country. As a different number of students is sampled from each country, this value can be looked across in the PISA 2012 Technical Report along with the list of participating countries and territories (OECD, 2014). Altogether, the main sample in the PISA 2012 cycle amounted to 485 490 students from 67 entities, however, only 322 667 of them filled in the overclaiming scale due to the PISA 2012 rotational questionnaire design (OECD, 2014, p. 61). Elimination of respondents who did not give response to any of the items of the analysed scale, further limited the number of analysed respondents to 310 965. The proportion between female and male participants was almost equal (50.57 % females, 49.43 % males) and the mean age was slightly more than 15 years and 9 months, both statistics are for the participants that were included in the following analysis.

Materials

In the PISA 2012 overclaiming scale was embedded into scale measuring familiarity with mathematical concepts (math familiarity). The scale comprises 16 items altogether, including 13 reals and three foils. Participants responded on a five-categorical rating scale, labelled with both numbers and descriptions, from “Never heard of it” (1) to “Know it well, understand the concept” (5). Respondents were asked to tick only one answer box in each row. Items' content, along with their means, standard deviations and intra-class correlations (ICC) is presented in Table 1.

Table 1:
Means, Standard Deviations and ICCs of items

Item	Content	Type	*M*	Difficulty	*SD*	*ICC*
Q01	exponential function	reals	2.38	hard	1.43	.19
Q02	divisor	reals	3.89	easy	1.35	.22
Q03	quadratic function	reals	3.50	easy	1.45	.16
Q04	**proper number**	**foils**	**3.01**	**hard**	**1.48**	**.17**
Q06	linear equation	reals	3.71	easy	1.43	.23
Q07	vectors	reals	2.98	hard	1.58	.25
Q08	complex number	reals	2.78	hard	1.44	.15
Q09	rational number	reals	3.77	easy	1.35	.21
Q10	radicals	reals	3.76	easy	1.42	.31
Q11	**subjective scaling**	**foils**	**1.94**	**hard**	**1.25**	**.12**
Q12	polygon	reals	3.79	easy	1.44	.31
Q13	**declarative fraction**	**foils**	**2.03**	**hard**	**1.31**	**.09**
Q15	congruent figure	reals	3.21	easy	1.59	.24
Q16	cosine	reals	3.18	hard	1.68	.22
Q17	arithmetic mean	reals	3.18	hard	1.62	.26
Q19	probability	reals	3.83	easy	1.38	.17

Note. N = 310 965. Foil items are in bold.

According to the PISA 2012 technical report foils were created by combining a real grammar term (e.g. “proper” or “declarative”) with a real mathematical term to form a foil item that in its entirety does not mean anything. This method of foil creation would be classified as yielding foils of high risk of confusion with existing mathematical concepts (cf. Franzen & Mader, 2019; Hargittai, 2005; Paulhus et al., 2003). Reals employed in the scale mainly stem from algebra and geometry (OECD, 2014, p. 57).

Procedure

Students participating in the PISA 2012 main survey first sit for cognitive assessment of three domains (reading, mathematics, science), which lasted approximately two hours, and then filled in the so-called background questionnaires for an additional half an hour. The assessment was organised in a self-paced, self-completion proctored session. A vast majority of the students participating in the assessment completed paper-and-pencil questionnaires (OECD, 2014).

As we only used the secondary PISA 2012 dataset publicly available on the OECD sites no issues relating to ethical standards were concerned.

Analysis

In order to contrast competing models of the PISA 2012 overclaiming scale internal structure we have employed two-level confirmatory factor analysis in several configurations. The outlines of the models compared are presented in Figure 1. Since our data have a two-level structure - students nested in countries - we used multilevel CFA models that allows us to account for two-level structure and model between-country structure (based on latent item means) and within- "individual"-level structure (based on individual responses) separately. Without accounting for multilevel structure of the data one could expect similar problems that naïve linear models would generate when fitted to multilevel data (see Muthén, 1991; Reise, Ventura, Nuechterlein, & Kim 2005). The data were weighted using weights that equally weighted each country regardless of the sample size (so-called senate weights; OECD, 2014, p. 396).

The W1 part of the CFA model is simply the one-dimensional structure on the within level with all items loading on one dimension. The W2 structure depicts a two-dimensional model where real items (*r*) and foil items (*f*) load on different correlated dimensions. The W3 depicts a two-dimensional model where easy (*e*) and hard (*h*) items load on different correlated factors. Models W4 and W5 are bifactor structures with one general within factor (*Mw*) and two specific orthogonal factors defined by real and foil items in model W4 (*Rw* and *Fw,* respectively) or by easy (*e*) and hard (*h*) items in model W5 (*Ew* and *Hw,* respectively). By fitting bifactor models we can also test assumptions that the PISA 2012 overclaiming scale is simply unidimensional and any hints of multidimensionality are driven by mainly spurious specific factors. To this end we have fitted two additional S-1 bifactor models: W6 with specific factor only for foils and W7 with specific factor only for hard items (cf. Gnambs & Schroeders, 2020). Finally, we also tested model W8: combined bifactor model with four specific factors comprising reals, foils, easy, and hard items that cross-load items (Gnambs, Scharl, & Schroeders, 2018).

However, in order to reduce confusion we have decided to assess the internal structure on the first level of analysis and only after doing that move to assessing structure on the between-country level. To this end all competing models for the within-level structure were estimated using the unrestricted (M0 - fully saturated or maximum) model for the between-level. This means that we did not impose any structure on the data on this level. Only after establishing the within-country structure we have moved to testing the between-country factor structure. It is one of the recommended approaches for testing multilevel CFA models, especially when establishing level-specific fit problems is important (Ryu, 2014; Wu, Lee, & Lin, & 2018). Models B1-B8 replicates the within structure on the between-country level with latent item means (*r*, *f*, *e*, *h*) and latent between factors (*Mb*, *Rb*, *Fb*, *Eb, Hb*).

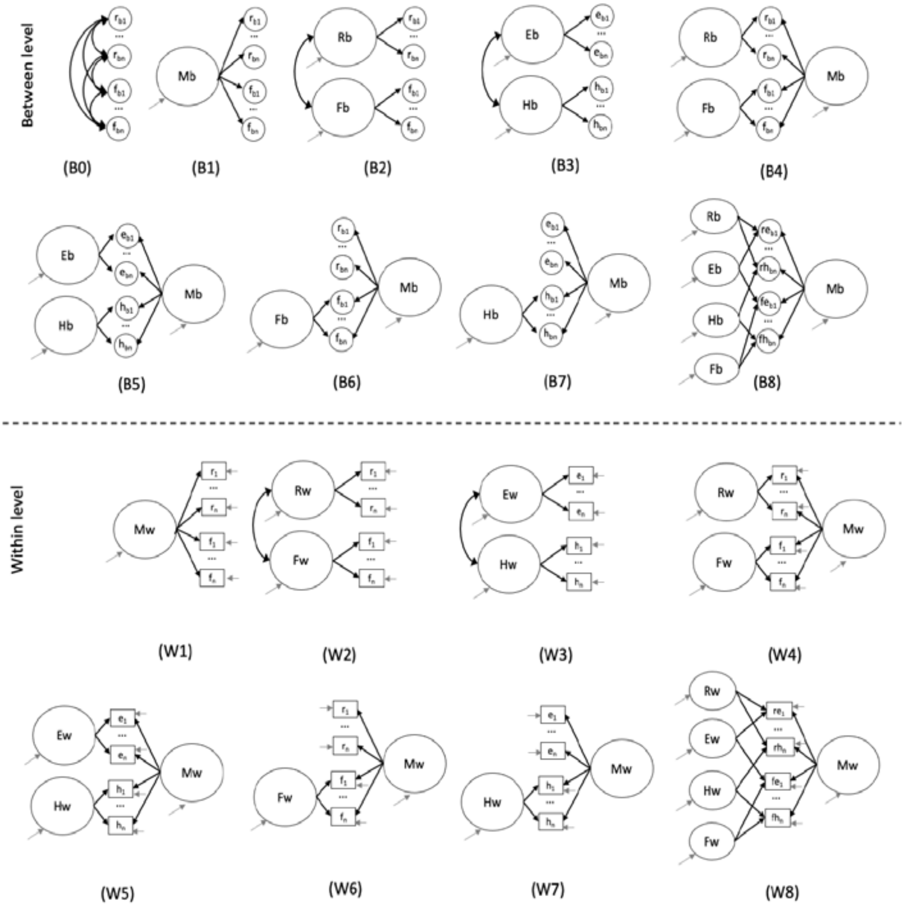

Figure 1:

Different specifications of within and between level part of the CFA model used in the analysis.

The items were assigned to reals and foils dimensions basing on their ontic status (OECD, 2014) and to hard and easy dimensions on the basis of their empirical difficulty - items with the mean value higher than the mean for all items were classified as hard and items below this criterion were classified as easy (see Table 1 for exact values).

The overclaiming technique items measured on a 5-point rating scale were treated as continuous indicators. We deemed this approach to be justified in the light of simulation studies that have shown that when the number of categories is at least four, parameters estimated using assumption of the continuous nature of indicators through maximum likelihood estimation are accurate, good approximation of categorical data modelling

(Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Johnson & Creech, 1983; Muthén & Kaplan, 1985; Pokropek, Davidov, & Schmidt, 2019; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Models with continuous indicators reduce the complexity of the estimation and avoid convergence problems for models with many dimensions (DiStefano, 2002; Dolan, 1994).

Model evaluation was based on the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC, Schwarz, 1978) as well as approximate fit indices of which we report the root mean square error of approximation (RMSEA), the comparative fit index (CFI), the Tucker-Lewis index (TLI) and the standardised root mean square residual (SRMR), the latter calculated on both within and between level (Kline, 2011, pp. 199-209). The value of the χ^2 was also reported.

Restricted maximum likelihood (MLR) estimator was used to estimate all the models presented. MPlus 8.5 was implemented for all the calculations (Muthén & Muthén, 1998-2017).

In order to improve the interpretation of the models the so-called bifactor ancillary statistical indices were calculated (Rodriguez, Reise & Haviland, 2016). To this end automatic scripts embedded in the Dueber's (2017) calculator were used. These indices are new developments and to our knowledge their properties were not thoroughly tested in multilevel models. However, they are used and reported in the multilevel bifactor models (e.g. Wang, Kim, Dedrick, Ferron, & Tan, 2018), thus we also decided to present them here in order to enhance results interpretation.

Results

Intraclass Correlations for overclaiming technique items

The analysed ICCs for overclaiming technique items showed that most of them have moderate intra-class correlations, in the range of 0.15-0.30, with the exception of two foils and Q08. It is noteworthy that the third foil, Q04, has a notably higher ICC than the other two.

Internal consistency analysis

The three foils treated together yielded Cronbach's $\alpha = 0.66$ with an average inter-item covariance of 0.70. The analysis indicated that the subscale's reliability would be higher with item Q04 eliminated. The remaining reals had $\alpha = 0.86$ with an average inter-item covariance equal to 0.70. If all the items would be treated as one scale the $\alpha = 0.86$, but the average inter-item covariance dropped to 0.61.

Confirmatory Factor Analysis (CFA)

The analysed model fit statistics indicated the combined bifactor solution (W8) as clearly the best fitting specification for the within-country model. Other bifactor models fit data slightly worse, the one with easy-hard specific factors (W5) being preferred over the reals-foils model (W4), and the one with only one specific factor grouping hard items (W7), as indicated by AIC, BIC and SRMR values. It is worth noting, that regarding simpler two-dimensional solutions reals-foils specification (W2) beats easy-hard specification (W3) in terms of model fit.

Regarding the between-countries structure the situation is a bit less clear: CFI, AIC, and SRMR (between) values favour the combined bifactor model (B8), whereas BIC indicates the two-factor model reals-foils (B2) as the best-fitting. TLI, and RMSEA are inconclusive, as their values hardly differ among the compared models. The latter model is more parsimonious, however, only the former model enables to meet the conventional standard of SRMR below 0.1 to achieve an “acceptable” fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Therefore, we decided to display and comment on factor loadings from the combined bifactor model (B8).

Table 2:
Fit statistics of the CFA models

Specification		**No. param.**	χ^2	**df**	***CFI***	***TLI***	***AIC***	***BIC***	***RMSEA***	***SRMR***	
Wi	**Bi**									**Wi**	**Bi**
W1	B0	184	3207.2	104	0.904	0.779	15307076	15309035	0.010	0.060	0
W2	B0	185	2486.0	103	0.926	0.828	15248034	15250004	0.009	0.063	0
W3	B0	185	2863.3	103	0.915	0.801	15275582	15277552	0.009	0.057	0
W4	**B0**	200	1573.5	88	0.954	0.875	15172139	15174268	0.007	0.037	0
W5	B0	200	1621.5	88	0.953	0.871	15157535	15159664	0.007	0.033	0
W6	B0	187	2067.0	101	0.939	0.856	15214718	15216709	0.008	0.046	0
W7	B0	191	1712.2	97	0.950	0.876	15179793	15181827	0.007	0.036	0
W8	B0	216	976.2	72	0.972	0.907	15119472	15121772	0.006	0.027	0
W8	B1	112	2337.1	176	0.933	0.909	15119593	15120786	0.006	0.027	0.134
W8	B2	113	2322.7	175	0.934	0.909	15119551	15120754	0.006	0.027	0.120
W8	B3	113	2326.6	175	0.933	0.909	15119564	15120767	0.006	0.027	0.129
W8	B4	128	2146.3	160	0.939	0.908	15119544	15120907	0.006	0.027	0.107
W8	**B5**	128	2130.4	160	0.939	0.909	15119527	15120889	0.006	0.027	0.115
W8	B6	115	2297.1	173	0.934	0.909	15119553	15120778	0.006	0.027	0.119
W8	B7	119	2247.1	169	0.936	0.909	15119545	15120812	0.006	0.027	0.110
W8	B8	144	1933.3	144	0.945	0.908	15119514	15121048	0.006	0.027	0.082

Note. N = 310 965, Number of groups = 67. Bolded models indicate that the model encountered problems with estimation resulting in negative residual variances for given items on a given level of analysis (W – within, B – between). In case of the W4 B0 model the item Q08W displayed such problems and in case of the W8 B5 model it was the Q03B.

Table 3:
Factor loadings on the individual level of the CFA model with combined bifactor structure on both levels (W8, B8)

	Factors				
	Mw	**Rw**	**Fw**	**Ew**	**Hw**
Q01 (h)	**0.53**	0.13			0.14
Q02	**0.58**	-0.06		0.06	
Q03	**0.68**	0.18		-0.15	
Q04* (h)	**0.47**		0.10		0.28
Q06	**0.67**	0.02		-0.05	
Q07 (h)	**0.57**	0.01			0.16
Q08 (h)	**0.50**	-0.18			**0.61**
Q09	**0.68**	**-0.53**		-0.06	
Q10	**0.59**	-0.24		0.00	
Q11* (h)	**0.31**		**0.45**		**0.39**
Q12	**0.55**	-0.13		0.27	
Q13* (h)	0.29		**0.72**		**0.37**
Q15	**0.52**	-0.03		0.27	
Q16 (h)	**0.59**	0.09			-0.01
Q17	**0.57**	0.02		0.27	
Q19	**0.54**	-0.04		**0.31**	

Note. N = 310 965, Number of groups = 67. Factor loadings with absolute value above .30 are in bold. Foil items are denoted with *. Hard items are denoted with (h).

On the within-country level the general factor represents variance of all the items quite well, maybe except Q11 and Q13 (both foils), which have very low loadings. The third foil, Q04, also does not have a substantial loading, however, it is of reasonable size, well above the customary threshold of 0.30. The specific factors are not represented well - the items ascribed to the easy factor in general were loaded under this dimension only below the mentioned threshold (with the sole exception of Q19). The specific factor for hard items accounts well for Q11, Q13 (foils), and Q08 (item that often loads on the same factor as foils), but yields a much smaller loading on Q04 (foil). The specific factor for foils is mainly represented by a sizeable loading on item Q13, much more modest loading on Q11 and a negligible loading on Q04. On the other hand, the specific factor for reals mostly loads on Q09 (negative loading) and is characterised by a group of much smaller loadings on other items.

On the between-country level the general factor represents the items to a lesser degree than on the within-country level, with many items of very small loadings. All three foils yielded a sizable loading on their respective specific factor, whereas items Q02, Q07, Q10, Q16 were loaded below even the minimal threshold under the specific factor for reals. Similar pattern concerns also loadings for easy and hard specific factors: some items are not loaded by them at all, whereas others display significant loadings.

Table 4:
Factor loadings on the between-country level of the CFA model with combined bifactor structure on both levels (W8, B8)

	Factors				
	Mb	**Rb**	**Fb**	**Eb**	**Hb**
Q01 (h)	-0.14	**0.55**			**0.36**
Q02	**0.51**	0.20		**0.53**	
Q03	**0.32**	**0.47**		0.04	
Q04* (h)	0.25		**0.54**		0.22
Q06	0.25	**0.30**		-0.10	
Q07 (h)	**0.72**	0.10			0.26
Q08 (h)	0.20	**0.34**			**0.80**
Q09	**0.60**	**0.48**		**0.41**	
Q10	0.10	0.23		**0.65**	
Q11* (h)	0.24		**0.51**		**0.48**
Q12	**0.45**	**0.52**		-0.10	
Q13* (h)	0.28		**0.73**		0.29
Q15	-0.06	**0.83**		**0.42**	
Q16 (h)	**0.50**	0.18			-0.03
Q17	**0.49**	**0.38**		0.15	
Q19	-0.01	**0.40**		**-0.39**	

Note. N = 310 965, Number of groups = 67. Factor loadings with absolute value above .30 are in bold. Foil items are denoted with *. Hard items are denoted with (h).

Bifactor ancillary indicators

The bifactor ancillary indices are presented in Table 5. The explained common variance (ECV 1 & 2) indicates the proportion of common variance explained by the general factor. Values above 0.80 are considered indicative of an essential scale's unidimensionality. In case of the PISA 2012 overclaiming scale the value of ECV for general factor (0.68) indicates that this scale should not be modelled by a one-dimensional model. The ECV 1 values show that specific factors account for some portions of the common variance (0.049 and 0.109 for easy items and hard items, respectively, 0.063 for reals and 0.102 for foils). Moreover, the ECV 2 values indicate that the specific factors account for a larger portion of the common variance in case of the items loaded by them: values of 0.087 and 0.247 for easy-hard and 0.080 and 0.490 for reals-foils are more than double in comparison to the ECV 1 values (save the value for reals; see Stucky & Edelen, 2015 for more information on ECV 1 & 2). Omega indices for both general and specific factors, representing internal reliability under the model fitted to the data, indicate good reliability of all factors specified in the model as they are above the 0.80 threshold or close to it (Reise, Bonifay, & Haviland, 2013), save the factor for foils, that falls below the assumed threshold. Omega hierarchical (ω_H), denoting what proportion of variance in total (factor) scores can be attributed to the general factor, yielded a value of 0.83. The comparison between the values of ω and ω_H in this model shows that only around of 7.5 % of the total variance can be attributed to the specific factors (ω-ω_H) and around 9.8 % to random error (1-ω ; see Rodriguez et al., 2016 for more on these indices). Omega HS (ω_{HS}) index reflects the reliability of a specific factor score after controlling for the variance attributable to the general factor (Rodriguez et al., 2016). The values ranging between 0.009 and 0.303 show that scores of the specific factors are not reliable and using them as outright variables, e.g. in an SEM model is not warranted. It seems that this is mainly due to the fact that little variance remains after accounting for the general factor (see DeMars, 2013 for more on score interpretation in bifactor models).

Table 5:
Bifactor ancillary indicators of scales in the CFA model with combined bifactor structure on both levels (W8, B8)

	ECV (1)	***ECV (2)***	**ω / ω_S**	**ω_H / ω_{HS}**	**Relative ω**	***H***	***FD***
Within-country level							
General Factor	0.678	0.678	0.902	0.827	0.917	0.883	0.938
Specific Factor 1- Real items	0.063	0.080	0.890	0.009	0.010	0.363	0.737
Specific Factor 2- Foil items	0.102	0.490	0.716	0.303	0.424	0.572	0.794
Specific Factor 3- Easy items	0.049	0.087	0.860	0.024	0.027	0.273	0.593
Specific Factor 4- Hard items	0.109	0.247	0.805	0.189	0.235	0.512	0.746
Between-country level							
General Factor	0.279	0.279	0.882	0.333	0.378	0.772	0.899
Specific Factor 1- Real items	0.290	0.363	0.874	0.485	0.555	0.814	0.918
Specific Factor 2- Foil items	0.133	0.661	0.777	0.520	0.669	0.658	0.836
Specific Factor 3- Easy items	0.152	0.282	0.839	0.091	0.108	0.636	0.828
Specific Factor 4- Hard items	0.147	0.317	0.817	0.322	0.394	0.710	0.875

Note. ECV – explained common variance. H – construct replicability. FD – factor scores determinacy.

The H values, often denoted as "construct replicability", namely how well is a given latent variable represented by the set of observable items and how well is this latent variable expected to replicate in other studies (Hancock & Mueller, 2001), indicate that only the general factor can be expected to replicate in future studies, as it exceeded the 0.70 minimum threshold and even the more demanding 0.80 threshold, as advocated by the authors of these measures. Neither of the H values for specific factors was close to reaching this value.

The factor determinacy (FD) statistic denotes the correlation between factor scores and the latent factors, with values above the 0.90 threshold assumed to indicate that respective factor scores can be used as variables on their own (e.g. in a regression or an SEM model; Rodriguez et al., 2016). Again the general factor reaches this goal while three out of four specific factors rank lower, with factor determinacy between 0.70 and 0.80. The specific factor for easy items stands out from the others with factor determinacy falling below 0.60.

When analysed for the between-country level these indices show that the general factor accounts for much less of the variance when compared with within-country level. They also show that almost 55 % of the total variance can be attributed to the specific factors - a value much higher than on the other level of analysis. Also the proportion of variance attributed to the random error is doubled here (18 %) in comparison with the within-country level.

Table 6:
Bifactor ancillary indicators of items on the within-country level in the CFA model with combined bifactor structure on both levels (W8, B8)

Item	Difficulty	IECV	ARPB
Q01	hard	0.888	0.010
Q02	easy	0.977	0.002
Q03	easy	0.893	0.089
Q04*	hard	0.713	0.096
Q06	easy	0.994	0.037
Q07	hard	0.925	0.025
Q08	hard	0.379	0.157
Q09	easy	0.621	0.010
Q10	easy	0.859	0.024
Q11*	hard	0.211	0.267
Q12	easy	0.772	0.033
Q13*	hard	0.114	0.293
Q15	easy	0.784	0.052
Q16	hard	0.975	0.036
Q17	easy	0.816	0.023
Q19	easy	0.748	0.024
Average ARPB			0.074

Note. IECV – item explained common variance. ARPB - absolute relative parameter bias. * denotes foils.

The item explained common variance (IECV) values act as a measurement of item-level unidimensionality with values above 0.80-0.85 indicating sufficient representation of the general dimension by the item variance (Stucky & Edelen, 2015). According to these values most of the easy items could be successfully modelled by a unidimensional model, with some exception of items Q12, Q15 and Q19. Among the hard items, some reals, e.g. Q01, Q07 and Q16 are good representations of the general factor, however all the foils and Q08 additionally, are not adequately accounted for by this dimension. It is interesting that among the three foils two are almost not represented by the general factor at all (Q11 and Q13), whereas the remaining Q04 is fairly well accounted for by the general dimension with the IECV value of 0.71.

The absolute relative parameter bias (ARPB) denotes differences between item loadings in the unidimensional solution and in the bifactor model (Dueber, 2017). Values around 0.10-0.15 are considered maximal for the average ARPB if a given model is to be accepted as a unidimensional one (Muthén, Kaplan & Hollis, 1987). The values obtained in this analysis denote that only Q11 and Q13 cannot be represented by a unidimensional model. Curiously enough, the third foil, Q04, was further from the above-mentioned threshold than many of the reals (e.g. Q12, Q15 or Q08).

Discussion

Overclaiming technique's internal structure on the within-country level

Our hypothesis was confirmed as the PISA overclaiming scale indeed better fitted to the multi-factor solution than to a unidimensional structure. However, the best fitting structure did not suit the assumption that foils are a pure measure of bias, and reals constitute a pure measure of math ability. The CFA revealed items loaded only to a minimal extent under factors to which they were supposed to be attached (e.g. Q04 or Q08) and yielded the combined bifactor solution as the best fitting model on both levels of analysis. This was counter to the theoretical assumption, where a much simpler structure was expected. The CFA bifactor solution obtained in this analysis yielded a different pattern, where most notably the specific factor for easy items very weakly represented variance of its items, however, the bifactor model with only one criterion used to create specific factors failed to yield better fit than the model with four cross-loading specific factors. While it is common for indices of response bias to measure both construct and bias variance (Khorramdel, von Davier, Roberts, Bertling, & Kyllonen, 2017), this often promotes bifactor CFA solutions (Reise et al., 2016; von Davier & Khorramdel, 2013).

The ancillary bifactor indices offered additional support for the bifactor structure, indicating that the unidimensional model was not appropriate for the data and that the general factor was a reliable and replicable representation of the general variance in the data. However, all the specific factors, especially the ones for real items and for easy items, displayed only low levels of reliability indicating that their scores (residual scores of the common variance controlled for the general factor; DeMars, 2013) should be treated with extra care when used as dependent or independent variables in other models, e.g. regression equations or SEMs.

Overclaiming technique's internal structure on the between-country level

The CFA analysis of the between-country structure also yielded the combined bifactor solution as best fitting the data. As it can be seen in Table 4 the items were loaded by general factor to a lesser degree than in case of the within-country level analysis, with some items (e.g. Q01, Q06, Q19) loaded only minimally. Moreover, it has to be emphasised that the between-country variance is accounted for by our models to a lesser degree than the within-country one. It can also be noticed that the SRMR (between) value indicated a lower fit of the between-country level than the within-country one. Such a pattern may be due to unmodelled variance sources, present at the country level, but absent (or less pronounced) at the individual level. Tendency to responding stylistically is a plausible candidate for this source, as countries are known to vary widely not only in the tendency to overclaim foil familiarity (Vonkova et al., 2018), but also in their tendencies to response styles (He & van de Vijver, 2015; 2016; Khorramdel, von Davier, & Pokropek, 2019). To our best knowledge no study analysed relations between overclaiming and response styles in a cross-country inquiry. This idea seems a promising idea for future studies.

Role of item difficulty in overclaiming technique

Therefore, it seems that item difficulty interfered with items' ontic status which resulted in obtaining only tentative factorial solutions that cannot comprehensively inform substantial questions. However, the solution obtained might suggest that item difficulty and plausibility can elicit distinct response processes. Hence, probably also other measurement tools are able to capture bias variance even without creating fake items (cf. Wiltermuth, 2011), that are notoriously difficult to develop (Franzen & Mader, 2019; Goecke et al., 2020; Ziegler, Kemper, & Rammstedt, 2013) and can easily change their status (Paulhus et al., 2003). On the one hand, this result is encouraging because it makes sense that for students that do not know the concepts presented in a scale the foils are not distinguishable from the reals and at least for hard items foils could be effectively used as a control instrument. On the other hand, the results indicate that using foils may be more complicated than it was believed before.

Moreover, it was suggested that participants change their response behaviour depending on the difficulty of the whole scale - when scale is easier, they overclaim more (Atir, Rosenzweig, & Dunning, in review). It is not known which overclaiming technique variations would result in more valid scores - the one from foils embedded in hard, or easy reals? The one with foils closely resembling real items, or the one with foils very distinct from them? Such questions, similarly as other issues concerning overclaiming scales construction, e.g. best reals-to-foils proportion, remain unanswered and to be tested. It is also to discern what other item characteristics could have played a role in shaping overclaiming scales internal structure. It is noteworthy, that such studies would be informative not only for the overclaiming research, but also for any survey measuring skills - and these abound in the applied fields, such as applicant selection - as these item characteristics interact with processes used to respond to foils and reals alike.

Overclaiming technique construction characteristics

The analyses presented emphasise the need of both reals and foils to be carefully piloted and matched on their difficulty and other characteristics such as word length, composition and similarity to other concepts (cf. Goecke et al., 2020; Hargittai, 2005). Linguistic studies on word recognition offer here an ample source of knowledge from which future designs of overclaiming technique should not hesitate to dip up. Moreover, such studies should also account for the crossed characteristics of items in the PISA 2012 math familiarity scale where all foils consisted of two words, whereas reals consisted of both one and two words. An experimental study that would systematically compare overclaiming technique scores for different types of foils (cf. Hargittai, 2005) and contrast them with responses to very difficult reals (e.g. concepts that could be hardly expected to be known in a given sample) would potentially bring valuable data on the processes observed in the above analysis and could help to avoid creating foils that are so easily endorsed by participants as Q04, and items that are reals but seem to be responded to as if they were foils (e.g. Q08). Other important item and scale characteristics should also be tested in a similar way in order to avoid problems with unwanted foil similarity to real concepts, as evidenced by the PISA 2012 item Q04 (“proper number”) that in many languages, including English, is a term present in everyday, even colloquial language (cf. with Ziegler et al., 2013 who removed foils strongly resembling real words from their version of the overclaiming technique; also see Goecke et al., 2020). This fact may explain why this item behaved more like a real than a foil in the conducted analyses. It is worthy to point out that also other problematic item from the scale, Q08 (“complex number”), could be, at least in some languages, interpreted as a term stemming from natural, not technical language, hence its associations could be different from the ones assumed.

It would be also interesting to verify how math ability level influences the observed relationships. Whether foils are just difficult reals for both high-achieving and low-achieving students? Interestingly, it was suggested before that existent, but just very difficult items may also work as foils in an overclaiming measure (cf. Hoffman, Diedenhofen, Verschuere, & Musch, 2015; Wiltermuth, 2011, but also note that these studies used a reward for top-scoring participants as an incentive to overclaiming). It seems that this effect has been somewhat replicated here as foils also show substantial loadings from the factor for hard items. Such a pattern of results points that more knowledge should be gathered on the implications of various foil-construction rules and that such differently-created foils should be tested in order to assess their validity (Franzen & Mader, 2019). Importantly, future overclaiming technique scales should contain not only more carefully piloted foils, but also a larger number of them, as it is known that the indices based on a small number of items are inherently in peril of low reliability (Dunlop et al., 2017). A good example of such a newly developed overclaiming instrument is the scale presented by Goecke and collaborators (2020). It would be interesting to verify findings of the above analyses on this instrument or on the most popular version of the overclaiming technique – the Paulhus’ overclaiming questionnaire (Paulhus et al., 2003).

Moreover, in case of deviation between the intended (theory-based) and empirical internal structure it is warranted to investigate the cause of this deviation. As it was put by Rost (2002, p. 108) “the source of model violation may lie with the items, the persons, the response categories, or the latent variable to be measured”. It is up for future research on overclaiming scales to verify what tool characteristics can cause such deviation as we have observed in our research. Analysing overclaiming scales responded to under different, experimentally manipulated conditions seems an especially promising avenue for this research. Different instructions, rating scales’ formats, time limits and paradata measures (e.g. measuring response times, tracking mouse movements or employing eye-tracking devices) stack out as important ideas to test and measure (Horwitz, Kreuter, & Conrad, 2017; Khorramdel, 2014; Khorramdel & Kubinger, 2006; Maricuțoiu, & Sârbescu, 2019). It would be also important to expand the research of Yang and colleagues (2019) and perform latent class analysis of an overclaiming scale, also with the aim to verify whether different latent classes would be characterised also by different scale dimensionality (Formann, 2002).

Covariates of overclaiming technique's internal structure

Another important direction for the future is to clarify what covariates influence overclaiming technique's internal structure. Does the scale have the same structure for low- and high-competent respondents (cf. Gnambs & Schroeders, 2020)? Do men and women differ in this regard? What is the level of the cross-country invariance of the scale? How can the now tangled interpretation of the correlation between reals and foils be explained? The above analyses show that this correlation probably has a substance (e.g. math ability) as well as bias (e.g. response styles) substrate, but its dependence on scale characteristics, response behaviour, and participants' covariates is a matter to pinpoint for future research. One of the conceptions was recently rejected, as general intelligence scores failed to moderate the reals-foils correlation (Goecke et al., 2020).

Future research attempts should also account for response styles in future studies on overclaiming technique's internal structure, especially those comparing cross-country data. Models presented by Aichholzer (2014; 2015) seem a promising avenue to control acquiescence's role in the formation of specific factors (for foils or hard items) and models developed by Khorramdel and von Davier (2014) can account for other response styles.

Conclusions

The presented analyses found that the PISA 2012 overclaiming scale internal structure is not unidimensional, but forms a bifactor structure with several specific factors on both within- and between-country levels. These factors cannot be simply interpreted as reals (construct variance) and foils (bias variance) as both types of items seem to measure both types of variance. Most probably, this is the main reason that a combined bifactor solution fitted the data best in the CFA. It does not necessarily mean that item difficulty elicits different response mechanisms and that the intuitive reals-foils distinction should be abandoned, rather it calls for care when using and interpreting subscores based on reals and foils as they are not pure measures of construct or bias, respectively. This evidence is not enough to claim about the nature of mechanisms engaged in responding to reals and foils, but it is probable that the scale is essentially unidimensional and measures only one cognitive process and the specific factors are mainly outcomes of some spurious variance, e.g. response styles (cf. with the research on the factorial structure of the Rosenberg self-esteem scale; Gnambs et al., 2018).

Careful overclaiming scale preparation is warranted, as it was evidenced that the simple ontic status does not guarantee the intended item measurement characteristics. Ideas for future overclaiming scales construction are presented, nevertheless, all of them have to be tested empirically. More internal structure studies should follow these future overclaiming scales in order to certify their designed measurement properties and interpretability, especially if any subscores are to be used. Overclaiming technique is a promising tool able to account for self-assessment imprecision, improve cross-country comparability, and screen for fakers in high-stakes contexts. It is desirable to develop this potential further on, but using more carefully prepared scales with clearer internal structure that the one employed in the PISA 2012. The investigation of covariates affecting overclaiming scale internal structure is also advised, with participants' cognitive abilities and response styles on top of that list.

Limitations

The main limitation of the study was its largely exploratory character. All the results presented here need to be interpreted with due cautiousness, especially the implications of overclaiming technique's internal structure on the interpretation of mechanisms driving its scores. A further study limitation is that we limited our analyses to only one measurement instrument, the PISA overclaiming technique scale, and one sample, 15-year-olds participating in this assessment. It is to be established by future studies if and how the above results generalise also to other overclaiming scales and other samples. Furthermore, we were limited by some shortcomings of the analysed scale, mainly low number of foils and some crossed item characteristics between reals and foils (e.g. foils always consisted of two words, e.g. "proper number", whereas reals were created by one or two words, e.g. "polygon", but also "complex number"). It is also to be determined in future studies whether overcoming such shortcomings would yield a clearer factorial structure.

We also did not account for possible factors associated with overclaiming technique's internal structure, most notably response styles. It is justified to claim that response styles are related to overclaiming technique's scores and that they can be especially influential on the between-country level. What is more, we have not tested the level of measurement invariance of the PISA overclaiming scale across the participating countries or language versions. The initial research attempt by Jerrim, Parker and Shure (2019) indicates that this topic is worthy to pursue. Finally, we do not present any analyses testing internal structure between different groups, e.g. male and female students or high- versus low-achieving participants.

Directions for Future Studies

Future studies should concentrate on developing better versions of the overclaiming scale, built according to advice voiced in this and other articles (especially Goecke et al., 2020). Many important items- (difficulty, length, content, word frequency, foil plausibility, etc.) and scale-characteristics (difficulty, proportion reals-to-foils, number of items, number of response categories, etc.) should be tested to enhance our understanding of their influence on overclaiming technique scores. Such research would greatly help to construct overclaiming scales with a clear and interpretable internal structure which is an indispensable prerequisite if this tool is to be further used to adjust self-report scores for response biases. Moreover, only scores obtained from such measurement tools could be fully informative on somewhat elusive mechanisms causing overclaiming.

When answering overclaiming items, especially foils, participants certainly use various strategies that may depend on their motivation, cognitive abilities or item difficulty. It seems that researching them with eye- and mouse-tracking devices, verbal protocols, cognitive interviews or metacognitive self-reports – as it is already done in intelligence tests (cf. Chuderski et al., 2020) – could bring very informative results, useful not only for overclaiming research, but also to the “questionnaire science” in general.

Other groups of studies should concentrate on possible relations between overclaiming technique scores and response styles. Such a study would not only inform substantial knowledge on overclaiming behaviour, but would also help to model the between-country internal structure as response styles are suspected to be responsible for much of the unmodelled cross-country variance. It is probable, that this mediation would be related to participants’ cognitive abilities (cf. Gnambs & Schroeders, 2020).

NOTE

This research is financed by the National Science Centre (NCN) research grant (2019/33/B/HS6/00937) *Understanding response styles in self-report data: consequences, remedies and sources*.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1–4.

Aichholzer, J. (2015). Controlling acquiescence bias in measurement invariance tests. *Psihologija, 48*(4), 409–429.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

American Educational Research Association American Psychological Association National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*(8), 1295-1303.

Atir S., Rosenzweig, E, & Dunning D. A. (in preparation). The Influence of Context on Overclaiming: When and Why Do People Claim to Know The Unknowable? https://www.stavatir.com/papers

Barber, L. K., Barnes, C. M., & Carlson, K. D. (2013). Random and Systematic Error Effects of Insomnia on Survey Behavior. *Organizational Research Methods, 16*(4), 616–649.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203.

Bing, M. N., Kluemper, D., Davison, H. K., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes, 116*(1), 148-162.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111-150.

Brutus, S., Gill, H., & Duniewicz, K. (2010). State of science in industrial and organizational psychology: A review of self-reported limitations. *Personnel Psychology, 63*(4), 907-936.

Brückner, H. (2009). Surveys. In P. Hedström, & P. Bearman, *The Oxford Handbook of Analytical Sociology*, pp. 666-687. New York, NY: Oxford University Press.

Campbell, A. C. (1963). Internal Structure of Items, Item Difficulty, and Solution Processes: Rejoinder to Silverstein and MCLain. *Psychological Reports*, 13(3), 753-754.

Chuderski, A., Jastrzębski, J., Kroczek, B., Kucwaj, H., & Ociepka, M. (2020). Metacognitive experience on Raven's matrices versus insight problems. *Metacognition and Learning*, 1-21.

Danek, A. H., & Wiley, J. (2017). What about false insights? Deconstructing the Aha! experience along its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology*, 7, 2077.

DeMars, C. E. (2013). A Tutorial on Interpreting Bifactor Model Scores. *International Journal of Testing, 13*(4), 354–378.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327-346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309-326.

Dueber, D. M. (2017). Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. https://doi.org/10.13023/edp.tool.01

Feeney, J. R., & Goffin, R. D. (2015). The overclaiming questionnaire: A good way to measure faking?. *Personality and Individual Differences, 82*, 248-252.

Ferrando, P. J. (2005). Factor analytic procedures for assessing social desirability in binary items. *Multivariate Behavioral Research, 40*(3), 331-349.

Formann, A. K. (2002). Identifying types, response errors, and unscalable respondents from personality questionnaires. *Psychological Test and Assessment Modeling, 44*(1), 78-93.

Franzen, A., & Mader, S. (2019). Do phantom questions measure social desirability?. *methods, data, analyses, 13*, 37-57.

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg Self-Esteem Scale. *Zeitschrift für Psychologie*, 226, 14-29.

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, *27*(2), 404-418.

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence, 81*, 101470.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social Science Computer Review, 23*(3), 371-379.

He, J., & Van de Vijver, F. (2016). Correcting for Scale Usage Differences among Latin American Countries, Portugal, and Spain in PISA. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa, 22*(1), 1-11.

He, J., & Van De Vijver, F. J. R. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly, 79*(S1), 267–290.

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, 62, 403-414.

Horwitz, R., Kreuter, F., & Conrad, F. (2017). Using mouse movements to predict web survey response difficulty. *Social Science Computer Review, 35*(3), 388-405.

Hülür, G., Wilhelm, O., & Schipolowski, S. (2011). Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. *Learning and Individual Differences, 21*(6), 742-746.

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence*, 40(5), 427-438.

Jerrim, J., Parker, P., & Shure, N. (2019). *Bullshitters. Who Are They and What Do We Know about Their Lives?*, IZA Discussion Papers, No. 12282, Institute of Labor Economics (IZA), Bonn. https://www.econstor.eu/bitstream/10419/196780/1/dp12282.pdf

Joseph, J., Berry, K., & Deshpande, S. P. (2009). Impact of emotional intelligence and other factors on perception of ethical behavior of peers. *Journal of Business Ethics, 89*(4), 539–546.

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 398-407.

Kam, C., Risavy, S. D., & Perunovic, W. E. (2015). Using Over-Claiming Technique to probe social desirability ratings of personality items: A validity examination. *Personality and Individual Differences, 74,* 177-181.

Khorramdel, L. (2014). The influence of different rating scales on impression management in high stakes assessment. *Psychological Test and Assessment Modeling, 56*(2), 154-167.

Khorramdel, L., & Kubinger, K. D. (2006). The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychology Science, 48*(3), 378-397.

Khorramdel, L., von Davier, M., Bertling, J. P., Roberts, R. D., & Kyllonen, P. C. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study. *Psychological Test and Assessment Modeling, 59*(1), 71-92.

Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology, 72*(3), 538-559.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*(4), 2025-2047.

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In: L. Rutkowski, M. von Davier, & D. Rutkowski. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 277-285. Boca Raton, FL: CRC Press.

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research, 45*(2), 271-293.

Lucas, R. E., & Baird, B. M. (2004). Extraversion and emotional reactivity. *Journal of Personality and Social Psychology, 86*(3), 473-485.

Maricuțoiu, L. P., & Sârbescu, P. (2019). The relationship between faking and response latencies. *European Journal of Psychological Assessment, 35*, 3-13.

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: The role of over-claiming, self-esteem, and emotional intelligence. *Psychology Science, 48*(3), 336-356.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Musch, J., Ostapczuk, M., & Klaiber, Y. (2012). Validating an inventory for the assessment of egoistic bias and moralistic bias as two separable components of social desirability. *Journal of Personality Assessment, 94*(6), 620-629.

Muszyński, M. (2020). Validity of the overclaiming technique as a method to account for response bias in self-assessment questions. Analysis on the basis of the PISA 2012 data. Unpublished doctoral thesis. Jagiellonian University, Krakow, Poland. https://ruj.uj.edu.pl/xmlui/bitstream/handle/item/251494/muszynski_validity_of_the_overclaiming_technique_as_a_method_2020.pdf?sequence=1&isAllowed=y

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Müller, S., & Moshagen, M. (2018). Overclaiming shares processes with the hindsight bias. *Personality and Individual Differences, 134*, 298-300.

Müller, S., & Moshagen, M. (2019a). Controlling for response bias in self-ratings of personality: A comparison of impression management scales and the overclaiming technique. *Journal of Personality Assessment, 101*(3), 229-236.

Müller, S., & Moshagen, M. (2019b). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment, 31*(2), 181-191.

OECD (2014). *PISA 2012 Technical Report.* PISA, OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (eds.), *Measures of social psychological attitudes, Vol. 1. Measures of personality and social psychological attitudes* (p. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (2002). Socially Desirable Responding: The Evolution of a Construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (eds.), *The role of constructs in psychological and educational measurement* (Issue 2002, pp. 49–69). Lawrence Erlbaum.

Paulhus, D. L. (2012). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 151-164). New York, NY: Oxford University Press.

Paulhus, D. L., & Dubois, P. J. (2014). Application of the Overclaiming Technique to Scholastic Assessment. *Educational and Psychological Measurement, 74*(6), 975–990.

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

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In: Richard W. Robins, R. Chris Fraley, Robert F. Krueger, *Handbook of research methods in personality psychology*, pp. 224-239. New York, NY: The Guilford Press.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of" social desirability" in survey studies. *American Journal of Sociology, 77*(5), 921-940.

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

Pokropek, A. (2014). Dekonstrukcja skal szacunkowych. Przykład skali znajomości pojęć matematycznych uczniów w PISA 2012. In: B. Niemierko & K. Szmigel (eds.) *Diagnozy edukacyjne. Dorobek i nowe zadania.* Gdańsk: PTDE. [Deconstructing rating scales. An example of PISA 2012 math familiarity scale]. http://www.ptde.org/pluginfile.php/879/mod_page/content/2/Archiwum/XX_KDE/pdf_2014/Pokropek.pdf

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(5)*,* 724-744.

Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics, 10*(11), 805-817.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126-136.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354-373.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150.

Rost, J. (2002). When personality questionnaires fail to be unidimensional. *Psychological Test and Assessment Modeling, 44*(1), 108-125.

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, *5*, 81.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8, 23-74.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20*(1), 51–68.

Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly, 24*(4), 402-433.

Steger, D., Schroeders, U., & Wilhelm, O. (2020). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*, 1073191120914970.

Stucky, B. D. & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In: S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*, pp. 183-206. New York, NY: Routledge.

von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R.E. Millsap, L.A. van der Ark, D.M. Bolt, & C.M. Woods. *New developments in quantitative psychology* (pp. 463-487). New York, NY: Springer-Verlag.

Vonkova, H., Papajoanu, O., & Stipek, J. (2018). Enhancing the cross-cultural comparability of self-reports using the overclaiming technique: An analysis of accuracy and exaggeration in 64 cultures. *Journal of Cross-Cultural Psychology, 49*(8), 1247-1268.

Wang, Y., Kim, E. S., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and Psychological Measurement, 78*(2), 253-271.

Wetzel, E., Böhnke, J. R., Brown, A. (2016). Response Biases. In: F.T.L. Leong, D. Bartram, F. M. Cheung, K.F. Geisinger, & D. Iliescu (eds.), *The ITC International Handbook of Testing and Assessment*, 349–363. New York, NY: Oxford University Press.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2), 157-168.

Woszczynski, A. B., & Whitman, M. E. (2004). The problem of common method variance in IS research. In: M. Whitman & A. Woszczynski (eds.), *The handbook of information systems research* (pp. 66-78). Igi Global.

Wu, J. Y., Lee, Y. H., & Lin, J. J. (2018). Using iMCFA to perform the CFA, multilevel CFA, and maximum model for analyzing complex survey data. *Frontiers in Psychology*, *9*, 251.

Yang, Z., Barnard-Brak, L., & Lan, W. Y. (2019). Examining the association of over-claiming with mathematics achievement. *Learning and Individual Differences, 70*, 30–38.

Ziegler, M. (2015). “F**** You, I won’t do what you told me!” - Response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*(3), 153–158.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences, 34*(1), 32–40.

Ziegler, M., MacCann, C., & Roberts, R. D. (2012). *New Perspectives on Faking in Personality Assessment*. New York, NY: Oxford University Press.