Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The cost of a reliability rule: single-valence scales and the suppressed social and emotional skills indices in PISA 2022

Pedro Velez-Pardo

pvelez@ces.edu.co

Universidad CES

Nadia Semenova Moratto-Vásquez

Universidad CES

Jesy Carolina Buitrago Salazar

Universidad CES

Research Article

Keywords: PISA 2022, social and emotional skills, mixed-worded scales, Cronbach's alpha, score suppression, matrix sampling, measurement equity

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10678718/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

PISA 2022 introduced seven scales measuring social and emotional skills. Each has ten items of mixed wording, and each student answers five. The OECD withholds a country's score when any of that country's country-by-language groups has a Cronbach's alpha below 0.60; across the seven scales the rule removed 158 of 513 administered country-by-scale indices. Whether the discarded information was recoverable, and on whom the loss falls, has not been examined.

Methods.

The OECD scaling and exclusion pipeline was reimplemented on the public-use file and validated against published item parameters, published group alphas and the individual suppression codes. Country-by-language groups were reconstructed with the OECD's own rule, giving 621 group-by-scale units. For every suppressed cell we computed the alpha a same-valence half-scale would have reached in every language group of that country, corrected for estimator bias and bounded by Feldt confidence intervals. This was compared with respondent screening, with a simulation containing no careless respondents, and with how each half's criterion validity varies with the country's rate of extreme invariant responding.

Results.

In 155 of the 158 suppressed cells the positively worded half alone would have reached 0.60 in every country-by-language group, 150 using the lower bound of the 95% interval. Respondent screening is not equivalent: discarding 15.5% of students recovers 87 cells and discarding 42.3% recovers 122; the positively worded half recovers 155 discarding none. A simulation without careless respondents reproduces the pattern: administered-form alpha falls from 0.775 to 0.511 while the positively worded half holds at 0.795. It also predicts an inverted U in the number of reverse-worded items, which the data confirm. In all seven scales it is the reverse-worded half whose relation to reading achievement weakens as extreme invariant responding rises.

Conclusions.

The threshold discards information a same-valence scale would have retained, and discards it unequally: the number of scales a country retains correlates 0.866 with its mean reading score, the lowest-performing quarter of countries retaining 1.47 of seven and the highest 6.95. Reporting decisions for mixed-worded scales should be evaluated against what a same-valence subscale would deliver, and the resulting missingness treated as informative.

Introduction

PISA 2022 added to its student questionnaire seven scales measuring social and emotional skills: assertiveness, co-operation, curiosity, emotional control, empathy, perseverance and stress resistance. Each scale contains ten items, some worded so that agreement indicates more of the attribute and some worded so that agreement indicates less. Each student answers five of the ten under a within-construct matrix sampling design. The resulting indices are already being used: a Scopus search for substantive work on the PISA 2022 social and emotional skills battery returns thirteen articles, none of which examines the measurement properties of the battery itself.

Those indices are not available for every country. The OECD applies an internal-consistency criterion before releasing them, and states it plainly in the technical report: "A group needed to have a Cronbach's alpha of at least 0.60 for a scale in order for the group's scale scores to be reported. Scale scores were suppressed for countries/economies in which one or more language groups had a Cronbach's alpha under 0.60 for the scale" (OECD, 2024, Ch. 19). The unit is the country-by-language group, and the aggregation is by the worst group: one failing group removes the country's score. For the seven scales considered here, the rule removed 158 country-by-scale indices, listed in Table 19.A.14 of that report.

A body of work explains why alphas are low on scales of this kind. Mixed item wording attracts respondents whose answers are internally inconsistent, and their presence distorts estimated dimensionality and reliability in international assessments (Steinmann et al., 2022). Inconsistent responding is not randomly distributed: it is more frequent among lower-achieving students, and among students in lower-performing systems (Steinmann et al., 2024; Wang, 2025). Explanations at the individual level point more to cognitive ability than to personality (Chen et al., 2024; Gnambs & Schroeders, 2020), and the phenomenon is not readily fixed by raising the stakes of the survey (Michaelides & Konstantinidou, 2026). Removing careless respondents improves fit and reliability in some settings (Antoniou & Alghamdi, 2024), but Bulut and Bulut (2022) found the improvement to be negligible precisely among the lowest-achieving students. Michaelides et al. (2026) document the prevalence and predictors of inconsistent responding on a mixed-worded PISA 2022 questionnaire scale across seven countries. Steinmann (2024) draws the operational conclusion for IEA studies and recommends against mixing positively and negatively worded items, particularly in populations with low reading proficiency.

None of this is in dispute here, and none of it is claimed as new. What has not been asked is the question that follows from it. If low alpha on these scales is a property of mixed wording interacting with the population, then a rule that responds to low alpha by withholding the score is making a choice with consequences, and those consequences can be quantified. Three questions follow. First, was the information the rule discarded recoverable? Would a scale composed only of same-valence items have met the OECD's own threshold in the cells where the mixed-valence form did not? Second, is the obvious alternative, screening out inconsistent respondents, an equivalent remedy? Third, who bears the loss?

This article addresses those questions, and two that support them. We ask whether the phenomenon requires careless respondents at all, or whether it follows from the reverse-worded items ceasing to load on the common factor; and we ask which of the two halves is the contaminated one, using external criteria rather than internal structure. We do not ask, and cannot answer, which half validly measures the construct; that is the article's principal limitation.

The specific research questions are:

- • **RQ1.** In the cells where the index was suppressed, would a same-valence scale have reached 0.60 in every country-by-language group of that country, in the unit in which the OECD applies its criterion?
- • **RQ2.** Is screening out inconsistent respondents an equivalent remedy?
- • **RQ3.** Does the phenomenon require careless respondents, or is it sufficient that the reverse-worded items stop loading on the common factor?
- • **RQ4.** Which half of the scale loses criterion validity where extreme invariant responding is more common?
- • **RQ5.** On which countries does the loss fall?

Methods

Data

The analyses use the PISA 2022 student questionnaire public-use file (file CY08MSP_STU_QQQ.SAV; OECD, 2024b). The seven social and emotional skills scales were administered in module 8 of the questionnaire to between 71 and 76 countries and economies each, giving 513 administered country-by-scale cells in total. Item responses are on a four-point agreement scale for the trend scales and a five-point scale for the module 8 scales; reverse-worded items are recoded before scaling, as in the OECD pipeline.

Sampling weights are used throughout. Senate weights are used for calibration, so that no country dominates the international item parameters, and final student weights are used for all within-country statistics.

The analyses cover the seven social and emotional skills scales of module 8 only. Two other questionnaire scales of mixed wording, sense of belonging at school and exposure to bullying, fall outside the scope of this article.

The rule, and the unit in which it is applied

The suppression rule and the two respondent-level exclusions that precede it are stated in the technical report and were reimplemented rather than inferred. Students who were administered items of both

valences and selected the same extreme category for all five were classified as extreme straightliners and removed from the analysis; students with fewer than three responses to a scale were likewise removed; and the alpha criterion was then applied to the surviving respondents, group by group.

The grouping rule matters and is easy to misapply. The technical report treats a language group within a country as an independent group when its sample size exceeds 150 and the sum of its weights exceeds 300. Applying that rule to the analytic sample of each scale yields between 86 and 90 groups per scale and **621 group-by-scale units**, corresponding to 92 distinct country-by-language groups. Table 19.A.11 of the technical report lists 103 groups; the eleven we do not separate fall below the size thresholds once the module-8 analytic sample is formed.

Throughout, the list of suppressed cells is taken from Table 19.A.14. The reconstruction of the rule serves only to check that the pipeline behaves as documented, and how well it does so is reported under Validation.

Estimation

Item parameters were estimated with a graded response model (Samejima, 1969) and, for the trend scales used in validation, a generalised partial credit model (Muraki, 1992), by marginal maximum likelihood with an EM algorithm and Gauss-Hermite quadrature (21 nodes in one dimension). Bifactor models with a method factor defined on the reverse-worded items were estimated on a 15 × 11 grid of 165 nodes. Person parameters were obtained by expected a posteriori scoring, and by weighted likelihood estimation (Warm, 1989) where a bias-reduced estimate was required. Item fit was assessed by RMSD, and metric linking by the mean/mean and Stocking-Lord (Stocking & Lord, 1983) methods.

Where the reverse-worded items are the majority, as in emotional control with seven of ten, the method factor spans most of the items and is not identified separately from the general factor. For that scale the alternative parameterisation, with the method factor defined on the minority valence, is the one reported.

Reliability under within-construct matrix sampling

Because each student answers five of ten items, coefficient alpha (Cronbach, 1951) cannot be computed in the usual way, and two distinct quantities must be kept apart.

The first is the **average alpha within each administered item set**, which is the quantity reported in Table 19.A.11 and the quantity to which the OECD threshold is applied. The second is **alpha computed from the pairwise covariance matrix** and projected by the Spearman-Brown formula to a five-item form, which estimates the alpha that a designated subset of items would have had if administered together. Only the second can express the counterfactual, because a same-valence five-item form was never administered.

The second estimator is optimistic. Measured against the published group alphas across the 608 groups where both are available, it exceeds the official value by a median of **0.017**, with a 90th percentile of 0.048. All counterfactual figures reported below are corrected by subtracting this bias, using the median within each scale, which ranges from 0.010 for curiosity to 0.021 for co-operation.

Interval estimates use Feldt's (1965) confidence interval for coefficient alpha. The sample size it takes is not the group's, because under the matrix design two items are both answered by the same student only when the five-item form drawn for that student happens to contain both. Once two items are fixed, the remaining three of the five are drawn from the other eight, so the proportion of students contributing to any one covariance is C(8, 3)/C(10, 5) = 56/252 = 22.2%. That proportion of the group, not the whole group, is the sample size the intervals use. Across the 203 groups belonging to suppressed cells, the median is 1147 students per item pair and the median interval width is 0.050.

The counterfactual

For each of the 158 suppressed cells, and separately for the positively worded and the reverse-worded items, we computed the bias-corrected five-item alpha **within every country-by-language group of that country** and took the minimum across groups. A cell counts as recovered when that minimum reaches 0.60 — that is, when the criterion the OECD applies, in the unit in which it applies it, would have been met. The same computation was repeated using the lower bound of the 95% confidence interval in place of the point estimate.

Because the positively worded subset is fixed in advance by item wording, the count based on it involves no selection. The count based on whichever half performs better is reported beside it, and identified as such.

Respondent screening as an alternative

More aggressive screening of inconsistent respondents might have rescued the same cells. We tested this with the mean absolute difference (MAD) index between the mean of the positively worded items and the mean of the recoded reverse-worded items, the screening statistic used by Steinmann et al. (2022). Students exceeding a MAD threshold were removed, the administered-form alpha was recomputed in every language group, and a cell was counted as recovered under the same rule as above. Four thresholds were used, from 2.5 down to 1.0.

Simulation

To test whether the phenomenon requires careless respondents, we simulated ten five-category items under a two-factor model with no careless or inconsistent respondents of any kind. The positively worded items load 1 on a general factor. The reverse-worded items load γ (gamma) on that same general factor, and 1 on a second, uncorrelated factor defined only on them, which stands for whatever

the reverse wording introduces. γ is the quantity of interest: at γ = 1 both valences measure the common attribute equally well, and at γ = 0 the reverse-worded items carry none of it. Item thresholds are identical across items. Responses were generated for 8000 simulees, five of the ten items were retained per simulee at random, and alpha was computed exactly as in the empirical analysis. Two sweeps were run: γ from 1.0 to 0.0 with five reverse-worded items, and the number of reverse-worded items from two to seven with γ fixed at 0.2. Each condition was run in **20 independent replicates**; means are reported, with the standard deviation across replicates never exceeding 0.009.

External criterion

For each scale and country, a graded response model was fitted separately to each half, every student was scored by EAP on both halves, and each half was correlated within country with six external criteria: reading and mathematics performance, economic, social and cultural status, grade repetition, skipping school, and educational expectation. None is a self-report of personality. Because the two halves are not equally reliable, and a comparison of raw correlations would favour whichever half is measured better, each correlation was divided by the square root of that half's marginal reliability, which is the usual correction for attenuation.

Achievement enters as the mean of the ten plausible values PISA supplies. The recommended procedure is to compute the statistic on each plausible value and combine the ten by Rubin's rules, and we verified that the shortcut does not matter here: across 352 country-by-scale units the two procedures correlate 0.9998, averaging inflates the absolute correlation by 0.007 at the median, and the between-country associations reported below shift by at most 0.021. Repeating the whole procedure for stress resistance with the halves refitted country by country gives the same answer, with shifts of 0.010 and 0.004.

The quantity of interest is not the size of these correlations, which is small for both halves. It is whether the relation to the criterion weakens where extreme invariant responding is more common. For each scale we correlated, across countries, the official rate of extreme straightlining with each half's disattenuated validity coefficient, and took the difference between the halves. Taking the difference is what makes the comparison informative: an explanation resting on the general structure of variances in low-performing populations would move both halves together and cancel.

Validation

The pipeline was checked against published output before any substantive analysis. Reproduction of the published mathematics anxiety index from the parameters in Table 19.A.27 gave r = 0.999991 with the published values and agreement to two decimals for 99.75% of 475792 students. Bifactor loadings were compared with those from the mirt package (Chalmers, 2012): in a well-conditioned country the correlation was 1.0000 on the general loadings and 0.9999 on the method loadings, with omega 0.6067 against 0.6082. In a country whose bifactor solution is close to degenerate, with the general loadings of the reverse-worded items near zero, both implementations reach the same degenerate solution and

agree on the resulting omega, 0.9940 against 0.9946. Our reconstruction of group alpha correlates 0.936 with Table 19.A.11 across the 539 groups where both values exist, with a median absolute difference of 0.027 and a systematic bias of − 0.025. Applied to the suppression rule itself, the reconstruction identifies 147 of the 158 cells of Table 19.A.14. Of the eleven it misses, nine have an official group alpha below 0.60 in a group we do reconstruct, so the shortfall there is the estimator's; the remaining two are Slovenian cells suppressed through a language group we do not separate. The straightlining rule reproduces the individual suppression codes for 99.4% of flagged students, and the minimum-responses rule reproduces Table 19.A.45 with r = 1.0000 across 80 countries.

The reconstruction of group alpha is approximate, and it is poor for curiosity (r = 0.642). Wherever an official alpha value is needed, the published value is used; our own is used only to calibrate the bias of the counterfactual estimator.

Software, code and use of AI assistance

Analyses were carried out in Python (numpy, pandas, scipy), with mirt in R used for independent validation. The estimator, the pipeline and every derived table are openly available; details are given in the data availability statement supplied with the submission. Claude (Anthropic) was used to assist in reviewing the analysis code and in reviewing and editing the manuscript text. The authors designed the study, specified and verified all analyses, and take full responsibility for the content and for the accuracy of all reported results.

Results

The scope of the suppression

Of 513 administered country-by-scale cells, 351 carry a published index and 162 do not. Table 19.A.14 attributes 158 of those 162 to the alpha criterion; the remaining four, all in co-operation, were suppressed because fewer than three items had international parameters. The remainder of this article concerns the 158.

The loss is not evenly spread across scales (Table 1). Curiosity, with two reverse-worded items, loses 4 of 76 countries. Emotional control, with seven, loses 12 of 73. The two scales with a five-five split lose 33 of 75 and 29 of 71.

Table 1
Administered, published and suppressed indices, by scale

Scale	Reverse-worded items	Administered	Published	Without index	Suppressed for alpha
Co-operation	5	75	38	37	33
Assertiveness	3	75	45	30	30
Stress resistance	5	71	42	29	29
Empathy	3	71	43	28	28
Perseverance	4	72	50	22	22
Emotional control	7	73	61	12	12
Curiosity	2	76	72	4	4
Total		513	351	162	158

The last column is the count of cells suppressed under the alpha criterion, from Table 19.A.14 of the PISA 2022 Technical Report. The four cells in the difference, all in co-operation, were suppressed because fewer than three items had international parameters.

Where the alpha goes

For each country we computed the difference between the mean of the two half-scale alphas and the alpha of the administered mixed-valence form. In suppressed countries that gap runs from 0.162 to 0.263 across the six scales with at least three reverse-worded items; in countries whose index was published it runs from 0.044 to 0.134. The gap separates suppressed from published countries almost perfectly in those six scales, with an area under the ROC curve between 0.880 and 0.988, although the separation is in part arithmetic: the administered-form alpha enters both the gap and the suppression decision. Figure 1 shows the same relation at the level of the country-by-language group, which is the unit the rule acts on. Of the 215 groups whose administered mixed-valence form falls below 0.60, **212 reach 0.60 on the positively worded half** and 170 reach it on the reverse-worded half.

Curiosity is the internal control. With only two reverse-worded items there is little mixture to lose: its gap is − 0.021 in suppressed countries and 0.009 in published ones, and its AUC is 0.510. The comparison rests on the four countries curiosity loses, so it carries little precision, but the point estimate shows none of the separation the other six scales show.

RQ1. The counterfactual

Table 2 gives the result. **In 155 of the 158 suppressed cells, the positively worded items alone would have reached alpha 0.60 in every country-by-language group of that country.** Using the lower bound of

the 95% confidence interval rather than the point estimate, 150 of 158.

Table 2
Suppressed cells in which a same-valence half-scale would have reached alpha 0.60

Scale	Suppressed cells	Positive half, point	Positive half, 95% LB	Reverse half, point	Reverse half, 95% LB	Either half, point	Either half, 95% LB
Co-operation	33	33	32	24	13	33	32
Assertiveness	30	29	29	20	16	29	29
Stress resistance	29	27	23	29	29	29	29
Empathy	28	28	28	14	6	28	28
Perseverance	22	22	22	22	16	22	22
Emotional control	12	12	12	12	12	12	12
Curiosity	4	4	4	1	1	4	4
Total	158	155	150	122	93	157	156

A cell counts as recovered when the half-scale reaches 0.60 in every country-by-language group of that country, which is the unit in which the OECD applies its criterion. All alphas are corrected for the bias of the matrix-sampling estimator using the median within each scale. LB = lower bound of the Feldt 95% confidence interval computed with the effective sample size. The "either half" columns involve selection after the fact and are reported for comparison only.

Three cells fail on the positively worded half: Georgia (0.554) and Peru (0.559) in stress resistance, and Cambodia (0.589) in assertiveness. All three are marginal. Cambodia in assertiveness is also the only cell that fails on **both** halves — its reverse-worded half reaches only 0.406, and it is an extreme case in its own right, with a minimum official group alpha of 0.210. Using the confidence bound, eight cells fail on the positively worded half: Georgia (0.509), Peru (0.521), Cambodia (0.549), Slovenia in co-operation (0.561), Uzbekistan (0.584), Kazakhstan (0.587), Kosovo (0.592) and the Philippines (0.593); six of the eight are stress resistance.

The two halves are not interchangeable. The reverse-worded half reaches the threshold in 122 of 158 cells on the point estimate and in 93 using the confidence bound, against 155 and 150 for the positively worded half. Counting a cell as recovered whenever *either* half reaches the threshold gives 157 of 158, and 156 using the bound. The two cells of difference are Georgia and Peru, both in stress resistance, which is the one scale where the reverse-worded half is the stronger of the two: it recovers 29 of 29 cells against 27 for the positively worded half. Because the positively worded subset is fixed by item wording while the better-performing half is identified only after the fact, 155 is the figure we rely on.

Forty-five of the 158 cells contain more than one language group, and in those the minimum is taken across groups, so the result is not an artefact of aggregating a country whose groups differ.

RQ2. Respondent screening is not an equivalent remedy

Table 3 gives the comparison. Applying no screening beyond the OECD's own, 11 of the 158 cells already reach 0.60 under our reconstruction; as reported under Validation, these are the cells the reconstruction fails to identify, not cells the rule treated wrongly. Removing students with MAD at or above 2.5 discards 5.9% of the sample and recovers 35 cells. At the threshold used by Steinmann et al. (2022), MAD at or above 2.0, screening discards 15.5% and recovers 87. To recover 122 cells it is necessary to discard 42.3% of respondents, a share no population survey can remove and still describe the population it sampled.

Table 3
Respondent screening compared with a same-valence half-scale

Rule applied	Respondents discarded	Suppressed cells recovered
None beyond the OECD's own exclusions	0.0%	11 of 158
Remove MAD ≥ 2.5	5.9%	35 of 158
Remove MAD ≥ 2.0	15.5%	87 of 158
Remove MAD ≥ 1.5	24.3%	118 of 158
Remove MAD ≥ 1.0	42.3%	122 of 158
Use the positively worded half-scale	0.0%	155 of 158

A cell counts as recovered when the administered-form alpha reaches 0.60 in every country-by-language group of that country, which is the rule the OECD applies. MAD is the mean absolute difference between the mean of the positively worded items and the mean of the recoded reverse-worded items.

The positively worded half recovers 155 cells while discarding no one.

This converges with what is already published. Bulut and Bulut (2022) found that removing careless respondents produced negligible improvement among the lowest-achieving students, who are precisely the population of the suppressed countries. The present result quantifies that in the OECD's own metric: the rule withholds the scale score of every respondent in 158 country-by-scale cells, while the alternative that works withholds none and uses half the items.

RQ3. The mechanism does not require careless respondents

Bifactor models fitted country by country, with a method factor on the reverse-worded items, show where the common variance goes (Table 4). In countries whose index was published, the median general-factor loading is 1.253 for the positively worded items and 0.665 for the reverse-worded ones. In suppressed

countries the positively worded items lose little, at 0.996, while the reverse-worded items fall to 0.232. The gap between suppressed and published countries in the general loading of the reverse-worded items holds in all seven scales and is large (Mann-Whitney $p < 10^{-40}$). Where the index was suppressed, the reverse-worded items have largely stopped measuring what the positively worded items measure.

Table 4
Median general-factor loading in bifactor models, by scale and by whether the index was published

Scale	Published: positive	Published: reverse	Suppressed: positive	Suppressed: reverse
Co-operation	1.23	0.52	1.59	0.33
Assertiveness	1.29	0.53	0.87	0.02
Stress resistance	0.90	0.77	0.85	0.22
Empathy	1.18	0.77	1.02	0.17
Perseverance	1.43	0.83	1.06	0.30
Emotional control	1.76	0.66	0.56	0.53
Curiosity	1.17	0.63	1.04	0.20
All scales	1.25	0.67	1.00	0.23

Bifactor models fitted country by country, with the method factor on the reverse-worded items. Entries are medians across countries of the median loading within each valence. The difference between suppressed and published countries in the general loading of the reverse-worded items is significant at $p < 10^{-40}$ (Mann-Whitney).

Whether that requires careless respondents is a separate question, and a simulation answers it. Mixed wording on its own does not reproduce the phenomenon: with the general loading of the reverse-worded items held at 1 and only an orthogonal method factor added, administered-form alpha moves from 0.78 to 0.75 even at large method loadings, and never approaches the threshold. What reproduces it is the condition the bifactor fits point to, the general loading of the reverse-worded items falling. Figure 2 shows that sweep. As γ falls from 1.0 to 0.0, the alpha of the administered mixed-valence form falls from **0.775 to 0.511**, crossing 0.60 between γ = 0.3 (0.624) and γ = 0.2 (0.590), while **the positively worded half stays at 0.795 throughout** and the reverse-worded half falls only from 0.876 to 0.794. The ten-item scale never falls below 0.707. The standard deviation across the 20 replicates never exceeds 0.009, so none of these separations is within simulation noise.

This is the empirical pattern, generated without a single careless respondent. Suppression does not require careless responding: it is enough that the reverse-worded items stop loading on the common factor. The careless-responding literature asks a different question about the same data, and the two accounts are compatible.

The prediction the simulation makes, and the data confirm

The same simulation makes a prediction that is not obvious in advance. Holding γ at 0.2 and varying how many items are reverse-worded, administered-form alpha traces an **inverted U**: the damage is greatest at a balanced split and smaller at both extremes, because a form dominated by one valence behaves like a single-valence form. Simulated alpha is 0.667 at two reverse-worded items, 0.625 at three, 0.599 at four, 0.590 at five, 0.600 at six and 0.628 at seven.

The seven scales differ in exactly this respect, and were designed that way for reasons unrelated to the present question. Figure 3 places the simulated curve against the observed percentages. Across the seven scales, the rank correlation between the distance from a balanced split and the percentage of countries suppressed is **− 0.805 (p = 0.029)**. The correlation is identical whether the percentage is computed on the alpha-suppressed cells of Table 19.A.14 or on all cells without a published index. Curiosity, with two reverse-worded items, loses 5.3% of its countries; emotional control, with seven, loses 16.4%; the two scales split five-five lose 44.0% and 40.8%.

RQ4. It is the reverse-worded half that degrades

The magnitudes of the criterion correlations decide nothing. They are small for both halves, with median absolute values between 0.02 and 0.24, and the differences between halves are small and inconsistent in sign. On correlation size alone, the two halves are indistinguishable.

What discriminates is whether the relation to the criterion weakens where extreme invariant responding is more common. Table 5 gives, for each scale, the across-country correlation between the official straightlining rate and each half's disattenuated validity coefficient for reading. **In all seven scales the difference between halves is negative**, from − 0.746 for emotional control to − 0.091 for assertiveness. For mathematics it is negative in five of seven, from − 0.420 to + 0.034.

Table 5
Straightlining rate and each half's validity coefficient for reading, across countries

Scale	Positively worded half	Reverse-worded half	Difference	Countries
Emotional control	0.240	−0.506	−0.746	73
Co-operation	0.640	0.131	−0.509	72
Perseverance	0.484	0.012	−0.472	72
Curiosity	0.236	−0.187	−0.423	76
Stress resistance	0.101	−0.272	−0.373	71
Empathy	0.530	0.181	−0.349	71
Assertiveness	0.252	0.160	−0.091	75

Spearman correlations, across the countries of each scale, between the official rate of extreme straightlining and each half's disattenuated criterion validity for reading. A negative difference means the reverse-worded half loses its relation to reading achievement, relative to the positively worded half, as the country's rate of extreme invariant responding rises.

The direction of that difference is what identifies the result, and it needs stating precisely. The reverse-worded half's validity coefficient falls as straightlining rises. The positively worded half's does not merely hold: it rises, in all seven scales, from + 0.101 to + 0.640. A common component moving both halves upward is expected here, because the correction for attenuation divides by a reliability that is itself lower where straightlining is higher, and inflates both coefficients in the same countries. That is precisely why the difference between halves, and not the level of either, is the quantity we read. An explanation resting on the general structure of variances in low-performing populations would move both halves together and cancel in the difference; it does not.

The pattern is not uniform across criteria (Table 6). It is clear for the two externally assessed achievement measures, negative in seven of seven scales for reading and five of seven for mathematics. It is weak for economic, social and cultural status, where the difference is negative in six of seven scales but with a median of − 0.027. And it is absent for the three single-item self-reports drawn from the same questionnaire: grade repetition, where the difference is positive in all seven scales, skipping school and educational expectation, each negative in two of seven. Those three criteria are themselves questionnaire responses, exposed to the same careless responding whose consequences we are trying to detect, so they are poor tests of the hypothesis. The claim we make is therefore restricted to the achievement criteria.

Table 6
Difference between halves in the association with the straightlining rate, by criterion

Criterion	Source	Scales with a negative difference	Median	Range
Reading performance	externally assessed	7 of 7	−0.423	−0.746 to −0.091
Mathematics performance	externally assessed	5 of 7	−0.249	−0.420 to + 0.034
Economic, social and cultural status	derived index	6 of 7	−0.027	−0.308 to + 0.003
Grade repetition	single questionnaire item	0 of 7	+ 0.324	+ 0.114 to + 1.011
Skipping school	single questionnaire item	2 of 7	+ 0.186	−0.083 to + 0.282
Educational expectation	single questionnaire item	2 of 7	+ 0.066	−0.340 to + 0.223

Each entry summarises the seven scale-level differences of Table 5 *computed for that criterion. A negative difference means the reverse-worded half loses its relation to the criterion, relative to the positively worded half, as the country's rate of extreme invariant responding rises.*

Three further comparisons point the same way, though they are not independent of one another, being functions of the same half-scale alphas. The positively worded half is the more reliable in five of seven scales; it reaches the threshold in 155 of 158 suppressed cells against 122; and it varies less across countries, with a mean standard deviation of 0.046 against 0.051.

Taken together, these results identify **the reverse-worded half as the contaminated one**. They do not establish that the positively worded half measures the construct cleanly.

Two further checks bear on how much of this is the country's reading level rather than its straightlining. The association between the decoupling of the two valences and the official straightlining rate is positive in all seven scales, from + 0.213 to + 0.841. Adjusting for the country's mean reading performance, it remains positive in all seven but reaches conventional significance in four: stress resistance (+ 0.427), emotional control (+ 0.570), assertiveness (+ 0.352) and co-operation (+ 0.240). This is consistent with the achievement gradient already reported by Steinmann et al. (2024) and Wang (2025), and we report it as a replication in a new battery rather than as an independent mechanism.

RQ5. The loss falls on low-performing countries

An analysis that uses all seven scales at once can work with **31 of 76 countries**.

The distribution is not arbitrary (Fig. 4). Across the 76 countries with a reading score, the rank correlation between the number of scales a country retains and its mean reading performance is **0.866** (p = 5.3 ×

10^{-24}). By quartile of reading performance, the mean number of scales retained is **1.47** in the lowest quartile, 4.00, 6.11 and **6.95** in the highest. The lowest-performing quarter of participating countries retains, on average, one and a half of the seven social and emotional skills indices; the highest-performing quarter retains essentially all of them.

For seven of these 76 countries part of the missingness reflects non-administration rather than suppression. Restricting to the 69 countries that administered all seven scales raises the correlation to 0.910, and using the proportion retained among those administered gives 0.888. The headline figure is therefore the conservative one.

Discussion

What this adds

Three components of this article are not new. The suppression rule is documented in plain language in the technical report, so implementing it is a reproducibility exercise rather than a discovery. That inconsistent responding is more common in lower-performing systems is established (Steinmann et al., 2024; Wang, 2025), and our replication of that gradient in a new battery is reported as a replication. The recommendation not to mix valences already exists (Steinmann, 2024); what this article adds to it is a measurement of what not following it cost in one specific reporting decision.

What is new is the counterfactual and what follows from it. The rule was applied 158 times, and in 155 of those the information required to meet its own criterion was present in the data, in its own unit of analysis, in a subset of items fixed in advance by wording. The rule did not encounter data incapable of supporting a reliable score. It encountered a form whose two halves behaved differently, and treated them as one.

The simulation and the external criterion each close an alternative reading. The simulation shows that careless responding is not necessary for the phenomenon to appear, and the screening comparison shows separately that removing careless respondents does not recover the scores; both converge with Bulut and Bulut's (2022) finding that screening helps least where it would be needed most. The external criterion shows that the two halves are not symmetric: it is the reverse-worded half whose criterion validity erodes as extreme invariant responding rises, in both achievement criteria and in every scale for reading.

The relation to Steinmann et al. (2022)

Steinmann et al. (2022) conclude that inconsistent responders biased estimated dimensionality and reliability, but that "the impact on external validity measures was limited and unsystematic". We report a systematic pattern in criterion validity. The two findings are not in conflict, because the questions differ in two ways.

First, they compare the external validity of the **complete** scale before and after **removing respondents**; we compare the validity of the **two halves** of the scale without removing anyone. Second, they examine the **level** of validity; we examine whether that level is **moderated by the country**. The second is a weaker signal in absolute terms but a cleaner one for this purpose, because the contrast between halves cancels whatever affects both halves equally. A design that removes respondents and then looks at the whole scale cannot see a difference that lives in the comparison between valences.

Implications

For the producers of the assessment, the finding is narrow and concrete. An internal consistency criterion applied to a mixed-worded form under matrix sampling is not a test of whether a country's students answered usefully; it is partly a test of how balanced the valence split of that scale is, as the inverted U shows. A reporting rule can be evaluated against what a same-valence subscale would have delivered before a score is withheld, and the information needed to do that is already in the calibration.

For secondary users the implication is different and more immediate. The missingness in these seven indices is strongly related to country performance, which means that any cross-national analysis using them is conducted on a sample of countries selected on a variable correlated with most outcomes of interest. The substantive articles that use this battery work on that sample. Reporting the retained-country set, and treating the missingness as informative, costs little and changes what the results can support.

Limitations

We do not establish that the positively worded half measures the construct cleanly. The evidence identifies the reverse-worded half as the contaminated one, through the interaction with extreme invariant responding. It does not follow that the remaining half is free of wording effects, acquiescence or any other threat, and it is well established that positively worded items carry their own susceptibilities. The counterfactual is a statement about internal consistency, which is what the OECD rule tests; it is not a validity claim. This is the principal limitation and it constrains what the article recommends: not that single-valence scales be adopted, but that reporting decisions be evaluated against what a single-valence subscale would deliver.

Our reconstruction of the official alpha is approximate. It correlates 0.936 with Table 19.A.11 across 539 groups but carries a systematic bias of − 0.025, and for curiosity the correlation is 0.642. This is why published alphas are used wherever an official value is needed.

The counterfactual estimator is optimistic and the correction is a median. The pairwise estimator exceeds the official value by 0.017 at the median and 0.048 at the 90th percentile. Correcting by the within-scale median leaves individual groups over- or under-corrected. The confidence-bound column in Table 2 is the appropriate reference for how far the result degrades under a conservative reading: from 155 to 150 of 158.

Coverage of language groups is not complete. We reconstruct 92 distinct country-by-language groups where the technical report lists 103.

Two scales do not fully converge. Curiosity and empathy had not converged at 60 EM iterations; the effect on the reported rank correlations was measured at approximately 0.03.

The half-to-half contrast is correlational across countries. It rests on 71 to 76 countries per scale and cannot separate straightlining from whatever else covaries with it at country level, beyond the adjustment for mean reading reported under RQ4.

The contrast holds only for the externally assessed criteria. Of the six criteria examined, the pattern is clear for reading and mathematics, weak for economic, social and cultural status, and absent for the three single-item questionnaire self-reports. We give a reason for expecting that — those three are themselves exposed to the responding behaviour under study — but the reason was not tested, and a reader is entitled to treat the restriction as a limit on the generality of the result.

Achievement enters as the mean of the ten plausible values. The recommended procedure is to combine the ten by Rubin's rules. The sensitivity analysis reported in the Methods shows the difference to be negligible for these quantities, but the reported standard errors of the underlying within-country correlations are not the ones Rubin's rules would give.

Conclusions

PISA 2022 withheld 158 country-by-scale social and emotional skills indices because a mixed-worded form did not reach an internal consistency threshold. In 155 of those 158 cells the positively worded items alone would have reached that threshold, in every country-by-language group, under the OECD's own criterion and unit of analysis, and using an item subset fixed in advance by wording rather than chosen after the fact. Screening out inconsistent respondents does not achieve the same result without discarding a share of the sample that a population survey cannot remove. The phenomenon does not require careless respondents: it follows from the reverse-worded items ceasing to load on the common factor, and the balance of valences predicts which scales lose the most countries.

The consequence is unequally distributed. The lowest-performing quarter of participating countries retains 1.47 of the seven indices on average and the highest-performing quarter retains 6.95, and only 31 of 76 countries retain all seven. A rule applied to protect internal consistency removes information disproportionately from the lowest-performing systems.

Two conclusions follow. Reporting decisions for mixed-worded scales can and should be evaluated against what a same-valence subscale would deliver before a score is withheld. And users of these indices should treat the resulting missingness as informative, as a variable strongly related to country performance, rather than as an incidental gap in coverage.

Abbreviations

AUC: area under the receiver operating characteristic curve · EAP: expected a posteriori · EM: expectation-maximisation · ESCS: index of economic, social and cultural status · GPCM: generalised partial credit model · GRM: graded response model · IEA: International Association for the Evaluation of Educational Achievement · ILSA: international large-scale assessment · MAD: mean absolute difference · MML: marginal maximum likelihood · OECD: Organisation for Economic Co-operation and Development · PISA: Programme for International Student Assessment · RMSD: root mean square deviation · WLE: weighted likelihood estimate

Declarations

Competing interests.

The authors declare that they have no competing interests.

Funding.

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

P.V.P. conceived the study, implemented the item response theory estimator and the analysis pipeline, conducted all analyses, produced the figures and tables, and drafted the manuscript. N.S.M.V. contributed to the design of the study and to the interpretation of the results, and critically revised the manuscript. J.C.B.S. contributed to the interpretation of the results and to the review of the literature, and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements.

The authors thank the OECD and the PISA 2022 consortium for releasing the student questionnaire microdata and the technical documentation of the scaling procedures in a form that makes independent verification possible. This article is a re-analysis of those public materials; the OECD bears no responsibility for the analyses or interpretations presented here.

Data Availability

This study is a secondary analysis of publicly available data, and it also generated derived data.Data analysed. The microdata are the PISA 2022 student questionnaire public-use file (CY08MSP_STU_QQQ.SAV), published by the OECD and freely available without registration or permission at https://www.oecd.org/en/data/datasets/pisa-2022-database.html. The scaling documentation and the Chapter 19 annex tables used for validation are published with the PISA 2022 Technical Report, https://doi.org/10.1787/01820d6d-en. The microdata are © OECD and are redistributed here only in derived, aggregated form.Data generated. The estimator, the complete analysis pipeline and every derived table cited in the article are openly available in the pisa-suppressed-scales repository, https://github.com/pedrovelezpardo/pisa-suppressed-scales. The derived tables include one row per suppressed country-by-scale cell with its counterfactual half-scale alphas and confidence intervals, the 621 country-by-language group reliabilities, the external-criterion correlations, the simulation grid and the respondent-screening comparison. Code is released under the MIT licence and the derived tables under CC BY 4.0. Programming language Python 3, with R (mirt) used for independent validation; platform independent; no restrictions on use by non-academics.

References

1. Antoniou, F., & Alghamdi, M.H. (2024). Confidence in mathematics is confounded by responses to reverse-coded items. Frontiers in Psychology, 15, 1489054. doi:10.3389/fpsyg.2024.1489054.
2. Bulut, H.C., & Bulut, O. (2022). Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? Studies in Educational Evaluation, 72, 101126. doi:10.1016/j.stueduc.2022.101126.
3. Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29. doi:10.18637/jss.v048.i06.
4. Chen, J., Steinmann, I., & Braeken, J. (2024). Competing explanations for inconsistent responding to a mixed-worded self-esteem scale: Cognitive abilities or personality? Personality and Individual Differences, 222, 112573. doi:10.1016/j.paid.2024.112573.
5. Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. doi:10.1007/BF02310555.
6. Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. Psychometrika, 30(3), 357–370. doi:10.1007/BF02289499.
7. Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. Assessment, 27(2), 404–418. doi:10.1177/1073191117746503.
8. Michaelides, M.P., & Konstantinidou, E. (2026). An experimental study on the impact of survey stakes on response inconsistency in mixed-worded scales. Educational and Psychological Measurement, 86(2), 266–286. doi:10.1177/00131644251395323.
9. Michaelides, M.P., Konstantinidou, E., Ivanova, M., & Steinmann, I. (2026). Inconsistent responding on a mixed-worded scale in the PISA 2022 questionnaire: Prevalence and predictors across seven

countries. Journal of Psychoeducational Assessment. Advance online publication. doi:10.1177/07342829261440806.

10. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159–176. doi:10.1177/014662169201600206.
11. OECD (2024). PISA 2022 Technical Report. Paris: OECD Publishing. doi:10.1787/01820d6d-en.
12. OECD (2024b). PISA 2022 Database. Organisation for Economic Co-operation and Development. https://www.oecd.org/en/data/datasets/pisa-2022-database.html. Accessed 7 August 2026.
13. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4, Pt. 2), 100.
14. Steinmann, I. (2024). To mix or not to mix positively and negatively worded items. Amsterdam: International Association for the Evaluation of Educational Achievement. https://www.iea.nl/publications/rd-outcomes/mix-or-not-mix-positively-and-negatively-worded-items. Accessed 7 August 2026.
15. Steinmann, I., Chen, J., & Braeken, J. (2024). Who responds inconsistently to mixed-worded scales? Differences by achievement, age group, and gender. Assessment in Education: Principles, Policy & Practice, 31(1), 5–31. doi:10.1080/0969594X.2024.2318554.
16. Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2022). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. Assessment in Education: Principles, Policy & Practice, 29(1), 5–26. doi:10.1080/0969594X.2021.2005302.
17. Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7(2), 201–210. doi:10.1177/014662168300700208.
18. Wang, M.D. (2025). Data quality disparities in large-scale assessments: Insufficient effort responding across student groups, schools, and cultures. Large-scale Assessments in Education, 13(1), 21. doi:10.1186/s40536-025-00260-z.
19. Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427–450. doi:10.1007/BF02294627.

Figures

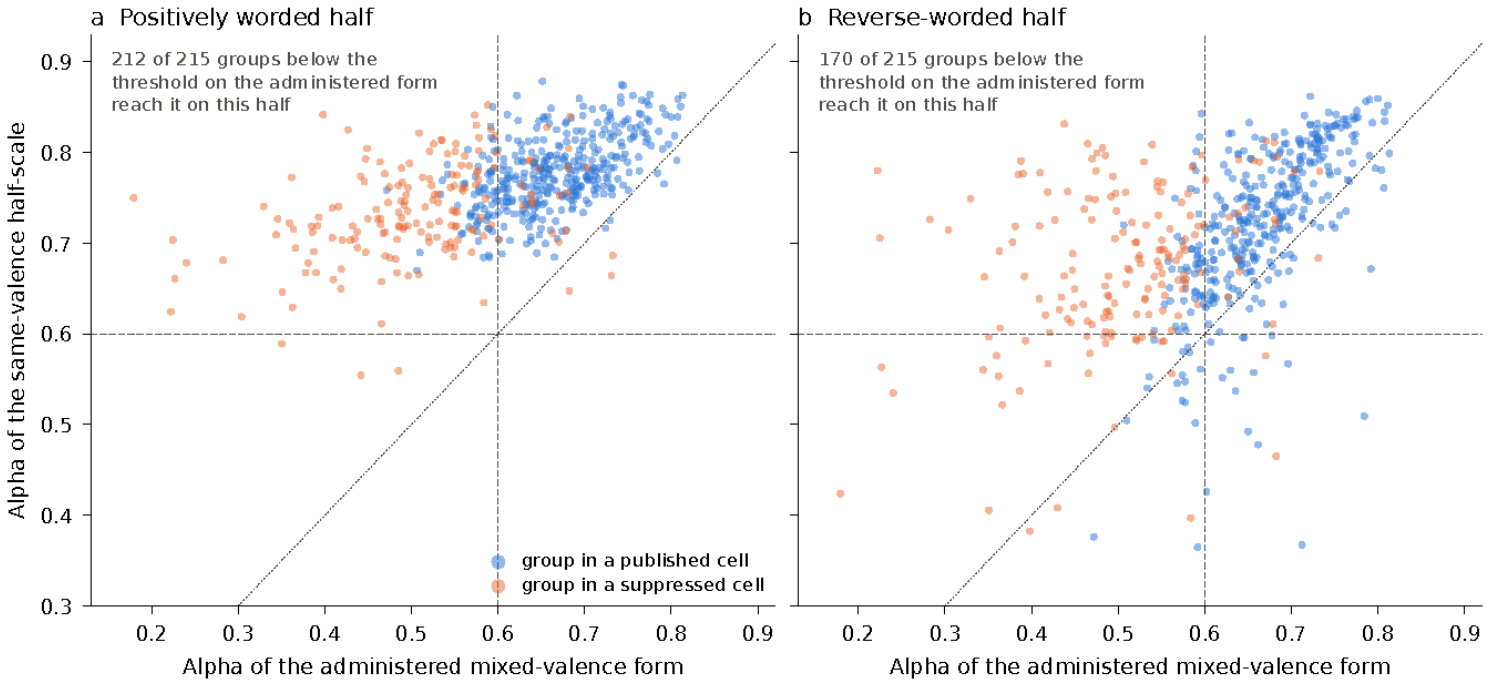


Figure 1

Administered-form alpha against same-valence half-scale alpha, by country-by-language group. Each point is one group-by-scale unit; the reference lines mark the 0.60 threshold. Groups belonging to suppressed cells are distinguished from the rest.

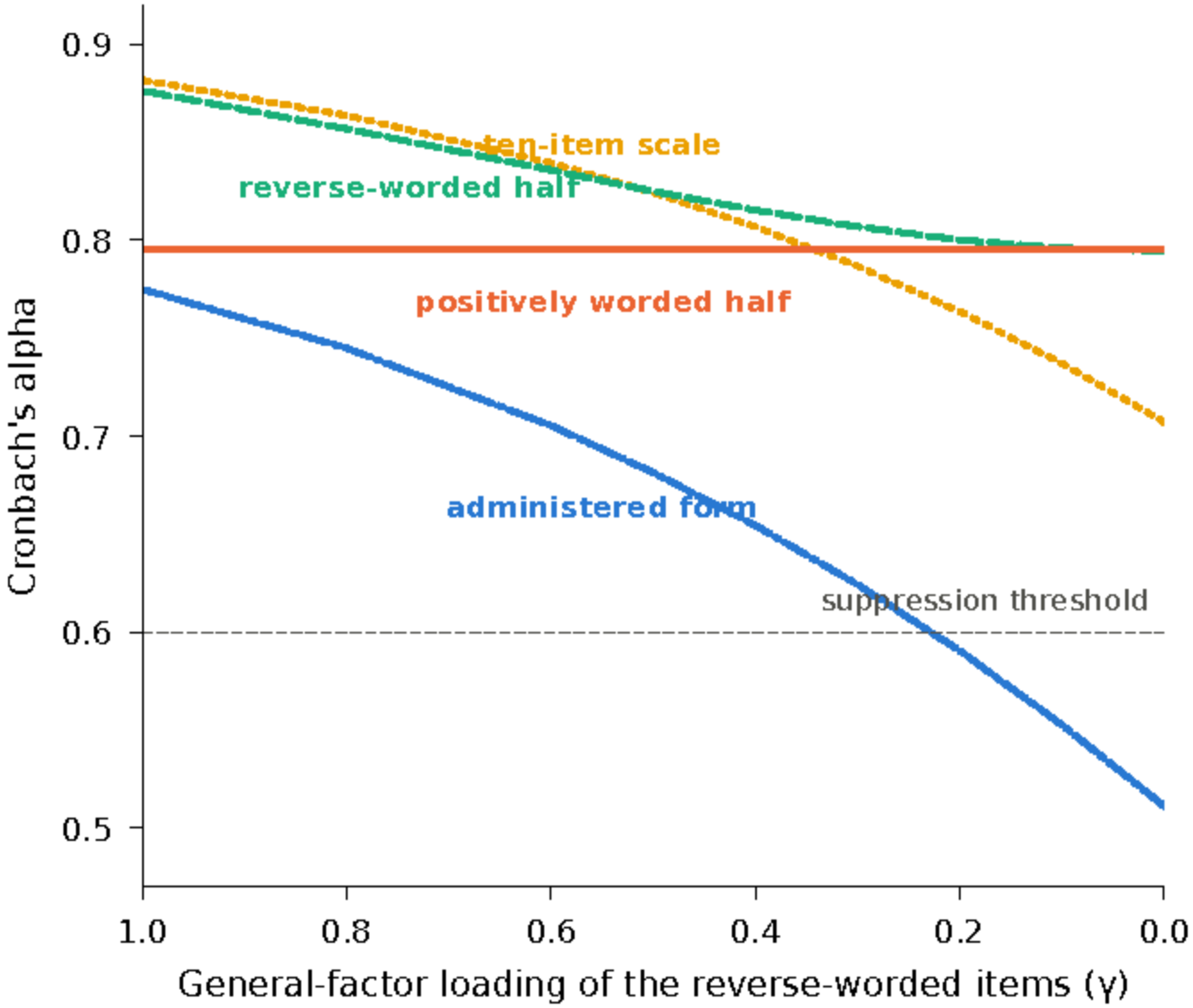


Figure 2

Simulated alpha as the general-factor loading of the reverse-worded items declines, with no careless respondents. Administered mixed-valence form, each same-valence half, and the complete ten-item scale. Means of 20 replicates.

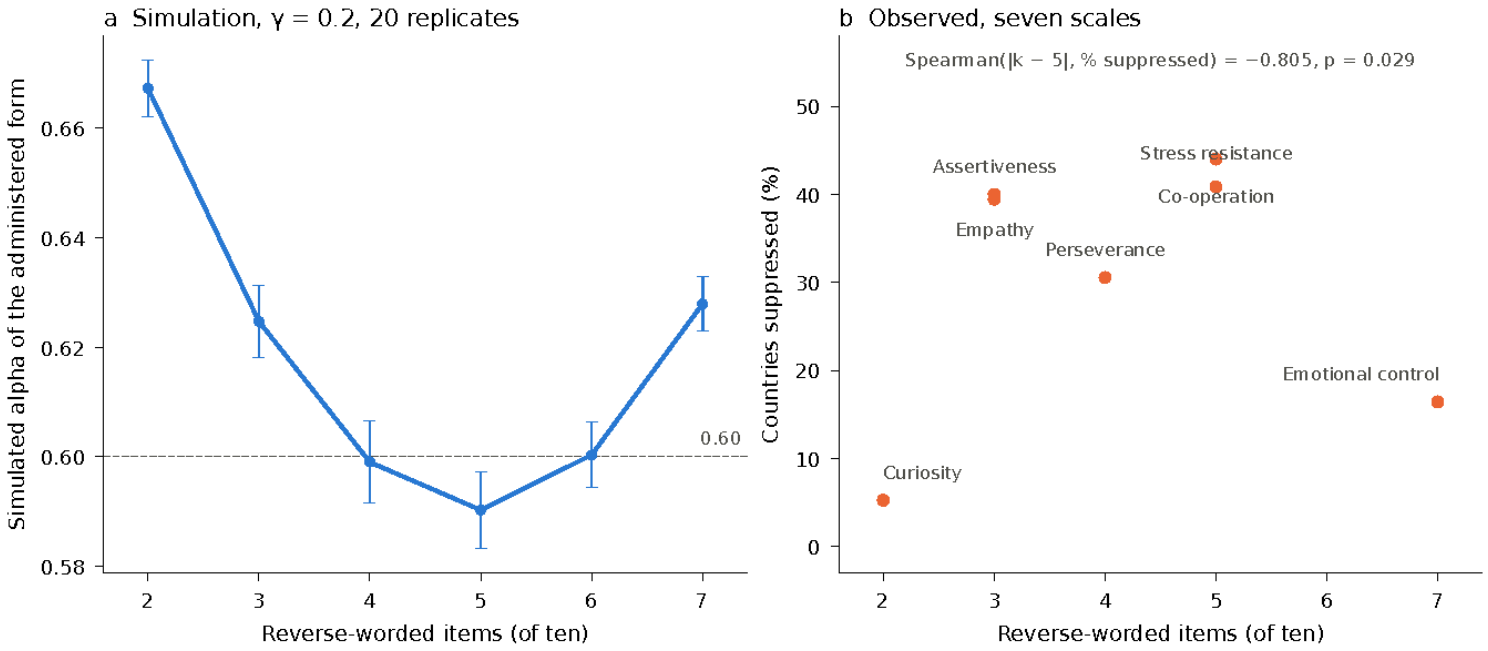


Figure 3

Simulated and observed effect of the balance of valences. Simulated administered-form alpha by number of reverse-worded items at a fixed general loading, against the observed percentage of countries suppressed in each of the seven scales.

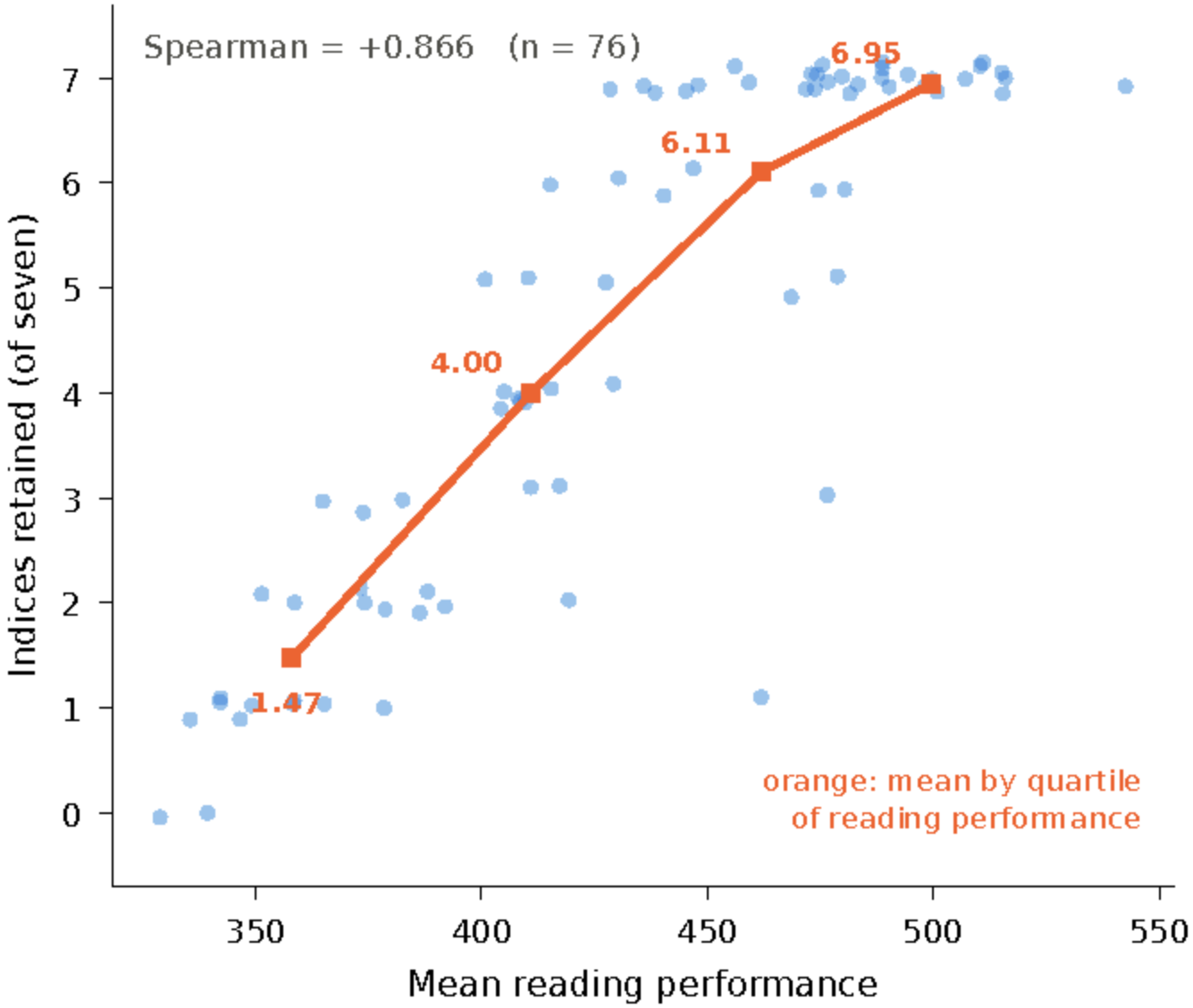


Figure 4

Number of social and emotional skills indices retained against mean reading performance, by country.