PISA 2012 as a Stress Test of a Difficulty- and Positive-Somers-Weighted Binary Scoring System

Mathematical specification, computational validation, bootstrap uncertainty, and cross-booklet replication

Anders Hellström

28 August 2026

AI use disclosure

OpenAI ChatGPT (GPT-5.6 Sol) was used to assist with mathematical formalization, statistical analysis, literature and web research, R/Python/LaTeX development, figure preparation, and drafting and editing of this report. The underlying scoring-system concepts and key methodological decisions were provided by the human researcher. AI-generated material was reviewed and revised as part of the research process; responsibility for the final content, interpretations, and conclusions remains with the human author.

New Scoring System project

Full research and technical-assistance disclosure is provided in Chapter 16.

This report documents and evaluates a sample-calibrated binary test-scoring method that combines empirical item difficulty with positive ordinal discrimination. For each dichotomous item, difficulty is represented by the number answering correctly, while discrimination is estimated using Harrell's Somers D_{xy} between the item response and an item-rest score. The empirical concordance probability C is retained for diagnostics, but the weighting stage uses the project-specific transformation $A = \max(0, D_{xy}) = \max(0, 2C - 1)$, so non-positive discrimination receives neither a reward nor a penalty. Two add-one padded logarithmic factors are multiplied to obtain a raw item weight. All retained item weights are then divided by the minimum raw weight, a purely multiplicative normalization that sets the smallest operational weight to one without changing weight ratios or the normalized examinee score.

PISA 2012 United States mathematics item responses are used as a real-data stress test rather than as a replacement for official PISA scaling. The public-derived `edmdata` matrix contains 4,978 students and 76 dichotomized items with a rotated administration design. A clean 435-student, 32-item administration-pattern group was analysed first. In that group the operational weights ranged from 1.000 to 8.621; number correct and weighted raw score correlated at $r = 0.9872$, but the weighted score produced 434 distinct values versus 31 distinct number-correct scores. The extreme item `pm995q02` combined 1.84% correct with $D_{xy} = 0.891$, receiving operational weight 8.621. A 50,000-replicate bootstrap produced a 95% percentile interval of 0.813–0.955 for D_{xy} and 12.87–21.39 for its raw weight; whole-form resampling placed it first in operational weight in 99.974% of replicates.

The rotated PISA design then allowed independent recalibration across 13 major response-pattern groups containing 4,948 students. All 76 items occurred in four major groups. Across items, the mean cross-group coefficient of variation (CV) of raw weight was 10.52% and the median 9.83%; six items exceeded 20%. `pm995q02` was comparatively stable (CV 12.43%) despite its large mean weight. The largest instability was observed for `pm943q02` (CV 27.77%), whose raw weight ranged from 9.45 to 18.90. A common-rest sensitivity analysis did not remove the instability, indicating that sample composition, small success counts, variation in D_{xy}, booklet/position effects, and the nonlinear sensitivity of the concordance factor near $D_{xy} = 1$ are all relevant.

The evidence establishes computational consistency and provides encouraging evidence that some extreme weights can replicate across PISA booklet groups. It does not establish external validity, fairness, reliability superiority, or suitability for high-stakes assessment. The strongest next steps are independent/frozen calibration, uncertainty propagation, rare-item regularisation studies, survey-design-aware PISA analyses, differential-item-functioning analyses, and prospective comparison with simpler classical and IRT scoring rules.

Contents

List of Figures

List of Tables

1 Purpose, scope, and evidential status

1.1 Research objective

The purpose of this report is twofold. First, it gives a complete mathematical and computational description of the current scoring system used in the New Scoring System project. Second, it uses PISA 2012 mathematics responses as an unusually informative external stress test because PISA combines a large real examinee sample, a broad item-difficulty range, and a rotated booklet design in which many items recur in independently assigned forms. This makes it possible to examine not only in-sample behaviour but also approximate replication of item weights across distinct respondent groups.

The report deliberately separates three questions that are often conflated:

1. **Implementation correctness:** does the software implement the stated equations?
2. **Empirical stability:** do estimated weights remain similar under resampling and across administration groups?
3. **Psychometric validity:** do the resulting person scores improve reliability, criterion validity, construct representation, fairness, and decision quality relative to appropriate alternatives?

The project has substantial evidence for the first question and new PISA evidence for the second. The third remains open.

1.2 What this report is not

This is not an official OECD/PISA analysis, and the proposed scoring system is not presented as a substitute for the PISA proficiency scales. PISA is designed primarily for population- and system-level inference, uses a complex survey design, and supplies plausible values that must be analysed with the appropriate sampling and imputation variance procedures (OECD 2014; OECD 2026b). The present analysis instead treats the U.S. mathematics response matrix as a psychometric test bed for a different scoring rule. The analyses reported here are unweighted within reconstructed administration-pattern groups; they are therefore not estimates of U.S. population proficiency.

Interpretive boundary

A high correlation with number correct, a stable item weight, or successful reproduction of the intended formula does *not* establish that the scoring rule measures mathematical literacy better than PISA's official IRT-based methodology. The PISA results in this report are evidence about the behaviour of the proposed weighting rule under real response patterns.

1.3 Current versus historical project versions

The project evolved through several earlier definitions, including a prevalence-matched rank-cut statistic and a temporary Spearman-based implementation. Those versions are historically relevant to the development process but are **superseded**. The authoritative method in this report is the current item-rest Somers D_{xy} version with positive-discrimination remapping, two different padded logarithmic factors, retained constant items, and minimum-to-one multiplicative normalization. Historical numerical conclusions from earlier rank-cut reports should not be attributed to the current method.

2 PISA 2012: assessment design and why it is useful here

2.1 PISA 2012 mathematics

PISA 2012 placed mathematical literacy at the centre of the assessment cycle. The OECD framework defines the construct in terms of students' capacity to formulate, employ, and interpret mathematics in a variety of contexts rather than as a narrow curriculum test (OECD 2013b). Approximately 510,000 students from 65 countries/economies participated in PISA 2012 overall (OECD 2014). The present project uses a U.S. mathematics item-response matrix distributed through the `edmdata` R package and traceable to the public PISA 2012 data resources (OECD 2026a; TMSA Lab 2026).

2.2 Rotated booklet structure

The PISA paper-based design did not administer every mathematics item to every student. Standard booklets combined half-hour item clusters, and the design rotated clusters across booklet positions so that a broad domain could be covered without requiring a single student to complete the entire pool. In the core PISA 2012 design, a cluster appeared in four booklets and was rotated through the four positions; the 13 standard booklets formed a balanced incomplete-block structure (OECD 2014).

This feature is methodologically valuable for the present project. If an item appears in four distinct booklet groups, then the same item can be recalibrated on four independently assigned student groups. That is more informative than a single random split because it reflects actual administration contexts and produces a natural cross-form replication study. It also introduces potential complications: position, fatigue/endurance, companion-item composition, and booklet difficulty can affect responses. PISA itself modelled booklet effects in calibration precisely to reduce confounding between item and booklet difficulty (OECD 2014).

2.3 Official PISA scaling is fundamentally different

PISA's official scaling is an IRT-based latent-variable system, not a weighted sum of observed item credits. The 2012 methodology used Rasch-family scaling for dichotomous responses and partial-credit modelling for polytomous responses, with international calibration and booklet effects in the relevant scaling models (OECD 2014; Rasch 1960; Warm 1989). Population analyses are based on multiple plausible values rather than treating a single student-level score as error-free (OECD 2026b).

The proposed project method answers a different question. It asks how to construct a sample-calibrated weighted sum in which a correct response is worth more when the item is both difficult and positively aligned with performance on the remaining items. This makes a PISA comparison informative as a stress test, but it is not a like-for-like contest between two estimators of the same latent model.

2.4 Derived U.S. response matrix and missingness coding

The `edmdata::items_pisa12_us_math` object contains 4,978 rows and 76 item columns. The package documentation maps PISA score code 7 (N/A) to `NA` and score code 8 (not attempted) to zero (TMSA Lab 2026). Consequently, zero in the binary matrix can represent either an observed no-credit response or a presented-but-not-attempted response. Structural N/A cells remain missing. This distinction must be remembered when interpreting "difficulty": the project weights operate on the binary scored representation, not a pure attempted-correct versus attempted-incorrect matrix.

The project reconstructed exact `NA`/non-`NA` masks and found 18 patterns. Thirteen large patterns contained 4,948 of the 4,978 students (99.40%) and aligned closely with the 13 standard-booklet structure. The remaining 30 rows had rare/anomalous patterns, including a 26-student all-missing

pattern and several singletons. For direct rectangular-matrix scoring, the largest common pattern contained 435 students and 32 administered items with no structural missingness.

Table 2.1: The 13 major PISA response-pattern groups reconstructed in the project.

Pattern	Students	Items	Pattern	Students	Items
01	435	32	08	414	11
02	433	30	09	394	22
03	429	34	10	297	9
04	421	22	11	295	32
05	419	11	12	289	32
06	418	24	13	287	11
07	417	34			

3 Complete specification of the current scoring system

3.1 Response matrix and item-rest criterion

Let $X = (x_{ij})$ denote a binary person-by-item response matrix with n test takers and J items:

$$x_{ij} \in \{0, 1\},$$

where 1 indicates credit/correct and 0 indicates no credit/incorrect in the input representation. For item j, define

$$c_j = \sum_{i=1}^{n} x_{ij}, \qquad n_{1j} = c_j, \qquad n_{0j} = n - c_j.$$

The default corresponding score for analysing item j is the *item-rest* score

$$T_{i,-j} = \sum_{k \neq j} x_{ik}. \tag{3.1}$$

A full total-score mode remains available as a compatibility option, but item-rest is preferred because including item j in its own criterion produces part-whole contamination. The long-standing item-analysis literature recognises that an item-total correlation is mechanically inflated when the item is contained in the total (Henrysson 1963). Item-rest analysis removes that direct overlap, although the criterion remains sample- and test-dependent.

3.2 Somers D_{xy} and concordance probability

For each nonconstant item, the program computes Harrell's Somers D_{xy} using the item-rest score as ordered predictor and the binary item response as outcome. The `Hmisc::somers2()` convention returns the concordance probability C_j and

$$D_{xy,j} = 2C_j - 1. \tag{3.2}$$

This statistic is particularly well matched to a binary item versus an ordered rest score: it compares the rest-score distribution of students receiving credit with the distribution of students receiving no credit. Somers' asymmetric ordinal association originates in Somers (1962); its use as an item-rest discrimination coefficient has been explicitly studied in educational measurement (Metsämuuronen 2020). The project uses the `Hmisc` implementation as the authoritative computational reference (Harrell et al. 2026).

3.2.1 Pair-count interpretation

Consider only cross-group pairs containing one correct respondent and one incorrect respondent. Let

- P_j be the number for which the correct respondent has the higher rest score;
- Q_j be the number for which the correct respondent has the lower rest score;
- T_j be the number tied on the rest score.

There are $n_{1j}n_{0j}$ cross-group pairs, hence

$$P_j + Q_j + T_j = n_{1j}n_{0j}. \tag{3.3}$$

The concordance probability gives half credit to cross-group score ties:

$$C_j = \frac{P_j + \frac{1}{2}T_j}{n_{1j}n_{0j}}. \tag{3.4}$$

Therefore

$$D_{xy,j} = \frac{P_j - Q_j}{n_{1j} n_{0j}}. \tag{3.5}$$

The same calculation is linked to the Mann–Whitney statistic

$$U_j = P_j + \tfrac{1}{2} T_j = C_j n_{0j} n_{1j}, \tag{3.6}$$

which provides an independent diagnostic identity (Mann and Whitney 1947). The program can report P, Q, T, U, n_0, n_1 in research-diagnostic mode and uses these identities to cross-check the Hmisc result.

3.3 Positive-discrimination remapping

The current project does *not* feed the raw concordance probability C_j directly into the weight. Instead it applies a piecewise-linear transformation:

$$A_j = \max(0, 2C_j - 1) = \max(0, D_{xy,j}). \tag{3.7}$$

Thus

$D_{xy,j}$	-1	-0.5	0	0.5	1
A_j	0	0	0	0.5	1

Negative or zero discrimination receives no discrimination reward and no discrimination penalty. Positive Somers discrimination is preserved linearly on $[0, 1]$. This truncation is a **project-specific scoring decision**; it is not part of the definition of Somers D itself.

3.4 The two padded factors

The method combines two dimensions: rarity/difficulty and positive discrimination. Their transformations deliberately use different complements.

3.4.1 Difficulty pair and first pad

Define the rarity pair

$$R_j = (c_j, n).$$

The first add-one pad is

$$p_{1j} = \frac{c_j + 1}{n + 1}. \tag{3.8}$$

The difficulty factor is

$$F_{1j} = 1 - \ln p_{1j} = 1 + \ln\left(\frac{n+1}{c_j + 1}\right). \tag{3.9}$$

Holding everything else fixed, F_{1j} decreases as more people answer correctly:

$$\frac{\partial F_{1j}}{\partial c_j} = -\frac{1}{c_j + 1} < 0. \tag{3.10}$$

Therefore a rarer item receives a larger difficulty factor.

3.4.2 Positive-discrimination pair and second pad

Define the concordance/discrimination pair

$$G_j = (nA_j, n). \tag{3.11}$$

The numerator nA_j is intentionally retained as a real number and is *not rounded*. The second pad uses the complement:

$$p_{2j} = \frac{n - nA_j + 1}{n+1} = \frac{n(1-A_j)+1}{n+1}. \tag{3.12}$$

The positive-discrimination factor is

$$F_{2j} = 1 - \ln p_{2j} = 1 + \ln\left(\frac{n+1}{n(1-A_j)+1}\right). \tag{3.13}$$

It is strictly increasing in A_j:

$$\frac{\partial F_{2j}}{\partial A_j} = \frac{n}{n(1-A_j)+1} > 0. \tag{3.14}$$

At $A_j = 0$, $F_{2j} = 1$; at perfect positive discrimination $A_j = 1$,

$$F_{2j} = 1 + \ln(n+1).$$

3.5 Raw item weight

The core raw weight is the product

$$\boxed{w_j^{\text{raw}} = \left[1 + \ln\left(\frac{n+1}{c_j+1}\right)\right]\left[1 + \ln\left(\frac{n+1}{n(1-A_j)+1}\right)\right].} \tag{3.15}$$

This product means that an item can receive a large weight only when one or both factors are large, and especially when difficulty and positive discrimination coincide. It is not an additive index: the multiplication creates an interaction in which the impact of increasing discrimination is larger when the difficulty factor is already large, and vice versa.

Because $0 \le c_j \le n$ and $0 \le A_j \le 1$, each factor lies in

$$1 \le F_{1j}, F_{2j} \le 1 + \ln(n+1),$$

so

$$1 \le w_j^{\text{raw}} \le [1 + \ln(n+1)]^2. \tag{3.16}$$

The upper bound is a formal endpoint; whether it is empirically attainable depends on the response configuration.

3.6 Minimum-to-one operational normalization

After all raw weights have been estimated within the calibration sample, define

$$w_{\min}^{\text{raw}} = \min_{1 \le k \le J} w_k^{\text{raw}}.$$

The operational item weight is

$$\boxed{W_j = \frac{w_j^{\text{raw}}}{w_{\min}^{\text{raw}}}.} \tag{3.17}$$

This guarantees that the smallest retained operational weight equals 1. It is a pure multiplicative rescaling and therefore preserves every raw-weight ratio:

$$\frac{W_j}{W_k} = \frac{w_j^{\text{raw}}}{w_k^{\text{raw}}}.$$

Unlike affine min–max scaling, it does not alter the normalized person score.

3.7 Person scoring

The weighted raw score of person i is

$$Y_i = \sum_{j=1}^{J} x_{ij} W_j, \tag{3.18}$$

and the maximum possible total is

$$W_{\max} = \sum_{j=1}^{J} W_j. \tag{3.19}$$

The final normalized score is

$$\boxed{S_i = \frac{Y_i}{W_{\max}}.} \tag{3.20}$$

All operational weights are positive, so $0 \leq S_i \leq 1$. If all item weights are multiplied by the same positive constant, both numerator and denominator change by that constant and S_i is unchanged. This is why the minimum-to-one step is interpretive rather than substantive for normalized person scores.

3.8 Constant items

Somers D_{xy} is undefined when an item is all zero or all one because one response group is absent. The current project policy is to *retain* such items while distinguishing empirical estimation from the operational rule:

$$D_{xy,j}^{\text{emp}} = \text{NA}, \qquad C_j^{\text{emp}} = \text{NA},$$
$$C_j^{\text{weight base}} = 0.5, \qquad D_{xy,j}^{\text{weight base}} = 0,$$
$$A_j = 0.$$

Thus a constant item receives $F_2 = 1$ and is weighted by difficulty alone.

This rule has nontrivial consequences. For an all-one item, $c_j = n$, so $F_1 = 1$ and $w_j^{\text{raw}} = 1$. For an all-zero item, $c_j = 0$, so $w_j^{\text{raw}} = 1 + \ln(n+1)$. The all-one item contributes the same positive amount to every examinee; the all-zero item contributes nothing to any numerator but can contribute a large term to the denominator. Under fixed constants these effects do not reorder students relative to the variable-item score, but they shift/compress the reported scale. This is an explicit design choice and should be reconsidered if the system is used operationally.

3.9 Asymptotic and sensitivity properties

For fixed nonzero proportion correct $p = c/n$ and fixed $A < 1$, as n grows,

$$F_1 \to 1 - \ln p, \qquad F_2 \to 1 - \ln(1 - A).$$

But two extreme cases grow with sample size. If c remains small while n grows, the difficulty factor grows approximately like $\ln n$. If $A = 1$, the discrimination factor is $1 + \ln(n+1)$. Thus extremely rare or perfectly separating items can receive larger raw weights in larger calibration samples unless their proportions/discrimination move away from the boundary.

Equation (3.14) also reveals a local sensitivity issue. For $n = 435$, the derivative is about 2 at $A = 0.5$, about 4.9 at $A = 0.8$, 9.8 at $A = 0.9$, and 19.1 at $A = 0.95$. Hence small sampling variation in a very high D_{xy} estimate can produce a much larger change in the discrimination factor.

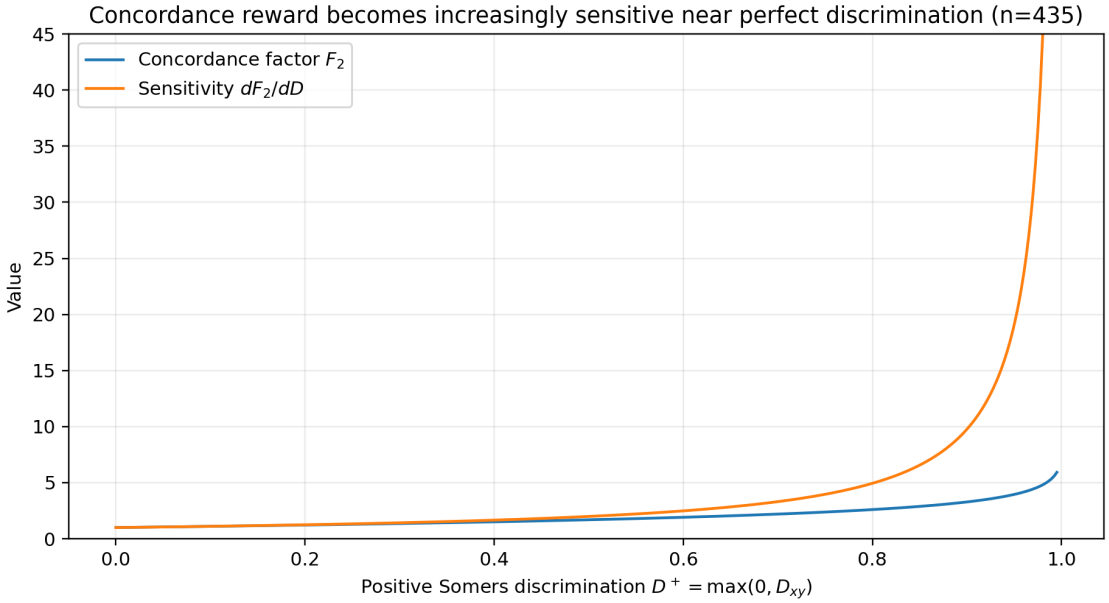

Figure 3.1: Concordance factor and its derivative for $n = 435$. The project-specific logarithmic reward becomes increasingly sensitive as positive Somers discrimination approaches one.

4 Relationship to established methods

4.1 Classical item difficulty

The basic empirical proportion correct

$$p_j = \frac{c_j}{n}$$

is the classical facility/easiness statistic. The project does not use p_j linearly; it applies an add-one-smoothed negative-log transform plus one. This gives the difficulty component an information-like shape: a change from 1% to 2% correct matters more multiplicatively than the same one-percentage-point change near 80%. The specific $(c+1)/(n+1)$ smoothing is a project design choice rather than a standard Bayesian posterior mean (which would ordinarily involve an $n+2$ denominator under a uniform Beta prior).

4.2 Item-rest discrimination

Using $T_{i,-j}$ rather than the full score follows the broad item-rest principle: the analysed item should not mechanically appear in the criterion against which it is evaluated. The choice does not make the criterion independent of the item-generating process; it remains the sum of other items in the same test and therefore assumes that those items supply a meaningful ordering of the target construct.

4.3 Somers D, AUC/concordance, and Mann–Whitney

The discrimination component has established statistical foundations. With a binary response and an ordered score, C is the probability that a randomly chosen positive/correct case outranks a randomly chosen negative/incorrect case, with half credit for ties. The Mann–Whitney U identity in Equation (3.4) and Somers relationship in Equation (3.2) show that the method is built on a rank-based cross-group ordering statistic rather than a conventional Pearson correlation. Somers D can attain ± 1 under complete separation/reversal even with a binary item, which is one reason it has been proposed for item discrimination (Metsämuuronen 2020).

The novel/project-specific aspects begin after D_{xy} is estimated: truncating negative values to zero, using nA as a pair numerator, applying the complementary add-one pad, multiplying the shifted logarithmic factors, retaining constants with an operational neutral base, and applying minimum-to-one normalization. A prior project literature search found close precedents for the components but did not identify the exact full pipeline. That observation should be treated as a targeted-search result, not a patent-grade or exhaustive novelty determination.

4.4 Comparison with IRT conceptually

Both the proposed method and IRT distinguish, in some sense, item difficulty from the relation between item response and proficiency. They do so in very different ways. In Rasch-family models, item difficulty is a latent-location parameter and the response probability is modelled conditionally on latent proficiency. In the project method, “difficulty” is sample proportion correct and “discrimination” is an ordinal association with the observed item-rest score. The resulting weight is then a deterministic function of those sample statistics. Consequently, the proposed item weight is explicitly cohort-dependent unless it is estimated on a calibration sample and frozen for later use.

5 Software architecture, diagnostics, and verification

5.1 R implementation

The current implementation is an R command-line program. Its core dependencies are `data.table` for data handling, `optparse` for the command-line interface, and `Hmisc` for authoritative Somers D_{xy} calculations. Research-only modes use `boot` for bootstrap resampling and `coin` for permutation inference. A separate PISA stability script reconstructs administration patterns, calibrates items independently within each major pattern, and constructs common-rest sensitivity scores.

The default analysis mode uses item-rest scores. A compatibility option allows full totals. The output preserves both empirical and operational values, including original C, empirical D_{xy}, the weighting-base values used for constants, $A = \max(0, D_{xy})$, rarity and concordance pairs, padded proportions, factors, raw weights, minimum-to-one weights, warnings, and status fields.

5.2 Internal identities and diagnostics

The program's extended validation mode checks, among other identities,

$$\begin{aligned} D_{xy} &= 2C - 1, \\ P + Q + T &= n_0 n_1, \\ U &= P + \tfrac{1}{2}T, \\ C &= \frac{U}{n_0 n_1}, \\ D_{xy} &= \frac{P - Q}{n_0 n_1}. \end{aligned}$$

It also cross-checks the Hmisc result against rank/pair-count reconstruction, warns when a response group is small, and retains fractional nA rather than rounding it to an integer.

5.3 Testing status

The current positive-discrimination version was run locally in the project environment with built-in self-tests and a frozen regression-output comparison. Both passed. These tests demonstrate that the program implements the specified equations consistently for the regression fixtures. They do not validate the scoring construct. This distinction is important enough to state formally:

$$\boxed{\text{software validation} \not\Rightarrow \text{psychometric validation.}}$$

5.4 Bootstrap and permutation modes

Bootstrap mode resamples respondents and recalculates item statistics, providing standard errors and percentile intervals for D_{xy}, the remapped discrimination, and item weights. The project also performed a whole-form bootstrap in which all item weights and the minimum-to-one divisor were recalibrated in each resample. This is essential because the operational weight depends on other items through the sample minimum. The bootstrap is a nonparametric case-resampling procedure in the usual sense (Efron and Tibshirani 1993).

Permutation mode uses a conditional randomisation framework through `coin`, providing a complementary null-inference tool for association (Hothorn et al. 2008). Neither bootstrap nor permutation solves sample-dependence; they quantify aspects of uncertainty under specified resampling/reference assumptions.

6 Largest PISA administration-pattern analysis

6.1 Sample and item distribution

The largest reconstructed administration pattern contains $n = 435$ students and $J = 32$ mathematics items. Observed number-correct scores range from 1 to 31. The item difficulties are broad: the most extreme observed item in this form, `pm995q02`, has only 8 correct responses (1.84%), while an easy item has 391 correct (89.89%).

All 32 items have positive empirical Somers discrimination in this pattern, with

$$0.2973 \leq D_{xy} \leq 0.8914.$$

Therefore the positive remapping is inactive in the sense that $A_j = D_{xy,j}$ for every one of these 32 items.

6.2 Item weights

The raw weights range from 1.8124 to 15.6240. Dividing by the minimum gives an operational range from 1.0000 to 8.6207, a mean of 2.1548, a sample SD of 1.4752, and a maximum total operational weight of

$$W_{\max} = 68.9522343.$$

Table 6.1: Ten highest-weighted items in the 435-student PISA pattern.

Item	n_1	n_0	D_{xy}	F_1	F_2	w^{raw}	W
pm995q02	8	427	.8914	4.880	3.201	15.624	8.621
pm406q02	41	394	.7988	3.340	2.594	8.665	4.781
pm923q04	47	388	.7688	3.206	2.457	7.877	4.346
pm803q01t	111	324	.7393	2.359	2.338	5.516	3.043
pm903q03	109	326	.7039	2.377	2.211	5.257	2.901
pm406q01	84	351	.6074	2.635	1.931	5.089	2.808
pm442q02	165	270	.7502	1.966	2.380	4.678	2.581
pm995q01	176	259	.7443	1.901	2.357	4.482	2.473
pm995q03	154	281	.6310	2.034	1.993	4.054	2.237
pm00fq01	142	293	.5673	2.115	1.835	3.880	2.141

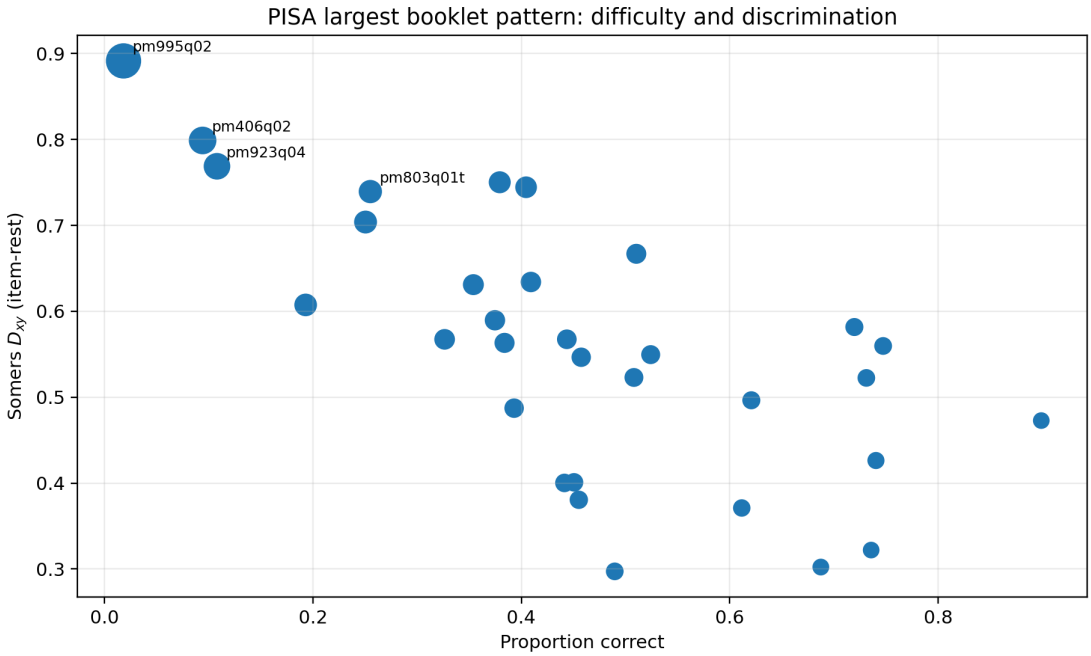

Figure 6.1: Difficulty versus item-rest Somers discrimination for the largest PISA pattern. Marker size reflects the operational item weight; the heaviest items combine low proportion correct with strong positive discrimination.

6.3 Person-score consequences

Number correct and weighted raw score are strongly but not perfectly related:

$$r_{\text{Pearson}} = 0.987195. \tag{6.1}$$

The least-squares fit is

$$\widehat{Y} = -5.45633 + 2.01495R, \tag{6.2}$$

where R is number correct and Y is the weighted raw score.

The main practical effect is tie resolution. Number correct has 31 distinct observed values; the weighted score has 434 distinct values among 435 students. This happens because response patterns with the same number correct generally yield different weighted sums.

The weighted ordering remains strongly anchored to number correct. Among 90,594 student pairs with different number-correct scores, 1,449 pairs (1.60%) are reversed by the weighted score. The largest number-correct gap involved in such an inversion is four items. Thus the method mostly refines the raw-score ordering but can sometimes elevate a lower raw score above a higher one when the solved-item profile is sufficiently high-weighted.

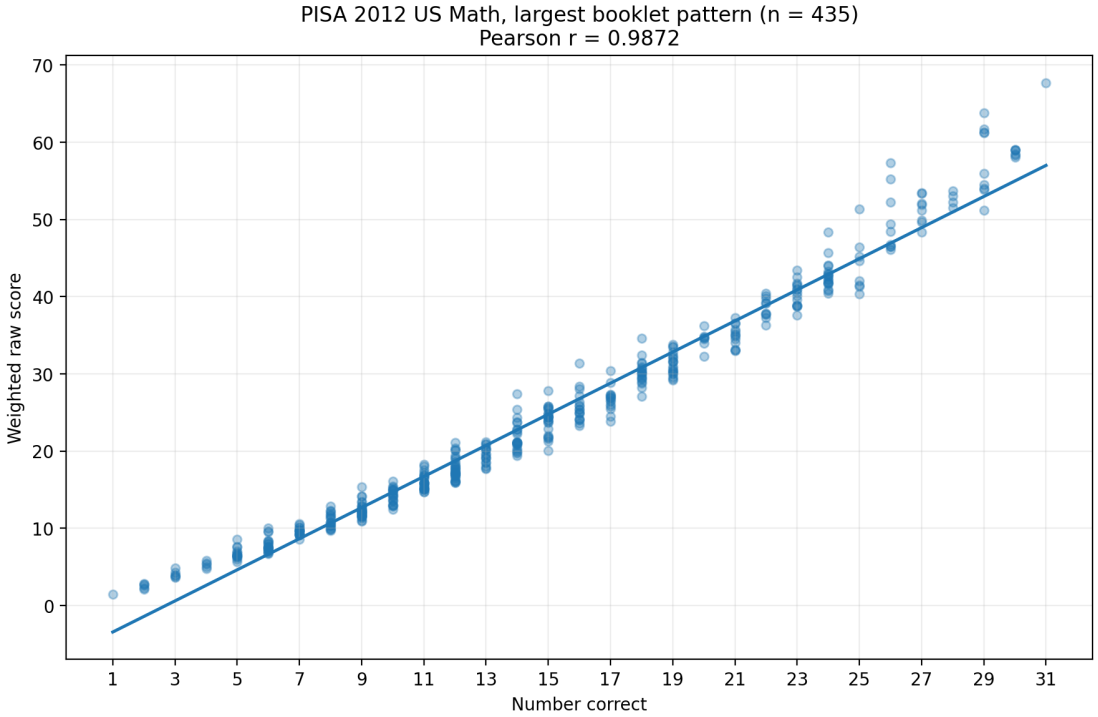

Figure 6.2: Number correct versus weighted raw score for the 435-student PISA administration pattern.

6.4 Interpretation

The near-unity correlation should not be treated as redundancy. A weighted sum constructed from the same binary responses is expected to correlate strongly with number correct, especially when all item weights are positive and lie within a moderate range. The more diagnostic questions are whether response-pattern distinctions are reliable, externally valid, stable across calibrations, and useful for decisions. PISA's repeated-booklet structure allows the stability question to be investigated directly.

7 Case study: the extreme item PM995Q02

7.1 Why the item received a large weight

In the largest pattern, `pm995q02` has

$$c = 8, \qquad n = 435, \qquad p = 0.01839, \qquad D_{xy} = 0.89139.$$

Its two factors are

$$F_1 = 4.8804, \qquad F_2 = 3.2014,$$

so

$$w^{raw} = 15.6240, \qquad W = 8.6207.$$

It accounts for about 12.5% of the largest-pattern maximum operational score by itself.

This is exactly the configuration the method is designed to reward: extreme rarity plus very strong positive rest-score ordering. The OECD identifies PM995Q02 as “Revolving Door 02”, a particularly difficult PISA 2012 mathematics item in the Space and Shape content area and Formulate process. OECD reported a U.S. solution rate of 2.26% and an OECD average of 3.47%, so the rarity observed in this 435-student pattern is not obviously an artefact of the extracted subgroup (OECD 2013a).

7.2 Itemwise bootstrap

The project conducted 50,000 ordinary case-bootstrap replicates for all 32 items in the largest pattern. For PM995Q02, 49,986 replicates retained both response categories; 14 became constant and therefore lacked an empirical Somers coefficient. The central results were:

Table 7.1: Bootstrap uncertainty for PM995Q02 in the 435-student pattern.

Quantity	Point estimate	Bootstrap result
Correct count n_1	8	95% range 3–14
Somers D_{xy}	0.8914	SE 0.0366; 95% CI 0.8128–0.9553
Raw weight	15.6240	mean 16.0985; SE 2.1857
Raw weight 95% CI	–	12.8679–21.3905
Raw-weight CV vs observed	–	14.0%

Even the lower percentile limit for D_{xy} remains high (about 0.813). This argues against interpreting the high discrimination estimate as being sustained only by a few pathological resamples, while the wide raw-weight interval demonstrates that the exact weight magnitude is uncertain.

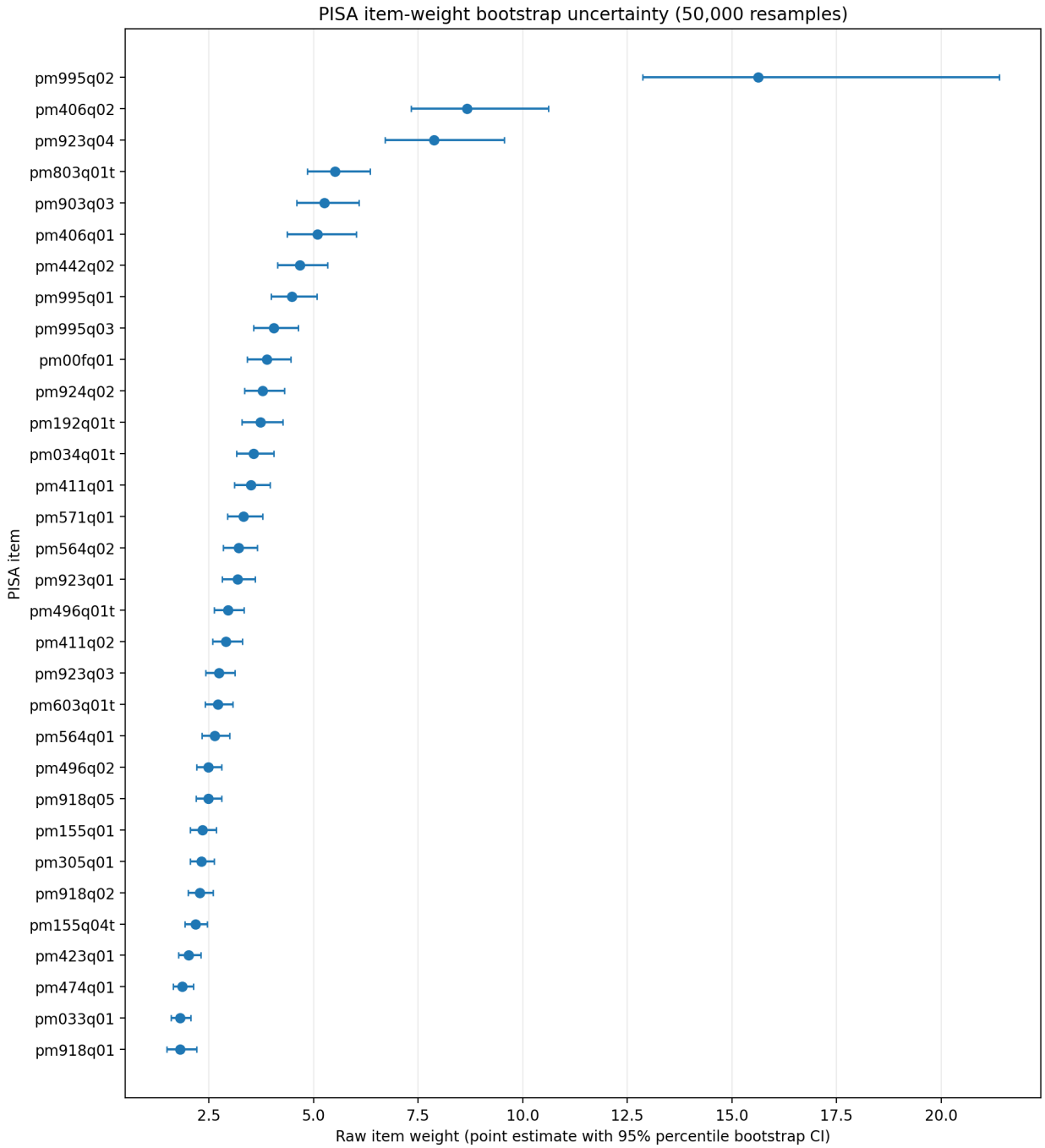

Figure 7.1: Bootstrap raw-weight intervals for the 32 items in the largest PISA pattern.

7.3 Whole-form bootstrap

A second 50,000-replicate analysis recalibrated all 32 items and the minimum-to-one divisor in every resample. PM995Q02's operational weight had mean 9.401, SD 1.424, and percentile interval 7.212–12.779. Its mean share of the total operational weight was 12.75% (95% interval 10.40–16.31%). It remained the largest operational weight in 99.974% of whole-form replicates, with the 2.5th percentile, median, and 97.5th percentile of rank all equal to 1.

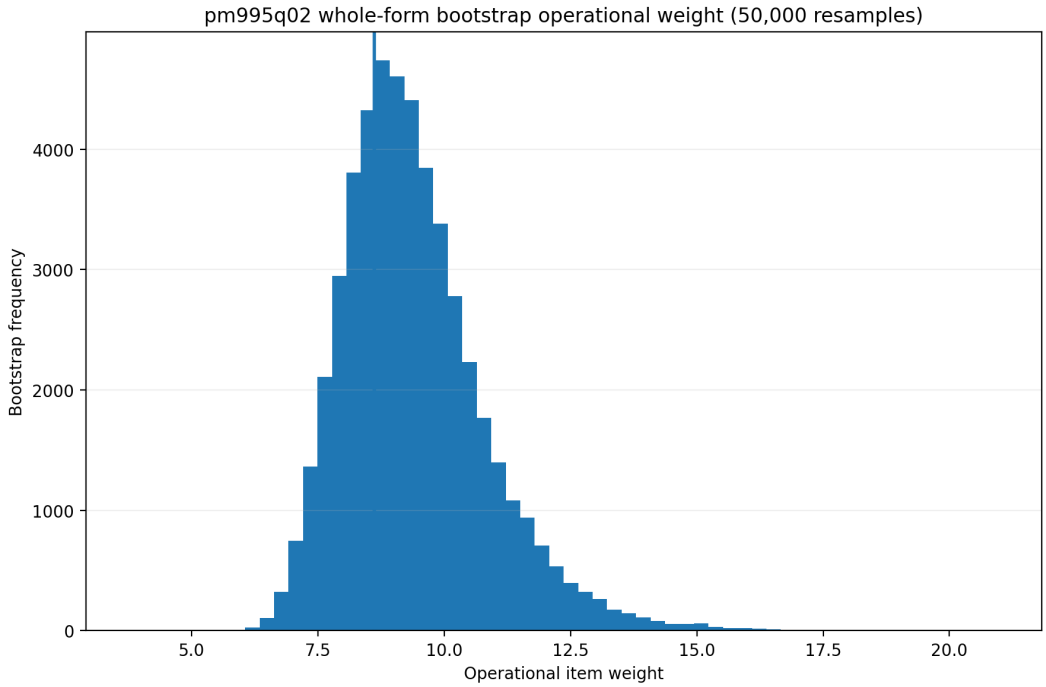

Figure 7.2: Whole-form bootstrap distribution of PM995Q02's operational weight.

7.4 Interpretation of the bootstrap

Two statements are simultaneously justified:

1. the exact numerical weight of PM995Q02 is sampling-sensitive; and
2. its status as an exceptionally influential item in this form is highly robust to case resampling.

Bootstrap resampling, however, only reuses the empirical population represented by one booklet-pattern group. It cannot show whether the same item behaves similarly under independently assigned forms. The cross-booklet analysis addresses that stronger question.

8 Cross-booklet replication of item weights

8.1 Design of the replication analysis

The 13 major reconstructed patterns contain 4,948 students. Every one of the 76 math items appears in four major patterns, closely matching the PISA rotation design. Each item was independently calibrated in each pattern using the form-specific item-rest score. The primary cross-pattern stability metric is the coefficient of variation of the *raw* item weight:

$$\mathrm{CV}_j = \frac{s(w_{jg}^{raw})}{\overline{w}_j^{raw}},$$

where g indexes the four patterns containing item j.

Raw weight is preferable to operational minimum-to-one weight for this comparison because the operational divisor varies by pattern. If an unrelated item becomes the minimum in one booklet, all operational weights in that booklet are rescaled even when the target item's raw weight is unchanged.

8.2 Common-rest sensitivity analysis

Form-specific item-rest discrimination can differ because the other administered items differ between booklets. To reduce that context confound, a second *common-rest* calculation was performed for each target item. It identifies items shared by all four major patterns containing the target item, excludes the target, and sums only this common anchor set. Somers D_{xy} and raw weight are then recomputed against the same anchor content in all four groups.

This control has a cost: the common-rest score is shorter, usually containing only 8–11 items, and therefore can be noisier. Consequently, a lack of CV reduction is not proof that booklet content has no effect; it means that equalising rest-score content did not produce a clear stability improvement under this particular short-anchor design.

8.3 Overall stability results

Across the 76 repeated items, the raw-weight CV distribution is:

Table 8.1: Cross-pattern raw item-weight stability across 76 PISA mathematics items.

Statistic	CV	Threshold	Number of items
Mean	10.52%	CV $> 10\%$	35/76
Median	9.83%	CV $> 15\%$	9/76
25th percentile	7.63%	CV $> 20\%$	6/76
75th percentile	12.89%		
Minimum	2.69%		
Maximum	27.77%		

The common-rest version has mean CV 10.19% and median 9.60%, only slightly lower overall. There is therefore no evidence of a dramatic stability improvement from the common-rest anchor.

Across items, raw-weight CV correlates strongly with the cross-pattern SD of form-rest D_{xy} ($r = 0.736$). It has a weaker negative correlation with mean proportion correct ($r = -0.321$), consistent with rare items tending to be somewhat more unstable, but variation in discrimination is the stronger direct correlate in these data.

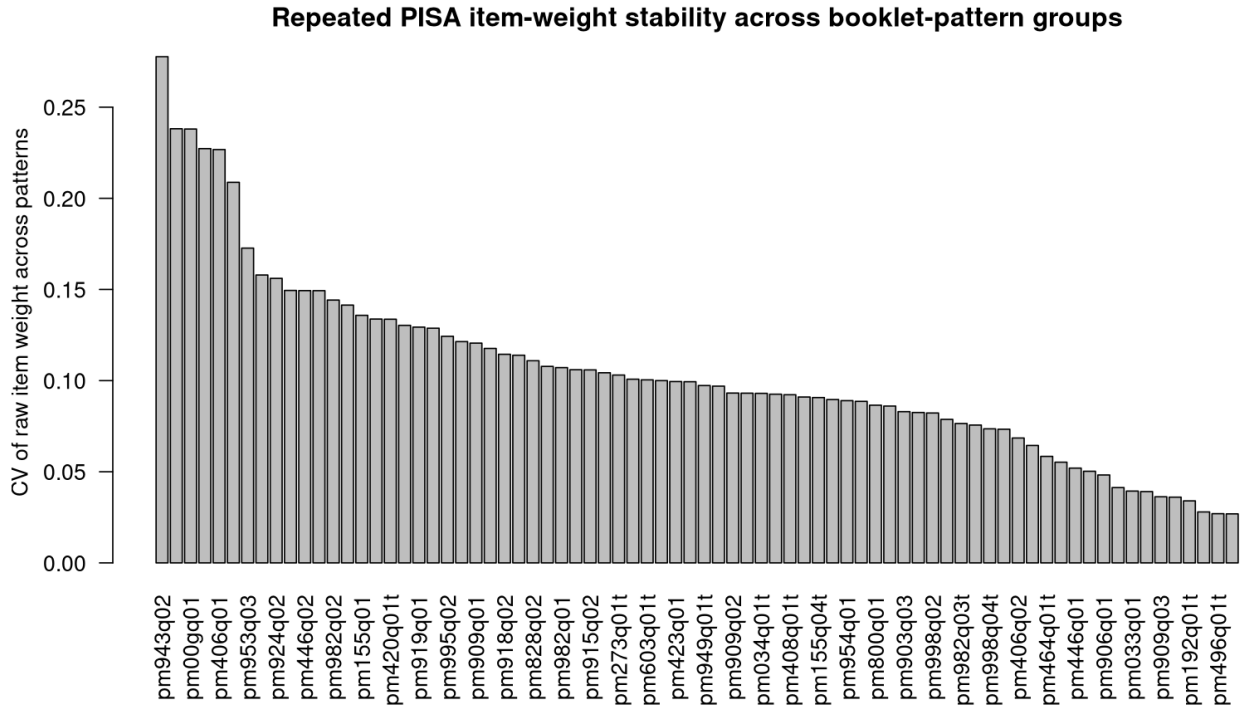

Figure 8.1: Coefficient of variation of raw item weight across the four major PISA pattern groups containing each item.

8.4 Most unstable items

Table 8.2: Six least stable repeated items by form-rest raw-weight CV.

Item	p_{min}	p_{max}	D_{min}	D_{max}	Raw-weight range	CV
pm943q02	.0115	.0382	.7139	.9463	9.446–18.899	27.77%
pm00kq02	.0628	.0761	.3165	.8172	5.149–9.486	23.82%
pm00gq01	.0482	.0647	.5937	.9023	7.556–13.063	23.80%
pm955q02	.1639	.2462	.4775	.7548	4.216–6.674	22.73%
pm406q01	.1394	.1931	.5777	.8423	5.089–8.038	22.68%
pm949q02t	.2664	.2843	.4706	.7810	3.694–5.798	20.88%

The table shows why “weight instability” should not be reduced to difficulty instability. PM00KQ02 has a narrow proportion-correct range (6.28–7.61%) but a very wide D_{xy} range (0.317–0.817), producing a 23.82% raw-weight CV. The discrimination estimate is often the main moving part.

9 PM995Q02 across four independent booklet-pattern groups

9.1 Replication results

PM995Q02 appears in Patterns 01, 06, 07, and 09. The form-specific results are:

Table 9.1: PM995Q02 across four major PISA booklet-pattern groups.

Pattern	n	Correct	p	D_{form}	w^{raw}_{form}	D_{common}	w^{raw}_{common}	W_{form}
01	435	8	.0184	.8914	15.624	.8601	14.410	8.621
06	418	7	.0167	.7560	11.916	.7556	11.909	7.276
07	417	15	.0360	.8602	12.588	.8018	11.122	7.332
09	394	5	.0127	.8334	14.419	.7239	11.829	7.848

The form-rest raw-weight mean is 13.637, SD 1.695, and CV 12.43%. This is only moderately above the median repeated-item CV of 9.83%. PM995Q02 ranks second among the 76 repeated items by mean raw weight but only about 21st by instability. Its pooled correct count across the four groups is 35/1664 (2.10%). The four D_{xy} estimates are all strongly positive (0.756–0.891).

An approximate test of equality of the four proportions gives $\chi^2(3) \approx 6.37$, $p \approx 0.095$. Given the sparse success counts, this does not provide conventional 5% evidence of different correct-response rates across the four groups. The result should be interpreted cautiously because the derived groups are not a simple independent-proportion experiment and PISA's survey design is not represented in this unweighted calculation.

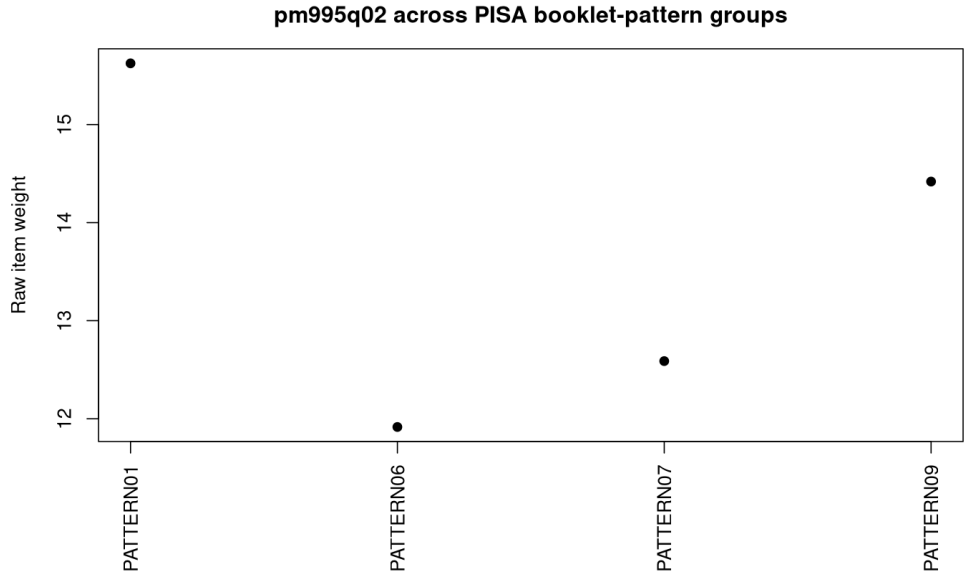

Figure 9.1: Form-rest raw weight for PM995Q02 across the four major pattern groups.

9.2 What the replication adds

The bootstrap established that PM995Q02 is not fragile to resampling within Pattern 01. The four-pattern analysis adds a stronger fact: the item remains both extremely difficult and strongly positively discriminating in three other independently assigned groups. The precise weight changes, but the qualitative conclusion survives. This is the most persuasive evidence in the project so far that at

least one extreme item weight reflects a reproducible response-property combination rather than a one-sample accident.

10 PM943Q02 and the upper end of instability

10.1 Observed variation

PM943Q02 is the least stable item in the current cross-pattern analysis:

$$\mathrm{CV}_{form} = 27.77\%.$$

Its form-rest raw weight ranges from 9.446 to 18.899. The four calibrations are:

Table 10.1: PM943Q02 across its four major PISA pattern groups.

Pattern	n	Correct	p	D_{form}	w^{raw}_{form}	D_{common}	w^{raw}_{common}	W_{form}
02	433	5	.0115	.8724	16.073	.6949	11.522	8.633
03	429	8	.0186	.9463	18.899	.9112	16.537	11.805
05	419	16	.0382	.7139	9.446	.7139	9.446	4.359
07	417	15	.0360	.8866	13.462	.7090	9.500	7.841

The common-rest raw-weight CV is 28.37%, not lower. Hence differing companion-item composition is not sufficient to explain the instability.

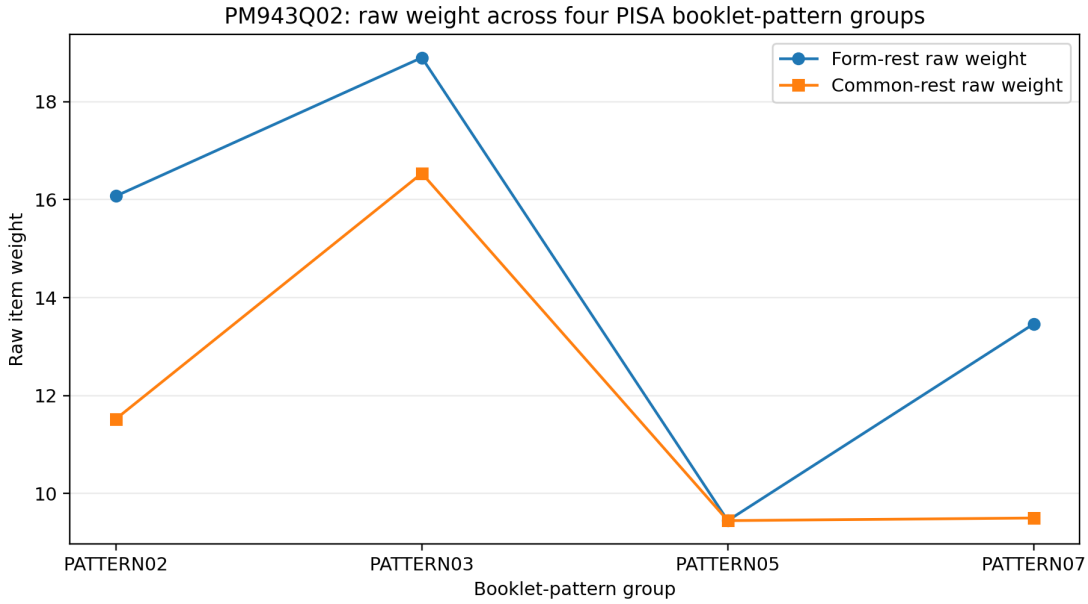

Figure 10.1: PM943Q02 form-rest and common-rest raw weights across its four pattern groups. Equalising the rest-score item set does not eliminate the instability.

10.2 Why this item is a difficult calibration case

Several mechanisms can operate simultaneously.

10.2.1 Very small positive group

The correct count ranges from only 5 to 16 students. Somers D_{xy} is then determined by how these few successful students rank against hundreds of unsuccessful students. Replacing only a few successful cases with students at different rest-score levels can materially change $P - Q$ and therefore D_{xy}.

10.2.2 Nonlinear amplification near $D = 1$

The item is in the region where Equation (3.14) is steep. Moving from $D = 0.71$ to $D = 0.95$ is not treated as a modest additive change: the log-complement factor accelerates as D approaches one. The difficulty factor is also large because the item is rare, so the product amplifies changes in the discrimination factor.

10.2.3 Booklet and position effects

PISA's own calibration methodology includes booklet effects because administration context can shift difficulty (OECD 2014). Cluster position also changes across booklets, allowing fatigue/endurance or order effects to influence response rates. For an item with only a handful of correct responses, a small absolute context effect can be large relative to the positive group.

10.2.4 Not-attempted coding

The derived `edmdata` matrix codes "not attempted" as zero (TMSA Lab 2026). If a late-position item has more nonattempts, this can alter both observed facility and its relation to the rest score. The current binary project files do not retain enough information to decompose each zero into attempted-wrong versus not-attempted, so this mechanism is plausible but not item-specifically quantified here.

10.3 Implication for the scoring formula

PM943Q02 exposes a general design tension: the formula gives the strongest rewards to rare items with very high D, but those are precisely the items for which discrimination can be estimated from very small positive groups and for which the F_2 derivative is largest. This does not prove the formula is unsuitable. It does imply that an operational version should consider uncertainty-aware regularisation, minimum-group thresholds, shrinkage, capping, or frozen calibration based on much larger samples.

11 Strengths revealed by the PISA experiment

11.1 The method behaves as intended on real items

The PISA largest-form results show a clean qualitative pattern: difficult items with stronger positive Somers discrimination tend to receive larger weights. The highest-weighted item has the lowest proportion correct and the highest D_{xy} in that form. Items with the same or similar difficulty can receive different weights when discrimination differs, which is a core design goal.

11.2 Positive discrimination is not an exotic edge case

All 32 items in the largest PISA pattern had positive D_{xy}, and all 24 Penn Matrix items and all 28 ECPE items analysed elsewhere in the project were also positive. Thus the $\max(0, D_{xy})$ truncation serves primarily as a safeguard against weak/reversed items in these real test datasets rather than dominating normal item weights. Nevertheless, the truncation has substantive consequences in noisy or deliberately reversed simulations and should remain explicit in any methodological description.

11.3 Strong tie-breaking resolution

The weighted score produces far more distinct person values than number correct. In the PISA pattern, 434 of 435 students receive distinct weighted raw scores, compared with 31 raw-score levels. For applications where rank resolution is desired, this is a genuine numerical advantage. Whether those distinctions reflect stable construct information rather than measurement noise is an empirical validity question, not something guaranteed by greater granularity alone.

11.4 Some large weights replicate

PM995Q02 is the clearest positive finding. It is not merely rare in one sample; official PISA evidence and the other three pattern groups also show extremely low solution rates. Its discrimination remains strong across all four groups, and its weight has only moderate cross-form CV relative to other repeated items. This is precisely the type of replication evidence needed for a sample-calibrated weighting system.

12 Limitations and threats to validity

12.1 Same-sample calibration and scoring

When weights are estimated on the same respondents who are then scored, each response can affect both the person's earned contribution and the calibration statistics that determine the weight. This creates a form of in-sample circularity. Operational use should estimate weights on a calibration sample, freeze them, and apply those fixed weights to a separate scoring cohort. Cross-fitting is another option for research evaluation.

12.2 Cohort dependence

Both c_j/n and D_{xy} are empirical sample properties. Different populations can generate different weights even when item content is unchanged. The cross-PISA-pattern CV distribution quantifies this problem under one real design. Direct comparison of person scores across administrations is not justified unless calibration is linked or weights are held fixed.

12.3 Extreme-item uncertainty

Rare items can receive very large difficulty factors while their discrimination estimates depend on a small positive group. The PM995 bootstrap and PM943 cross-pattern results illustrate both sides: an extreme weight can be robust, but extreme items can also be the least stable. A fixed threshold such as $\min(n_0, n_1) < 5$ is useful as a warning but is not a general solution. Effective sample size, interval width, and expected impact on total score are more informative.

12.4 Constant-item policy

Retaining all-zero/all-one items with neutral pre-remap $C = 0.5$ is a deliberate project choice, not a standard psychometric convention. An all-zero item can receive a large difficulty-only weight and thereby increase the denominator without anyone earning the contribution. An all-one item contributes to everybody and may anchor the minimum raw weight at one. These items do not contribute discrimination information. For operational use, excluding them from scoring, treating them as unscored, or defining a separate policy would be easier to defend unless there is a substantive reason to retain them.

12.5 Dimensionality and construct dependence

The item-rest score is meaningful as a criterion only if the other items provide a coherent ordering of the proficiency relevant to the target item. Multidimensional item pools, local dependence, speededness, differential strategy use, or heterogeneous formats can make D_{xy} reflect more than target ability. PISA mathematics itself spans content categories and processes; a one-dimensional rest-score ordering is therefore an approximation.

12.6 Missingness and nonattempts in PISA

The PISA experiment uses a binary derived matrix in which not-attempted responses can be coded zero. It also reconstructs booklet patterns from missingness masks rather than from an explicit booklet identifier in the reduced object. This is sufficient for a methodological replication exercise, but a publication-grade PISA analysis should return to the official scored response files, booklet identifiers, missing-response categories, final student weights, and replicate weights.

12.7 Complex survey design ignored in current descriptive analyses

The project analyses reported here do not use PISA final sampling weights or replicate weights. That is acceptable for describing the mechanics of a scoring rule in the selected response groups, but not for making nationally representative statements. Any claim about population distributions, subgroup comparisons, or generalised U.S. performance would require survey-design-aware analysis and the standard plausible-value machinery (OECD 2026b).

12.8 No demonstrated superiority to simpler scores

The current evidence does not show that the weighted person score is more reliable, more valid, fairer, or more predictive than number correct, a Rasch score, a 2PL/IRT score, or another classical weighted score. The high correlation with number correct means any incremental benefit will need strong criterion-based evidence. A useful future design would prespecify external outcomes and compare out-of-sample predictive/construct validity while propagating calibration uncertainty.

13 Recommended next research programme

13.1 Freeze and cross-validate weights

The highest priority is to separate calibration from scoring. Estimate W_j in one large calibration sample, freeze the weights, and evaluate person scores in another sample. For PISA, the four booklet groups can support cross-form experiments, but the unequal companion content makes a linked design preferable to naive interchangeability.

13.2 Hierarchical or empirical-Bayes shrinkage for rare items

The strongest weights often arise from very small correct groups. Instead of directly using raw p_j and $D_{xy,j}$, future versions could shrink difficulty and discrimination toward population/item-pool distributions, with the amount of shrinkage determined by information. This would preserve the principle of rewarding rare/high-discrimination items while reducing boundary volatility.

13.3 Compare capped and uncapped concordance rewards

Because F_2 has rapidly increasing derivative near $A = 1$, test versions with a cap such as $A^* = \min(A, a_{max})$, a softened logistic transformation, or a lower-curvature link. Compare stability and external validity rather than choosing a cap purely for aesthetic reasons.

13.4 Propagate weight uncertainty into person-score uncertainty

Current person scores treat estimated item weights as fixed. A full uncertainty analysis should bootstrap or otherwise model both response noise and calibration uncertainty, producing a conditional or posterior uncertainty interval for S_i. If weights are frozen from an independent large calibration, only the calibration component can be estimated once and reused.

13.5 Differential item functioning and subgroup stability

An item can have high overall D_{xy} while behaving differently across demographic or linguistic groups. PISA is particularly suitable for examining this issue, but the analysis must use the official design and ethically appropriate subgroup definitions. Weighting an item more heavily magnifies any DIF it contains, so fairness analysis is not optional for high-stakes use.

13.6 Prospective score comparisons

A rigorous comparison should include at least:

- number correct / percent correct;
- fixed classical difficulty/discrimination weights;
- Rasch scoring;
- a discrimination-varying IRT model where appropriate;
- the current project method;
- regularised/capped variants of the project method.

Evaluation outcomes should include test–retest or alternate-form reliability, predictive validity, rank stability, classification consistency, subgroup/DIF effects, robustness to missingness, and calibration transportability.

14 Conclusions

The PISA analyses provide a much stronger empirical basis for evaluating the current scoring system than small synthetic examples alone. The method is mathematically coherent as currently specified: item-rest Somers D_{xy} supplies an ordinal discrimination measure; negative discrimination is truncated to zero reward; the rarity and discrimination components use distinct add-one pads; shifted negative logarithms produce positive factors; their product defines a raw weight; and minimum-to-one normalization preserves weight ratios and normalized person scores.

On the 435-student PISA pattern, the method strongly differentiates response patterns while remaining highly correlated with number correct. PM995Q02 demonstrates that a very large weight can be stable under 50,000-replicate bootstrapping and can replicate qualitatively across four independent booklet-pattern groups. Cross-form calibration across all 76 repeated items, however, shows nontrivial variability: the median raw-weight CV is about 9.8%, and several items exceed 20%. PM943Q02 demonstrates the method's most important current vulnerability: rare items with very high but unstable D_{xy} sit in a region where the logarithmic concordance factor is locally very sensitive.

The appropriate conclusion is therefore neither that the method is validated nor that it fails. The data support a narrower assessment: **the current implementation behaves as designed and yields some reproducible item-weight signals in a large real assessment, but calibration uncertainty and cross-form variation are large enough that operational use should require independent/frozen calibration, regularisation studies, uncertainty propagation, and direct validity/fairness comparisons with simpler scoring models.**

15 Reproducibility and project integration

15.1 Primary project files used

The report integrates the current R scoring outputs and the later PISA research stages produced in this project. The companion directory delivered with this report contains:

- `pisa12_math_item_weights.csv`: largest-pattern item audit output;
- `pisa_bootstrap_item_weights_50000.csv`: 50,000-replicate itemwise bootstrap results;
- `pisa_pm995q02_bootstrap_summary.csv`: detailed PM995Q02 itemwise and whole-form bootstrap summary;
- `pisa_major_pattern_summary.csv`: 13 major administration-pattern groups;
- `pisa_all_pattern_item_calibrations.csv`: four-pattern item calibration records;
- `pisa_repeated_item_stability.csv`: cross-pattern stability summary for 76 items;
- `pisa_booklet_stability_analysis.R`: the independent cross-pattern analysis script.

Earlier project documents describing the rank-cut or Spearman phases were used only for development history and contribution provenance; their formulas are not treated as current.

15.2 Representative commands

The largest rectangular PISA pattern was scored with the current program using item-rest scores by default and extended validation, conceptually:

```
Rscript item_analysis.r pisa12_math_largest_pattern.csv \
  --validation \
  --research-diagnostics \
  --items-output pisa12_math_item_weights.csv \
  --persons-output pisa12_math_person_scores.csv
```

The repeated-booklet analysis was then run with:

```
Rscript pisa_booklet_stability_analysis.R
```

The report itself was generated with XeLaTeX/KOMA-Script and BibLaTeX/Biber; the build script is included as a companion artifact.

15.3 Implementation pseudocode

```
for each item j:
    y <- item responses (0/1)
    score <- rowSums(all other items)  # item-rest default
    c <- sum(y)
    n1 <- c; n0 <- n-c

    if item is nonconstant:
        C, Dxy <- Hmisc::somers2(score, y)
    else:
        empirical C, Dxy <- NA
        C_weight_base <- 0.5

    D_weight_base <- 2*C_weight_base - 1
    A <- max(0, D_weight_base)

    p1 <- (c + 1) / (n + 1)
    p2 <- (n*(1-A) + 1) / (n + 1)
    raw_w[j] <- (1 - log(p1)) * (1 - log(p2))

scale <- min(raw_w)
W <- raw_w / scale

for each person i:
    Y[i] <- sum(response[i,j] * W[j])
    S[i] <- Y[i] / sum(W)
```

16 Full AI-use disclosure

16.1 Disclosure statement

Artificial-intelligence assistance was used extensively in the development, research, analysis, and preparation of this report. The AI system was **OpenAI ChatGPT, model GPT-5.6 Sol**, used during August 2026. This disclosure is intentionally broader than a minimal writing-assistance statement because AI support affected mathematical formalisation, software work, research design, literature retrieval, statistical calculations, figure/report generation, and manuscript drafting.

This follows the transparency principle articulated by the ICMJE: authors should disclose which AI tool was used and for what purpose; AI tools should not be listed as authors because they cannot accept responsibility for accuracy, integrity, and originality (International Committee of Medical Journal Editors 2026). Nature Methods similarly requires LLM use to be documented and does not treat LLMs as authors (Nature Methods 2026).

16.2 Human-originated contributions

The substantive project concept and successive design decisions originated with the human project owner. In particular, the human contribution includes:

- the objective of rewarding difficult items more strongly while also rewarding items whose correct responses are concentrated among stronger test takers;
- the two-factor/padded-log product architecture and the correction that the difficulty and discrimination pads must have different directions;
- the decision to use an established rank-association statistic rather than the earlier custom rank-cut statistic;
- the selection of Somers D_{xy} / concordance as the discrimination basis after comparison with Spearman and Kendall alternatives;
- the project-specific remapping $A = \max(0, D_{xy})$, so non-positive discrimination receives no reward;
- the decision to retain constant items with an operational neutral pre-remap $C = 0.5$ rather than silently claiming an empirical coefficient;
- the minimum-to-one multiplicative normalization that preserves weight ratios;
- the use of PISA, ECPE, Penn Matrix, and simulation datasets to stress-test the method;
- the request for bootstrap and cross-booklet replication of extreme PISA item weights.

16.3 AI-assisted contributions

ChatGPT's contributions included:

- translating iterative verbal specifications into formal mathematical notation and deriving algebraic properties, bounds, monotonicity, and sensitivity results;
- comparing candidate discrimination measures (Spearman, Kendall, rank-biserial/Somers) and identifying Hmisc's Somers D_{xy} implementation as a suitable established basis;
- drafting and refactoring R implementations, command-line interfaces, diagnostics, validation identities, self-tests, bootstrap/permutation modes, and stability-analysis scripts;
- identifying edge cases such as constant items, small response groups, part-whole contamination, sample dependence, and the consequences of different normalizations;
- web searching and synthesising documentation/literature concerning PISA design, PISA scaling, Somers D, item-rest analysis, bootstrap methods, and AI disclosure;
- computing or independently checking descriptive statistics from project CSV outputs, including correlations, regressions, item-weight summaries, bootstrap intervals, and cross-pattern CVs;
- generating figures and the XeLaTeX/KOMA-Script manuscript structure, bibliography, equations, tables, and explanatory prose.

16.4 What AI did not do

The AI system is not an author and has no legal or scholarly responsibility for this work. It did not collect the original PISA responses, administer the assessment, or create the underlying PISA measurement framework. The response data are derived from publicly distributed PISA materials/package representations. The AI did not independently execute the user's local R runs that produced all project CSVs; many computations were performed by the human project owner in a local R environment and then supplied to the analysis. This report also does not claim that a human expert has independently peer-reviewed every derivation or source statement.

16.5 Verification and responsibility

AI-generated output can contain errors. Accordingly, the report distinguishes direct project-output calculations, published-source claims, and methodological interpretation. Key numerical results were recomputed from the locally available CSV files when possible; the current scoring implementation had previously passed its built-in self-tests and frozen regression-output test in the project environment. Official PISA claims in the report were checked against OECD sources, and the statistical definition of Somers D_{xy} was checked against Hmisc documentation and the literature.

Before journal submission, public release, high-stakes use, or any claim of psychometric superiority, the complete analysis should receive independent human statistical/psychometric review. Final responsibility for any use, interpretation, or publication of this report rests with the human project owner, consistent with current AI-publishing guidance (International Committee of Medical Journal Editors 2026).

16.6 Disclosure reproducibility note

The AI interaction was iterative rather than a single prompt. The broad instruction for this report was to use a deep-research-like process, integrate relevant project material, explain the scoring system completely, use web research, produce a XeLaTeX/KOMA-Script report with narrow margins, and include a full AI-use disclosure. During the underlying project, many intermediate prompts refined formulas, software behaviour, diagnostics, datasets, and analyses. Because the full conversational history is lengthy and includes superseded specifications, reproducing every prompt in the manuscript would obscure rather than clarify the current method. The mathematical specification in Chapter 3, the companion data outputs, and the disclosed division of human/AI contributions are intended to provide the scientifically relevant provenance.

Bibliography

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC. DOI: 10.1201/9780429246593.

Harrell, F. E. J., C. Beck, and C. Dupont (2026). *Hmisc: Harrell Miscellaneous*. R package version 5.2-6. DOI: 10.32614/CRAN.package.Hmisc. URL: https://CRAN.R-project.org/package=Hmisc.

Henrysson, S. (1963). "Correction of Item-Total Correlations in Item Analysis". In: *Psychometrika* 28.2, pp. 211–218. DOI: 10.1007/BF02289618.

Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis (2008). "Implementing a Class of Permutation Tests: The coin Package". In: *Journal of Statistical Software* 28.8, pp. 1–23. DOI: 10.18637/jss.v028.i08.

International Committee of Medical Journal Editors (2026). *Use of Artificial Intelligence by Authors*. URL: https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html (visited on 08/28/2026).

Mann, H. B. and D. R. Whitney (1947). "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1, pp. 50–60. DOI: 10.1214/aoms/1177730491.

Metsämuuronen, J. (2020). "Somers' D as an Alternative for the Item–Test and Item-Rest Correlation Coefficients in the Educational Measurement Settings". In: *International Journal of Educational Methodology* 6.1, pp. 207–221. DOI: 10.12973/ijem.6.1.207.

Nature Methods (2026). *Preparing Your Material*. URL: https://www.nature.com/nmeth/submission-guidelines/preparing-your-submission (visited on 08/28/2026).

OECD (2013a). *Lessons from PISA 2012 for the United States*. Paris: OECD Publishing. DOI: 10.1787/9789264207585-en. URL: https://doi.org/10.1787/9789264207585-en.

– (2013b). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing. DOI: 10.1787/9789264190511-en. URL: https://doi.org/10.1787/9789264190511-en.

– (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing. DOI: 10.1787/6341a959-en. URL: https://doi.org/10.1787/6341a959-en.

– (2026a). *PISA 2012 Database*. URL: https://www.oecd.org/en/data/datasets/pisa-2012-database.html (visited on 08/28/2026).

– (2026b). *PISA Research Documentation: How to Prepare and Analyse the PISA Database*. URL: https://www.oecd.org/en/about/programmes/pisa/how-to-prepare-and-analyse-the-pisa-database.html (visited on 08/28/2026).

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen and Lydiche.

Somers, R. H. (1962). "A Similarity between Goodman and Kruskal's Tau and Kendall's Tau, with a Partial Interpretation of the Latter". In: *Journal of the American Statistical Association* 57.300, pp. 804–812. DOI: 10.1080/01621459.1962.10500818.

TMSA Lab (2026). *items_pisa12_us_math: PISA 2012 United States Mathematics Item Responses*. URL: https://rdrr.io/github/tmsalab/edmdata/man/items_pisa12_us_math.html (visited on 08/28/2026).

Warm, T. A. (1989). "Weighted Likelihood Estimation of Ability in Item Response Theory". In: *Psychometrika* 54.3, pp. 427–450. DOI: 10.1007/BF02294627.

A Selected algebraic checks

A.1 Minimum-to-one normalization leaves the normalized person score unchanged

If $W_j = w_j^{raw}/m$ with $m > 0$, then

$$\frac{\sum_j x_{ij} W_j}{\sum_j W_j} = \frac{\sum_j x_{ij} w_j^{raw}/m}{\sum_j w_j^{raw}/m} = \frac{\sum_j x_{ij} w_j^{raw}}{\sum_j w_j^{raw}}.$$

Therefore the final minimum-to-one normalization changes the displayed item-weight scale but not S_i.

A.2 Relationship of C, D_{xy}, and U

From

$$C = \frac{P + T/2}{n_0 n_1}$$

and $P + Q + T = n_0 n_1$,

$$\begin{aligned} 2C - 1 &= \frac{2P + T - (P + Q + T)}{n_0 n_1} \\ &= \frac{P - Q}{n_0 n_1} = D_{xy}. \end{aligned}$$

Also $U = P + T/2$, so $C = U/(n_0 n_1)$.

A.3 Boundary behaviour of the discrimination factor

For $0 \le A \le 1$,

$$F_2(A) = 1 + \ln\left(\frac{n+1}{n(1-A)+1}\right).$$

Hence

$$F_2(0) = 1, \qquad F_2(1) = 1 + \ln(n+1),$$

and

$$F_2'(A) = \frac{n}{n(1-A)+1}, \qquad F_2''(A) = \frac{n^2}{[n(1-A)+1]^2} > 0.$$

The factor is therefore increasing and convex. The convexity formalises the empirical sensitivity seen in high-D PISA items.

B Selected data tables

B.1 All 32 largest-pattern items

Table B.1: All 32 items in the largest PISA administration pattern, sorted by operational weight.

Item	n_1	p	D_{xy}	F_1	F_2	w^{raw}	W
pm995q02	8	0.0184	0.8914	4.880	3.201	15.624	8.621
pm406q02	41	0.0943	0.7988	3.340	2.594	8.665	4.781
pm923q04	47	0.1080	0.7688	3.206	2.457	7.877	4.346
pm803q01t	111	0.2552	0.7393	2.359	2.338	5.516	3.043
pm903q03	109	0.2506	0.7039	2.377	2.211	5.257	2.901
pm406q01	84	0.1931	0.6074	2.635	1.931	5.089	2.808
pm442q02	165	0.3793	0.7502	1.966	2.380	4.678	2.581
pm995q01	176	0.4046	0.7443	1.901	2.357	4.482	2.473
pm995q03	154	0.3540	0.6310	2.034	1.993	4.054	2.237
pm00fq01	142	0.3264	0.5673	2.115	1.835	3.880	2.141
pm924q02	178	0.4092	0.6341	1.890	2.001	3.783	2.087
pm192q01t	163	0.3747	0.5895	1.978	1.887	3.732	2.059
pm034q01t	167	0.3839	0.5632	1.954	1.825	3.566	1.968
pm411q01	222	0.5103	0.6669	1.670	2.095	3.499	1.931
pm571q01	193	0.4437	0.5674	1.810	1.835	3.321	1.832
pm564q02	171	0.3931	0.4872	1.930	1.666	3.215	1.774
pm923q01	199	0.4575	0.5465	1.779	1.788	3.181	1.755
pm496q01t	228	0.5241	0.5495	1.644	1.795	2.950	1.628
pm411q02	221	0.5080	0.5230	1.675	1.738	2.911	1.606
pm923q03	192	0.4414	0.4002	1.815	1.510	2.740	1.512
pm603q01t	196	0.4506	0.4009	1.794	1.511	2.711	1.496
pm564q01	198	0.4552	0.3806	1.784	1.478	2.637	1.455
pm496q02	270	0.6207	0.4964	1.476	1.684	2.484	1.371
pm918q05	313	0.7195	0.5817	1.328	1.868	2.482	1.369
pm155q01	325	0.7471	0.5596	1.291	1.817	2.346	1.294
pm305q01	213	0.4897	0.2973	1.712	1.352	2.314	1.277
pm918q02	318	0.7310	0.5224	1.312	1.737	2.279	1.258
pm155q04t	266	0.6115	0.3711	1.490	1.462	2.180	1.203
pm423q01	322	0.7402	0.4263	1.300	1.554	2.020	1.115
pm474q01	299	0.6874	0.3023	1.374	1.359	1.867	1.030
pm033q01	320	0.7356	0.3221	1.306	1.388	1.813	1.000
pm918q01	391	0.8989	0.4728	1.106	1.638	1.812	1.000

Note. The appendix table is a compact report table; the companion CSV is authoritative for full precision and all diagnostic columns. Minor rounding in displayed factor values can cause displayed products to differ slightly from the listed raw weight.

C Companion-file provenance

The generated report directory contains the source, bibliography, build script, report-analysis asset script, figures, and selected project CSVs. A companion `SHA256SUMS.txt` file can be generated with `sha256sum` for archival or publication. This report does not embed or redistribute the full OECD PISA database; it includes only project-derived summaries/selected response-analysis outputs already present in the working environment.