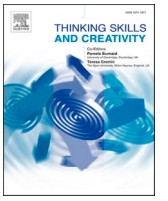
Thinking Skills and Creativity 62 (2026) 102228

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Solving the paradox between the generalized partial credit model and PISA's rating instructions: A generalized IRTree approach to within-rating multidimensionality

Nils Myszkowski [a,*], Martin Storme [b]

[a] Department of Psychology, Pace University, NY, USA
[b] IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Économie Management, 59000 Lille, France

ARTICLE INFO

Keywords:
Psychometrics
Creativity
Item response theory
Measurement

ABSTRACT

In the PISA 2022 Creative Thinking test, students produce open-ended responses that are scored by human raters as no credit, partial credit, or full credit. As in many large-scale assessments, PISA scales these ratings using the generalized partial credit model. We show that the rating instructions conflict with the GPCM assumption of within-rating unidimensionality: raters are asked to consider different attributes at different thresholds, such as appropriateness and originality or diversity. Using multidimensional generalized item response tree (IRTree) models, we account for the sequential nature of the ratings and disentangle the attributes contributing to each threshold. Multidimensional models outperformed unidimensional ones, indicating within-rating multidimensionality. The latent traits were strongly correlated, but only at levels comparable with those observed between creative thinking and mathematics, reading, and science. We discuss whether collapsing attributes remains reasonable, and other implications for score validity and interpretation.

Creativity is increasingly recognized as an important skill across a variety of domains, most notably in education. The inclusion of creative thinking in PISA by the OECD testifies to this growing recognition and reflects an increasing interest in assessing creativity on a global scale. As with any large-scale project that seeks to capture such a multifaceted construct (Barbot et al., 2019; Myszkowski, 2024), certain trade-offs, compromises, and shortcuts are inevitable. This paper focuses on one such issue that has been raised theoretically (Benedek & Beaty, 2025; Cuesta-Hincapie & Camargo Salamanca, 2025; Myszkowski & Storme, 2025) but not yet explored empirically: Within-rating multidimensionality.

Before discussing it specifically, we shall note that, in addition to the complexities inherent to large-scale international assessment, the psychological construct that is creativity poses multiple inherent challenges related to its multidimensionality (Barbot et al., 2019; Myszkowski, 2024; Myszkowski et al., 2024). Therefore, the structure of the PISA 2022 Creative Thinking assessment itself reflects several layers of multidimensionality. At the broadest level, the assessment spans four domains of application — written expression, visual expression, social problem solving, and scientific problem solving — each intended to capture different contexts in which creative thinking may manifest. Within these domains, items are organized around three distinct facets of creative thinking: generating diverse ideas (GDI), generating creative ideas (GCI), and evaluating and improving ideas (EII). Each facet emphasizes a different cognitive demand, from producing multiple appropriate and varied responses, to providing a single original solution, to refining an

* Corresponding author.
E-mail address: nmyszkowski@pace.edu (N. Myszkowski).

https://doi.org/10.1016/j.tsc.2026.102228
Received 11 October 2025; Received in revised form 1 April 2026; Accepted 10 April 2026
Available online 13 April 2026

existing idea in novel ways. Taken together, this design illustrates the complex, multifaceted nature of creative thinking as assessed in PISA (Barbot & Kaufman, 2025) and raises important questions about whether the construct can ultimately be treated as unitary (Barbot & Kaufman, 2025; Hernández-Ramos & Araya, 2025) or whether its different layers require differentiated psychometric treatment. Building on these layers of complexity, we now turn to a new challenge: the possibility that the rating scale may, itself, capture multiple underlying dimensions, a phenomenon we refer to here as within-rating multidimensionality.

1. Within-rating multidimensionality in PISA creative thinking ratings

In the assessment, students engage in different activities depending on the facet being measured, whether generating diverse ideas, generating creative ideas, or evaluating and improving ideas. Across these facets, their productions are judged by trained raters using standardized scoring guides (OECD, 2024b). Ratings are made on an ordinal three-point scale, with 0 point indicating no credit, 1 point indicating partial credit, and 2 points indicating full credit. This approach is consistent with long-standing practices in creativity assessment, where product ratings, which are typically captured through ordinal scales (Myszkowski & Storme, 2019), are widely used as the basis for a person's creativity measurement. This approach to creativity measurement is commonly referred to in creativity psychology as the consensual assessment technique (Amabile, 1982) although the psychometric community more commonly refers to it as rater-mediated assessment (e.g., Engelhard Jr. et al., 2018; Wind & Engelhard Jr., 2016). In PISA these ratings serve as the foundation for scaling responses within each facet (GCI, EII and GDI).

At first glance, such ratings suggest a straightforward ordinal progression along a single underlying dimension of creative performance, with higher scores reflecting higher levels of creative thinking. This interpretation is reflected in PISA's use of the generalized partial credit model as the response model for scaling, which notably relies on what we will refer to here as the assumption of within-rating unidimensionality. We define this assumption as the idea that, for a given item and rater, a single latent attribute (i.e., here, creative thinking) is sufficient to account for the progression from 0 to 1 to 2 points. In other words, the observed rating is modeled as the manifestation of one underlying trait, with higher levels of that trait increasing the probability of receiving higher scores across all thresholds of the scale. This assumption can clearly be seen in the formulation of the Generalized Partial Credit Model (Muraki, 1992), previously discussed in the PISA Creative Thinking context by Myszkowski and Storme (2025). Importantly, the assumption that we discuss here only concerns the interpretation of thresholds within a rating scale and does not preclude the possibility of other forms of multidimensionality in the assessment, such as those that we have previously discussed here (e.g., induced by variability across domains or facets).

Creativity is commonly defined as the combination of originality (or novelty) and appropriateness (or usefulness/effectiveness), rather than originality alone (Runco & Jaeger, 2012). While there is general agreement on the importance of these two components, their relationship is less clearly understood. A long-standing view is that originality and appropriateness may stand in tension, such that highly original ideas are often less well adapted to task constraints, whereas highly appropriate ideas are often more conventional (Diedrich et al., 2015; Dumas et al., 2025). At the same time, other work suggests a more asymmetric or conditional relationship, whereby originality becomes diagnostic of creativity primarily once a response satisfies a minimum level of appropriateness. Consistent with this view, experimental studies show that originality and appropriateness contribute differently to creativity judgments and do not combine in a purely additive way (Runco & Charles, 1993). Taken together, these perspectives suggest that creativity judgments may rely on multiple attributes that do not operate uniformly across all levels of performance, rather than on a single underlying dimension.

This perspective is reflected in the PISA scoring guides, which suggests that the assumption of within-rating unidimensionality may not hold. As noted in the PISA Technical Report (OECD, 2024b), the transition from no credit to partial credit generally reflects whether a response is appropriate to the task (although for the GDI items it also requires diversity), while the transition from partial to full credit only requires originality or diversity. The term "diversity" is used here to remain consistent with the terminology of the PISA scoring rubrics, but it should be noted that this dimension is conceptually related to what is often referred to as flexibility in the creativity literature — see Forthmann (2026) for a discussion on this topic. This means that different thresholds along the same scale are, per the instructions, tied to distinct (sets of) constructs (Benedek & Beaty, 2025; Cuesta-Hincapie & Camargo Salamanca, 2025; Myszkowski & Storme, 2025). More specifically, in a Generate Diverse Ideas (GDI) task, the first threshold (0→1) combines two attributes: the set of responses must include at least one appropriate idea, and the responses must show some degree of diversity. Full credit (1→2) then requires that the ideas are also sufficiently different from one another to demonstrate diversity. In a Generate Creative Ideas (GCI) task, a response that repeats a common theme in a conventional way would be scored as appropriate (1 point) but not original, while a response that departs from conventional themes or elaborates on them in an unusual way would earn full credit (2 points). Finally, in an Evaluate and Improve Ideas (EII) task, a basic improvement that addresses the problem in an appropriate but predictable way is sufficient for partial credit (1 point), whereas a more original improvement — one that introduces a novel feature or perspective while remaining appropriate — is required for full credit. Across all three facets, the scoring rubrics therefore combine appropriateness with either diversity or originality, with different sets of attributes required to pass the two thresholds. Using the generalized partial credit model, which assumes that a single latent dimension predicts the progression across thresholds, is thus inconsistent with the rating instructions (Myszkowski & Storme, 2025), which explicitly rely on different attributes at different points of the scale.

2. Generalized item response tree models

An alternative framework that directly addresses these concerns (Myszkowski & Storme, 2025) is offered by (generalized) Item

Response Tree (IRTree) models (De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). The central idea of IRTree models is to represent an observed categorical response not as the consequence of a single continuous latent trait, but as the outcome of a sequence of choices along a decision tree, with each of these choices involving a specific set of latent traits. In this set of models, the overall rating is obtained by passing through the decision nodes (also referred to as pseudo-items) in sequence (e.g., here, passing the first threshold, then, if passed, passing the second). Because IRTree models allow different nodes to be linked to distinct (though possibly correlated) latent attributes, they allow the disentangling of several attributes from a single item response or rating. For instance, in the current situation, a three-point rating can be modeled as two successive decisions that account for the rating instructions: First, whether a response meets a minimum threshold of appropriateness (and for GDI, diversity), and second, conditional on passing the first threshold, whether it additionally demonstrates originality or diversity (Myszkowski & Storme, 2025). This decomposition has the advantage of disentangling the processes that underlie different thresholds, making explicit the within-rating multidimensionality that is otherwise concealed within a single rating scale.

An important note that had not been made in Myszkowski and Storme (2025) is that, in the case where multidimensionality is not assumed but suspected (like here), the (multidimensional) IRTree model can be empirically compared with its unidimensional particular case, the Tutz sequential model (Tutz, 1990), in which all nodes are affected by the same attribute. Comparisons between a multidimensional IRTree model and a unidimensional IRTree model have precedent (e.g., Partchev & De Boeck, 2012), including in creativity psychology (Forthmann et al., 2019), and, in this case, such a comparison provides a principled decision-making strategy when one wants to choose whether to retain within-rating unidimensionality or multidimensionality. Based on previous work by Myszkowski and Storme (2025), we propose an illustration of the models for such comparisons in Fig. 1 for the GDI items and in Fig. 2 for the GCI and EII items. The multidimensional IRTree models were specified to reflect the structure of the PISA scoring rubrics. For GCI and EII items, the first rating threshold is defined exclusively in terms of appropriateness, with originality considered only at the second threshold; accordingly, the effects of originality were fixed to zero at node 1 for these facets. In contrast, for GDI items, diversity is explicitly involved in the initial decision process, which motivated a compensatory specification at node 1. A simpler multidimensional specification in which diversity loaded only on the second node was also estimated for GDI, but showed poorer fit and was therefore not retained.

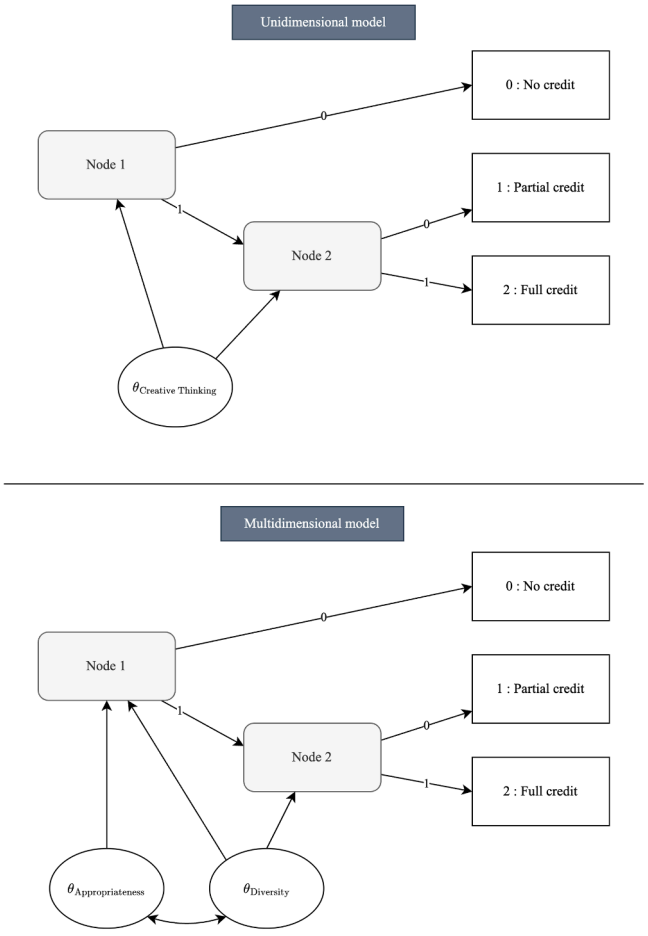


Fig. 1. Schematic representation of Generalized IRTree models for the GDI items.

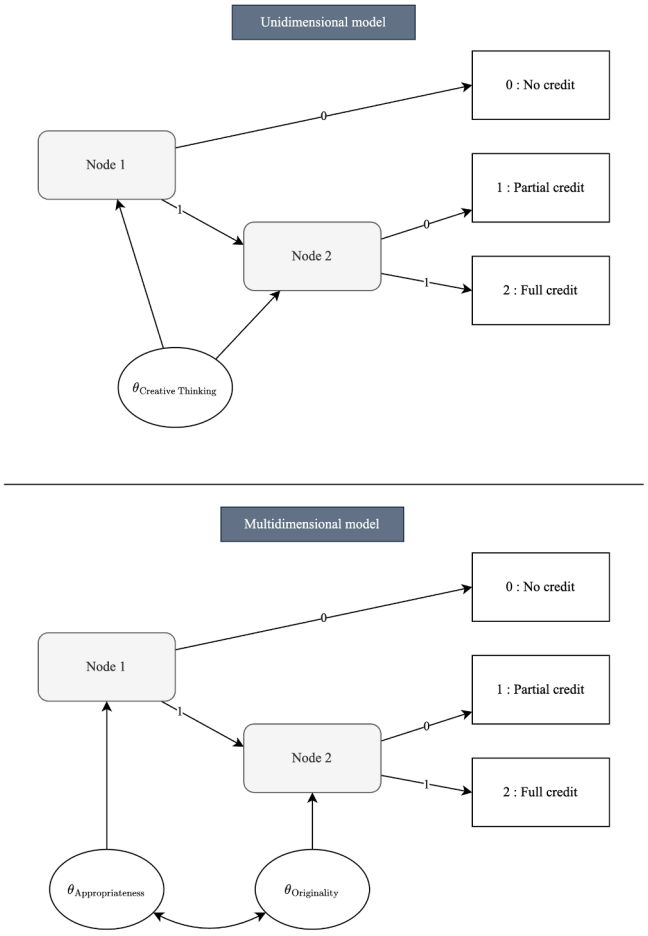


Fig. 2. Schematic representation of Generalized IRTree models for the GCI and EII items.

The latent traits represented in these figures are grounded directly in the PISA 2022 competency model and scoring procedures rather than introduced as independent theoretical constructs. Across all three facets, the scoring rubrics implement a sequential evaluation process in which appropriateness constitutes a necessary first criterion, followed (conditional on passing this threshold) by the evaluation of either originality or diversity (OECD, 2024a). Appropriateness refers to whether a response respects task requirements and demonstrates a minimum level of usefulness or relevance. Originality is defined in the PISA framework as statistical infrequency relative to the response pool, serving as a proxy for novelty in tasks requiring a single idea (GCI, EII), whereas diversity captures the heterogeneity of multiple responses within a single task and is conceptually related to ideational flexibility (GDI). These dimensions align with familiar distinctions in creativity research, but are modeled here only to capture the rater decision criteria specified in the PISA coding guides. The latent traits therefore represent process-level scoring components rather than full theoretical constructs of creative thinking.

3. Aims of this paper

In this paper, we aimed to investigate empirically the assumption of within-rating unidimensionality (i.e., given a rating, the assumption that a single attribute underlies all rating scale thresholds), which is both implicitly formulated by OECD's use of the generalized partial credit model, and contradicted by the rating instruction. To our knowledge, this paradox, recently noted by Myszkowski and Storme (2025), has not been empirically investigated. We thus propose to estimate the models proposed in this previous paper on the three facets (GDI, GCI, EII) (which assume within-rating multidimensionality) and to compare them with models with a similar structure but which assume within-rating unidimensionality (i.e., the Tutz sequential model). Aligned with this previous paper and the rating instructions, we hypothesized that the ratings would indeed be multidimensional, and therefore that multidimensional models would provide a better fit to the PISA data. To note, given that the GDI items embed both appropriateness and diversity in the same initial decision process, the underlying multidimensionality is less distinct than in GCI or EII. Consequently, we expected the multidimensional specification to provide only modest improvement over the unidimensional model for this facet.

Although we are primarily concerned with fit, an important consequence of disentangling ratings into processes involving different

attributes, is that the second node (which is supposed to capture diversity or originality) is only evaluated if the first node was successfully passed. In other words, it is considered that a node 2 response is missing every time that no credit was given (Myszkowski & Storme, 2025). Consequently, for a given person, there is a smaller or equal number of node 2 observations than of node 1. The sequential nature of the ratings therefore implies a loss of information (and thus, of reliability) for the attributes measured at node 2. Whereas it is still possible that this loss could be compensated by second nodes having stronger discrimination parameters (which yields more information), we saw no reason to expect such a compensatory effect, so we anticipated lower reliability (and thus larger standard errors) for attributes measured at node 2 in comparison with node 1.

The present study operates at the level of item responses and rating processes, rather than at the level of reported PISA scale scores or plausible values. Our analyses focus on the measurement models used to scale the trichotomous ratings assigned by human raters, and on the assumptions these models make about the latent attributes underlying different rating thresholds. Plausible values, which are generated as downstream products of the scaling model, are therefore not analyzed directly here. Instead, the goal of this study is to evaluate whether the measurement assumptions underlying the model used to produce these plausible values are consistent with the scoring instructions and the cognitive attributes they invoke.

4. Methods

4.1. Dataset

We re-analyzed the PISA Creative Thinking dataset, which is publicly available from the OECD, and contains item scores for the respondents. The only alteration that was made to the data was the exclusion from analysis of items that were scored on a binary scale rather than ordinal, because the models are not applicable to these items. This resulted in three subsets of items, with 8 items for GDI, 11 items for GCI and 9 items for EII. The sample of cases with at least one item response on the remaining items was comprised of 142,431 respondents, after excluding respondents with completely missing responses across these items. Such missingness is expected in PISA due to the rotated booklet design, whereby students are administered different subsets of items. The average age in this sample was 15.8 years, with 50.0% male and 50.0% female according to the standardized gender variable provided by the OECD (more demographics are made available notably through the student survey data, also available from OECD).

Due to the latent traits being measured differing (e.g., diversity vs. originality) and because some items of different facets came from the same task (thus creating local dependencies), we decided to primarily study separately the three facets (GDI, GCI, EII), however pooling together across the domains to keep enough items for analysis. Analyses were conducted on the pooled international sample, thereby implicitly assuming measurement invariance across countries. This assumption was adopted as a pragmatic modeling choice in the present study and reflects the fact that no demographic variables were included in the analyses. Concretely, this implies that the same item parameters and latent trait structure were assumed to apply across countries, such that observed differences are interpreted as reflecting differences in the underlying latent traits rather than differences in how items function or are scored across contexts. This assumption should not be interpreted as evidence for, or a substantive claim about, measurement invariance across countries, but rather as a simplifying working hypothesis adopted for the purposes of the present analyses. The potential impact of violations of measurement invariance is addressed in the Discussion and outlined as an important direction for future research.

For each trichotomous item, we created 2 separate node variables, where node 1 represented passing the first threshold and node 2 represented passing the second. Node 1 was coded 0 if a respondent received no credit, and 1 if they received any credit (partial or full), while node 2 was coded missing if a respondent received no credit, 0 if they received partial credit, and 1 if they received full credit (see Myszkowski & Storme, 2025 for a table representation of this mapping).

The PISA 2022 Creative Thinking data used in this study are publicly available from the OECD (https://www.oecd.org/pisa/data/). To support transparency and reproducibility, all data preprocessing steps (starting from the raw OECD data files), model specifications, and analysis scripts used to produce the results reported in this paper are documented and shared in an open repository on the Open Science Framework (OSF; https://osf.io/z9tgk/).

4.2. Models

For each facet, we estimated two item-response tree models: A within-rating unidimensional model (i.e., Tutz model) and a within-rating multidimensional model. While the former represents the assumption of within-rating unidimensionality, the latter was specified to mimic the rating instructions, which depends on the facet considered. To lighten the notation, throughout, g represents the logit (link) function. Also, throughout, i represents the respondent, j the item and X_{ij} a respondent's (original) score (i.e., 0, 1 or 2). Finally, we labelled the latent traits (and their discrimination parameters) with the attributes that they are supposed to represent (i.e., creative thinking, appropriateness, diversity, originality). This is to facilitate understanding, but our labeling should not be interpreted as evidence or claim that the estimated dimensions necessarily correspond exactly to these constructs. Rather, the labels are theoretical placeholders that reflect the intended meaning of each node within the scoring framework.

4.2.1. Unidimensional IRTree model

The baseline model that represented the assumption of within-rating unidimensionality was a (generalized) Tutz model (Tutz, 1990), in which each node decision was modeled using a 2-parameter logistic model. Applying this model to the present data, the probability to pass the first node $P(X_{ij} > 0)$, depends on a person's ability $\theta_{\text{Creative Thinking},\, i}$, item-node discrimination $a_{\text{Creative Thinking},\, j}$

and item-node easiness b_j:

$$P(X_{ij} > 0) = g^{-1}\left(a_{\text{Creative Thinking, }j}\theta_{\text{Creative Thinking},i} + b_j\right),$$

The second node probability conditional upon passing the first node $P(X_{ij} = 2|X_{ij} > 0)$, depends on the person's ability $\theta_{\text{Creative Thinking, }i}$, item-node discrimination $a'_{\text{Creative Thinking, }j}$ and item-node easiness b'_j:

$$P(X_{ij} = 2|X_{ij} > 0) = g^{-1}\left(a'_{\text{Creative Thinking},j}\theta_{\text{Creative Thinking},i} + b'_j\right).$$

It is important to note that, although both node probabilities involve different node-item parameters, the person parameter $\theta_{\text{Creative Thinking, }i}$ is here the same across nodes.

4.2.2. Multidimensional IRTree model

The multidimensional models used were as described and argued for in Myszkowski and Storme (2025). More specifically, for the GDI items, the probability to pass node 1 involves both appropriateness and diversity in a compensatory model:

$$P(X_{ij} > 0) = g^{-1}\left(a_{\text{Appropriateness},j}\theta_{\text{Appropriateness},i} + a_{\text{Diversity},j}\theta_{\text{Diversity},i} + b_j\right),$$

while the (conditional) probability to pass node 2 is given by a unidimensional model:

$$P(X_{ij} = 2|X_{ij} > 0) = g^{-1}\left(a'_{\text{Diversity},j}\theta_{\text{Diversity},i} + b'_j\right).$$

For the GCI and EII items, the probability to pass node 1 is given as:

$$P(X_{ij} > 0) = g^{-1}\left(a_{\text{Appropriateness},j}\theta_{\text{Appropriateness},i} + b_j\right),$$

while the probability to pass node 2, conditional upon having passed node 1 is given as:

$$P(X_{ij} = 2|X_{ij} > 0) = g^{-1}\left(a'_{\text{Originality},j}\theta_{\text{Originality},i} + b'_j\right).$$

4.2.3. Unconditional probabilities

Because the sequential models are previously (and more conveniently) described using conditional probabilities to pass the nodes, we shall note that, for all models, the (unconditional) probability of receiving no credit $P(X_{ij} = 0)$ is given as:

$$P(X_{ij} = 0) = 1 - P(X_{ij} > 0).$$

The unconditional probability of receiving partial credit $P(X_{ij} = 1)$ is given as:

$$P(X_{ij} = 1) = P(X_{ij} > 0)\left[1 - P(X_{ij} = 2|X_{ij} > 0)\right].$$

Finally, the unconditional probability of receiving full credit $P(X_{ij} = 2)$ is given as:

$$P(X_{ij} = 2) = P(X_{ij} > 0)P(X_{ij} = 2|X_{ij} > 0).$$

5. Estimation

All models were estimated on the recoded (i.e., item node level) data, using the expectation-maximization (EM) algorithm implemented in the package "mirt" (Chalmers, 2012) for R. All models were identified by fixing the latent variance to 1, and covariances between latent traits (if any) were freely estimated.

6. Model comparisons

Because the Tutz model was not nested within the multidimensional models, we used the information criteria returned by "mirt" and commonly used to compare non-nested IRT models (Myszkowski, 2021), which are the Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), Sample-Adjusted Bayesian Information Criterion (SABIC; Sclove, 1987) and Hannan-Quinn criterion (HQ; Hannan & Quinn, 1979), in order to decide on which model had better fit. For all these criteria, a lower value indicates better fit. Due to the amount of missing data produced by the assessment design, limited information goodness-of-fit indices (Maydeu-Olivares & Joe, 2006) were not available.

7. Model parameters

The models were inspected using node-item response function plots for the item parameters. We also inspected correlations between latent variables, as we considered them to be informative regarding the relevance of using multidimensional models. It was of interest to examine the correlations between latent traits in the multidimensional models, as, beyond the fit of the models, these could indicate whether it may be reasonable/pragmatic to collapse these traits. A useful benchmark for comparison is the observed correlations between PISA 2022's creative thinking scores and the other performance tests (mathematics, reading and science), which

ranged between 0.66 and 0.67 and still considered distinct as a result (OECD, 2024a).

8. Impact on reliability and relations with achievement

Since we hypothesized better reliability for traits measured at node 1 as opposed to node 2, we computed empirical reliabilities for the different latent traits. Bootstrapping was used (1000 replications) to obtain confidences intervals, as has been previously used for comparisons of reliabilities computed from item response theory models (e.g., Myszkowski & Storme, 2018, 2024; Storme et al., 2019). Since all traits had the same variance of 1 (fixed for identification), their standard errors are comparable, and therefore, we produced overlayed density plots of the standard errors of the traits involved for each set of items.

In addition, to examine relations with PISA core domains, Pearson correlations were computed between the estimated IRTree latent traits and students' mathematics, reading, and science performance. Following recommended practices for plausible values, correlations were computed separately for each of the ten plausible values per domain and combined using Rubin's (1987) rules.

9. Joint modeling of rating processes across facets

In addition to the facet-specific models described above, we estimated a joint IRTree model across all facets in order to examine relations among the latent components underlying GDI, GCI, and EII within a unified framework. Rather than introducing facet-level latent variables, we focused on the rating processes implied by the scoring instructions. Specifically, two global latent traits were specified: one capturing appropriateness, defined as the ability to pass the first decision node across all facets, and one capturing originality/diversity, defined as the ability to pass the second decision node across all facets. In line with the scoring rubric for GDI items, where diversity already contributes to the first decision, the first node of GDI items was allowed to depend on both latent traits. The two traits were allowed to correlate. This specification allows all facets to be modeled jointly while preserving the process-level interpretation of the IRTree framework. Facet-level unidimensional joint models were also explored but proved empirically unstable due to near-collinearity between facet factors, and were therefore not retained. Model fit was evaluated using the same information criteria as for the facet-specific models.

In addition, to examine relations among facet-specific dimensions, a second joint IRTree model was estimated in which separate latent variables were specified for each facet-by-process combination (i.e., appropriateness and originality/diversity within GDI, GCI, and EII). In this model, items were assigned to latent traits according to their position in the IRTree (node 1 vs. node 2), with GDI node 1 loading on both appropriateness and diversity in accordance with the scoring rubric. All latent traits were allowed to correlate freely. This specification enables the estimation of model-implied correlations among all facet-specific dimensions within a unified measurement model.

10. Results

10.1. Model fit comparisons

The fit indices of the different models are presented in Table 1. As expected, for GDI, the results were mixed, with the BIC and SABIC favoring the unidimensional model, and the AIC and HQ favoring the multidimensional model. For the EII and the GCI subtests however, all fit indices favored the multidimensional models. Overall, these results indicate that the multidimensional models were better fitting. This pattern extended to the joint analysis across subtests, for which all information criteria favored the two-factor model.

10.2. Correlations between latent traits

The latent traits were strongly and positively correlated in all subtests. More specifically, for GDI, latent appropriateness and diversity were correlated at 0.619. For EII and GCI, latent originality and appropriateness were correlated respectively at 0.695 and 0.691. These correlations are however similar to the correlations between creative thinking and reading, mathematics and science, which are still considered distinct in PISA (OECD, 2024a).

To further examine the relations among all estimated dimensions across facets, a full correlation matrix of the latent traits was

Table 1
Model fit comparisons.

Subtest	Model	Log-Likelihood	AIC	BIC	SABIC	HQ
GDI	Unidimensional IRTree	−400,612.1	801,288.3	801,603.2	801,501.5	801,382.5
	Multidimensional IRTree	−400,573.4	801,228.7	801,632.2	801,501.9	801,349.4
EII	Unidimensional IRTree	−420,185.1	840,442.3	840,796.8	840,682.4	840,548.3
	Multidimensional IRTree	−420,139.9	840,353.8	840,718.2	840,600.6	840,462.7
GCI	Unidimensional IRTree	−541,482.2	1083,052.4	1083,486.2	1083,346.4	1083,182.0
	Multidimensional IRTree	−541,363.2	1082,816.4	1083,260.1	1083,117.1	1082,949.0
Joint	Unidimensional IRTree	−1307,830.5	2615,885.0	2616,990.1	2616,634.2	2616,215.2
	Multidimensional IRTree	−1307,127.0	2614,495.9	2615,689.8	2615,305.3	2614,852.6

obtained from a joint multidimensional IRTree model in which all facet-specific dimensions were estimated simultaneously (see Table 2). This model included separate latent variables for appropriateness and originality/diversity within each facet (GDI, GCI, and EII), with all latent traits freely correlated. Because these correlations are model-implied, they reflect the relationships among latent variables directly, without relying on factor score approximations.

As shown in Table 2, correlations were uniformly positive and strong within each facet (i.e., between appropriateness and originality/diversity), but substantially higher correlations were observed across facets within attribute (e.g., appropriateness across GDI, GCI, and EII, originality/diversity across GDI, GCI, and EII). This pattern indicates that the multidimensional structure embedded within ratings–distinguishing appropriateness from originality/diversity–captures systematic variation that is not reducible to a single underlying dimension. At the same time, the strong cross-facet correlations suggest that these components generalize across task types, reflecting stable but partially distinct processes underlying creative performance.

10.3. Node item response functions

We provide the node item response function plots for the GDI, EII and GCI respectively in Fig. 3, Fig. 4 and Fig. 5. Although all nodes had (expectedly) positive discrimination parameters, these plots show that there were some items that had sufficient discrimination ($a \geq 0.5$) at node 1 with low discrimination ($a' < 0.5$) at node 2: GCI 11, EII 03, EII 04, EII 06, EII 07, EII 08, EII 09. This indicates that some items better captured appropriateness than originality.

11. Reliability

We present the reliability estimates of the different traits in Table 3. Originality (captured at node 2) measures yielded larger standard errors/lower reliability than appropriateness for GCI and EII. It was not the case for GDI, which may be attributed to the fact that both traits are already measured at node 1. This is in line with our hypothesis that the sequential nature of the ratings would lead to lower reliability for attributes of node 2, due to node 2 being missing by design when node 1 is not passed. Fig. 6 displays the empirical distributions of standard errors for the latent traits estimated at node 1 and node 2. Because all latent variances were fixed to 1 for identification, these standard errors are directly comparable across traits. The figure therefore provides a distributional view of measurement precision, highlighting the shift toward larger uncertainty for node-2 traits relative to node-1 traits for GCI and EII.

12. Correlations with other PISA domains

We present the Pearson correlations between the IRTree latent traits and PISA mathematics, reading, and science performance in Table 4. All facet-level traits showed strong correlations with performance in PISA core domains, while the global appropriateness and originality/diversity traits exhibited somewhat stronger associations. Differences across facets and domains were small and highly consistent. These results indicate that the latent rating processes captured by the IRTree models are systematically related to general academic performance without suggesting strong differentiation between latent traits defined by node position or by subtest.

13. Discussion

In this study, we compared models that assumed within-rating unidimensionality and models that assumed within-rating multidimensionality using, as previously suggested (Myszkowski & Storme, 2025), generalized IRTree models. Questioning the assumption of within-rating unidimensionality in this way was substantively motivated by the rating instructions of PISA, in which raters are instructed to focus on different attributes for different thresholds of the rating scale (Benedek & Beaty, 2025; Cuesta-Hincapie & Camargo Salamanca, 2025; Myszkowski & Storme, 2025). Overall, as suspected, model fit comparisons indicated that ratings were more multidimensional than unidimensional, suggesting that ratings did capture different traits at different thresholds. Despite this, correlations between latent attributes (appropriateness and diversity/originality) and within facets (GDI, EII, GCI) were strong, indicating that, even though the attributes were distinct, they overlapped. However, these associations were consistently lower than across facets for the same attribute, reinforcing the interpretation that ratings reflect multiple partially distinct dimensions rather than a single latent continuum. Finally, we observed some imbalance in the reliability of the traits, with originality being less reliably measured than appropriateness for the EII and GCI traits, which was expected because originality is considered for the second

Table 2
Latent correlations.

	1	2	3	4	5
1. GDI - Appropriateness					
2. GDI - Diversity	0.593				
3. GCI - Appropriateness	0.932	0.811			
4. GCI - Originality	0.581	0.768	0.701		
5. EII - Appropriateness	0.805	0.879	0.938	0.759	
6. EII - Originality	0.532	0.637	0.617	0.855	0.586

Note. All correlations significant at $p<.001$.

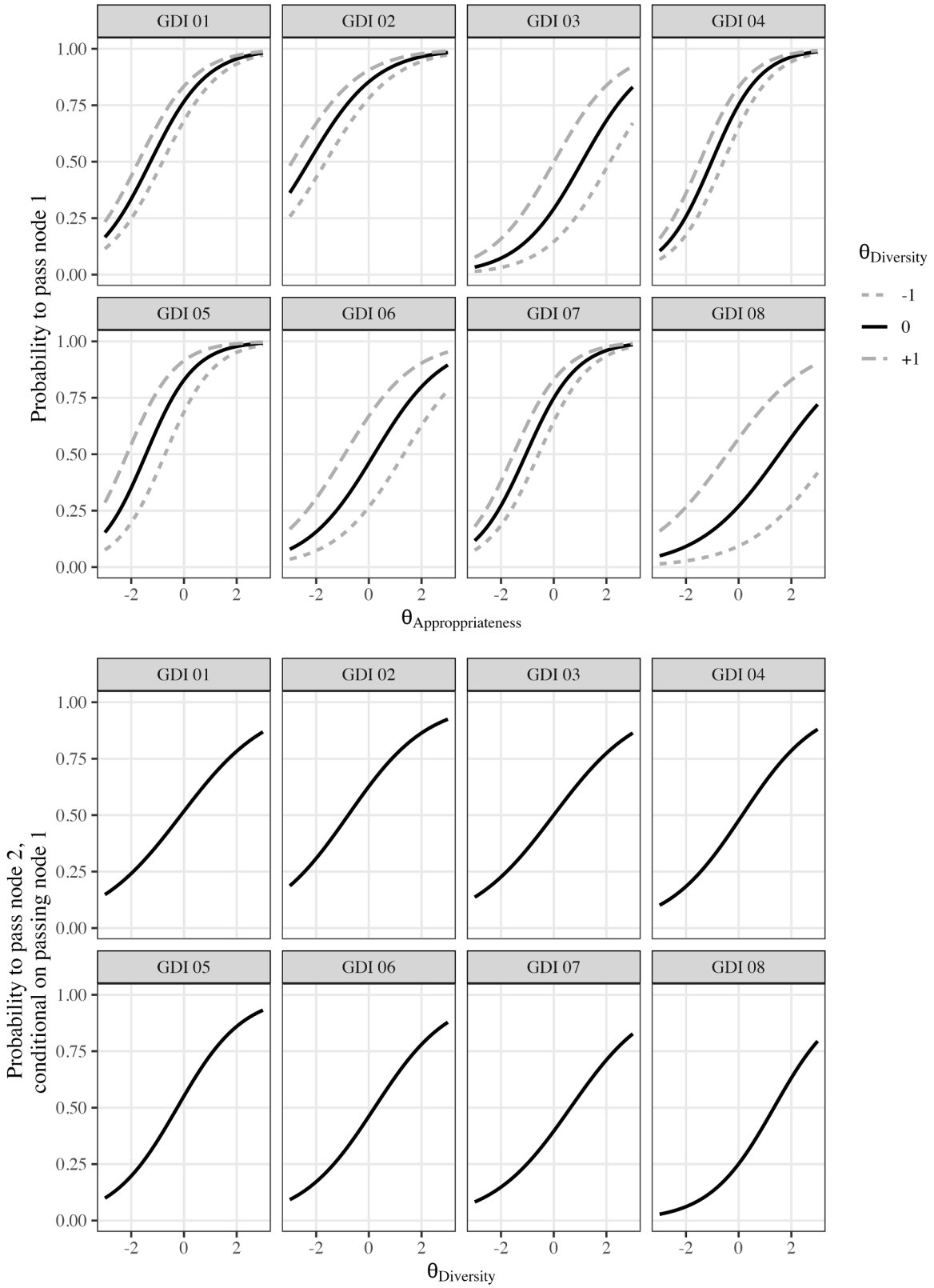


Fig. 3. Node item response functions of the GDI items.

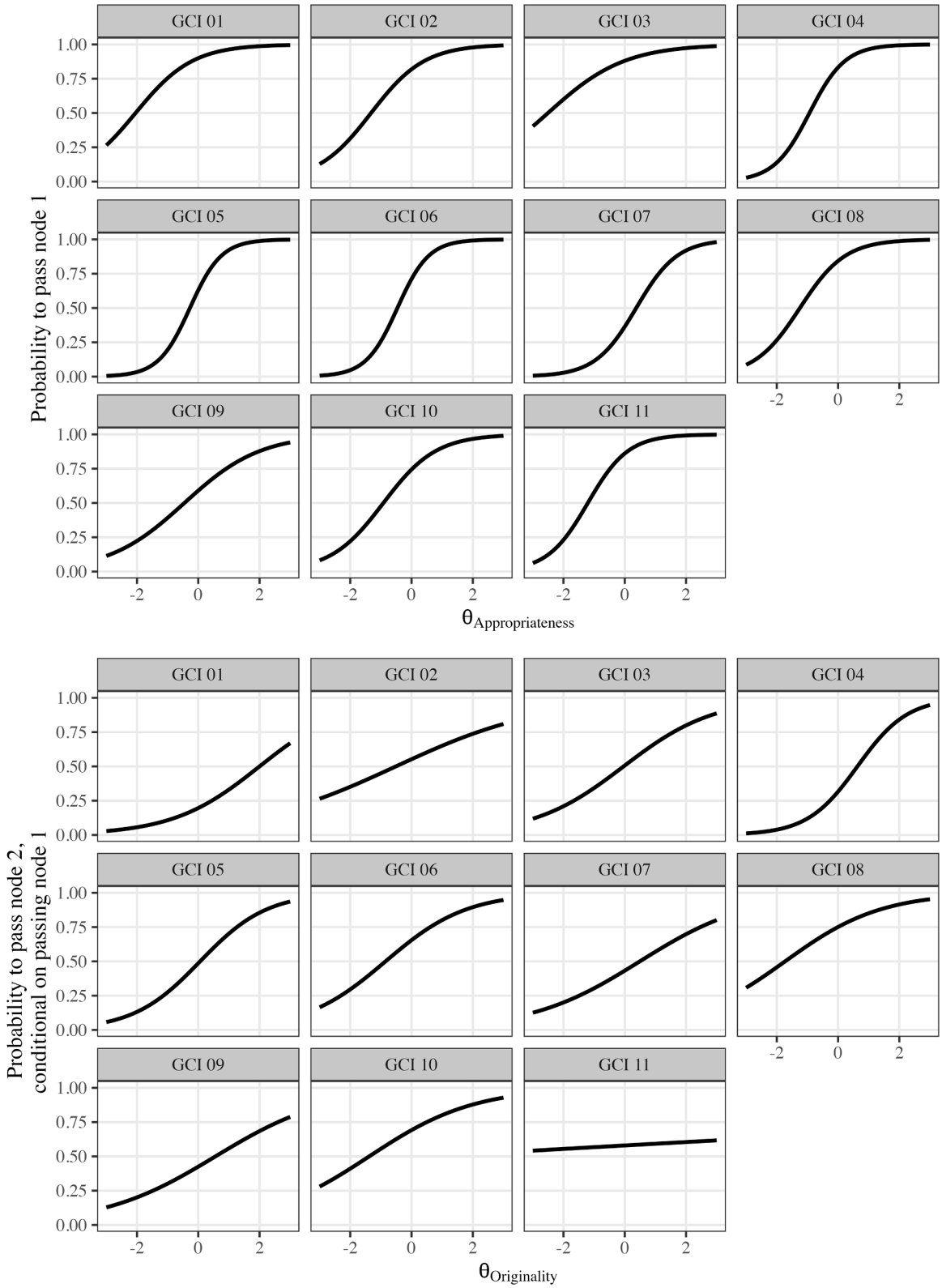


Fig. 4. Node item response functions of the GCI items.

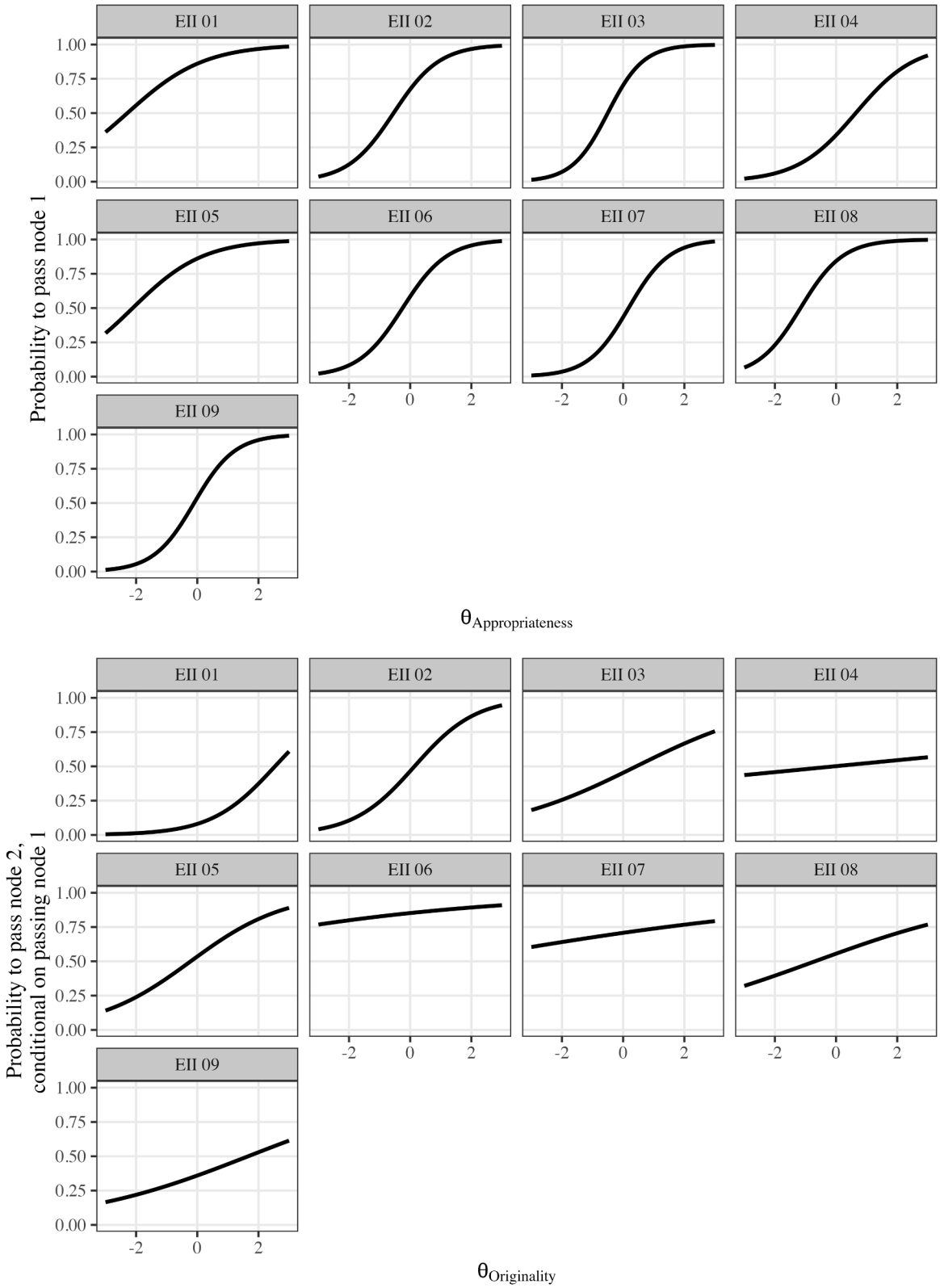


Fig. 5. Node item response functions of the EII items.

Table 3
Empirical reliability comparisons.

Subtest	Reliability of $\theta_{\text{Appropriateness}}$	95% Bootstrapped CI	Reliability of $\theta_{\text{Diversity/Originality}}$	95% Bootstrapped CI
GDI	.397	[.395,.398]	.415	[.414,.417]
GCI	.519	[.518,.521]	.362	[.361,.363]
EII	.461	[.460,.462]	.280	[.279,.281]
Joint	.746	[.745,.757]	.637	[.636,.639]

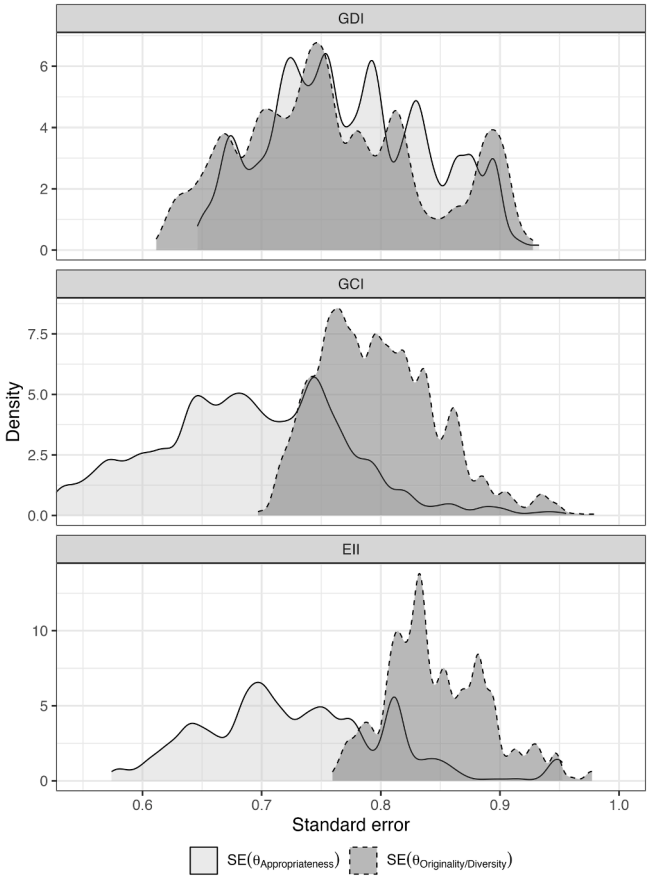


Fig. 6. Density plots of standard errors in the sample.

Table 4
Correlations between factor scores and performance in PISA core domains.

IRTree latent trait	Mathematics	Reading	Science
GDI - Appropriateness	0.535	0.531	0.531
GDI - Diversity	0.546	0.538	0.541
GCI - Appropriateness	0.533	0.551	0.538
GCI - Originality	0.523	0.534	0.527
EII - Appropriateness	0.514	0.528	0.518
EII - Originality	0.497	0.509	0.500
Global - Appropriateness	0.641	0.652	0.643
Global - Originality / Diversity	0.647	0.653	0.647

Note. All correlations significant at $p<.001$.

threshold, which is only reached if some credit is given, and therefore less frequently observed than appropriateness for these scales.

In spite of the low estimates of reliability, it is important to note, however, that the PISA Creative Thinking assessment is not intended for individual-level diagnosis or decision-making. Rather, its primary purpose is to support population-level and group-level comparisons. In this context, modest reliability at the individual level is less problematic, as aggregated estimates can still be

informative for comparative and descriptive purposes. Nevertheless, low precision limits the interpretability of individual differences and reinforces the need for caution when drawing fine-grained conclusions about specific creative attributes. This applies to the correlations reported in this very study between the computed factor scores and other achievement variables.

These findings have important implications for the validity of the original creative thinking plausible values obtained using the GPCM on these ratings. Indeed, as previously raised (Benedek & Beaty, 2025; Cuesta-Hincapie & Camargo Salamanca, 2025; Myszkowski & Storme, 2025), in the presence of within-rating multidimensionality (which this study shows), it becomes unclear which attribute each rating really represents, and therefore which blend of attributes a factor scores (or their distributions) ultimately represent. Further, it implies that the relative contribution of appropriateness and originality to their score differs across individuals, which is a violation of measurement invariance.

13.1. Choosing between unidimensional and multidimensional models

Although model fit comparisons favored within-rating multidimensionality, we noted that strong positive correlations were still observed between the latent traits, for all subtests. The pattern of associations with mathematics, reading, and science performance followed the same direction, with highly similar correlations observed across nodes and subtests. In other words, the capacity to produce appropriate responses and the capacity to produce original/diverse responses, although distinct, largely overlapped. This overlap can be seen as reassuring for the current use of the GPCM, as it suggests that treating the ratings as indicators of a single underlying dimension may provide a reasonable approximation of overall creative performance.

A first issue with these strong correlations is that they may be reflections of the measurement rather than the constructs themselves. Notably, they may indicate that raters did not clearly differentiate between appropriateness and originality or diversity when assigning scores, effectively conflating the two attributes. In that case, the observed unidimensionality would reflect rater behavior rather than the true structure of the construct, underscoring the need for clearer rating instructions or rater training to ensure that distinct aspects of creative thinking are evaluated independently.

Yet, even if partly artefactual, is it reasonable to collapse attributes here? We would argue that it shall probably depend on the researcher's objectives. Indeed, being able to separate originality and appropriateness can be interesting in specific research contexts where these dimensions are examined separately. For example, one could be interested in the trade-offs individuals make between the generation of novel ideas and their relevance/feasibility. However, in other contexts, researchers could prefer to rely on an overall measure of creative thinking rather than separate scores. In the broader IRTree literature, it is common to combine multiple correlated dimensions into a single score when they are conceptually similar (e.g., Celik et al., 2022; Lang et al., 2019; Lievens et al., 2018; Storme et al., 2020). For creativity assessment, however, this aggregation raises a challenge: originality and appropriateness interact and possibly trade-off rather than simply add up, but a clear formula on how to exactly combine these two attributes remains to be defined (Myszkowski & Storme, 2025). This lack of clarity could tempt researchers who are motivated by pragmatism to rely on a unidimensional modeling approach that yields a single creative thinking score, even though, from a theoretical perspective, the assessment process is clearly multidimensional. They should however remain aware that a unidimensional scoring approach aggregates distinct individual capacities, which may compromise the meaning of the resulting scores. Further, the fact that the correlations observed between originality/diversity and appropriateness were of similar magnitude to those found between creative thinking and mathematics, reading, and science in PISA raises a conceptual problem: if one were to justify collapsing originality and appropriateness into a single creative thinking dimension on the basis of their correlations, then, by the same logic, one should also collapse creative thinking with mathematics, reading, and science.

13.2. Reliability issues

The node-level item response functions highlighted additional problems. More specifically, discrimination was uneven across thresholds (a pattern that cannot be accommodated within a model that relies on a single item-level discrimination parameter like the GPCM). Indeed, several items (GCI 11 and EII 03–09), displayed satisfactory discrimination at node 1 but poor discrimination at node 2. This pattern may suggest that raters were able to judge appropriateness more accurately than originality or diversity. Appropriateness tends to be defined by clearer and more objective criteria, whereas originality and diversity are rarer and more subjective, making them harder to evaluate reliably. Since each response was rated by a single judge, these discrepancies may also reflect differences in item clarity or in how easily the scoring criteria could be applied, rather than solely differences in rater accuracy. In any case, this imbalance in reliability may indicate that the overall reliability of the instrument may be driven primarily by the measurement of appropriateness, while the originality component contributes relatively less information. Although total scores may therefore appear psychometrically sound, they may underrepresent the originality dimension of creative thinking, which raises important questions about how to balance reliability with construct validity in large-scale creativity assessments. The IRTree framework makes this imbalance visible by separating the two decision processes, whereas the GPCM assumption masks it within a single latent trait.

These discrepancies are further compounded by the sequential nature of the rating process. Because node 2 is only observed when a response has already passed node 1, fewer observations are available for estimating originality or diversity. This structural feature of the assessment reduces the precision of measurement at node 2 and contributes to larger standard errors for the associated trait estimates. In practice, this means that the information recovered from originality or diversity is systematically lower than that recovered from appropriateness, not only because raters find originality harder to judge, but also because the data design inherently provides fewer observations for this threshold. Together, these factors create an imbalance in reliability that favors appropriateness, further

reinforcing the concern that total scores may overstate the extent to which originality is being measured.

13.3. Suggestions for future research

Future research could examine the measurement invariance of the multidimensional rating structure, notably across countries. One of PISA's goals being to compare countries, ensuring that the underlying constructs of originality and appropriateness are interpreted and function similarly across cultural contexts is crucial. In the present study, we did use data from different countries, but we pooled data across countries, thus assuming invariance, an assumption we did not yet test with the IRTree models. Future studies could therefore extend our work by formally testing invariance across countries, as well as across languages or domains, to determine whether the multidimensional structure holds consistently. This would provide important information as to whether the scores of appropriateness, diversity and originality that are obtained through IRTree models can be used for comparisons across countries (as well as other demographic variables). Although this is important, we should however note that we found no particular reason to suspect differential item functioning, especially since the original items have been selected partly with the objective of ensuring invariance.

Although this special issue focuses on item-level analyses, future research could move beyond psychometric considerations to investigate whether known predictors and outcomes of creative thinking, as studied by PISA and researchers who reanalyzed the data, generalize to the more specific node-level attributes of appropriateness, diversity, and originality. Differences in how these traits are distributed might notably lead to alternative interpretations of national or group-level comparisons in creative thinking performance (e.g., some countries may hold a high or low creative thinking ranking largely due to the capacity, or lack thereof, of their students to generate appropriate responses).

Another possible avenue for improving measurement does not involve changing the model but rather revising the rating process itself (Myszkowski & Storme, 2025). From the perspective of the original Rasch philosophy, the goal is not necessarily to adapt the model to fit the data but to refine the measurement instrument until it conforms to the requirements of the model. Applied here, this would mean reconsidering the rating procedures or the scoring rubrics so that they align more closely with the assumption of unidimensionality. For example, providing clearer rating criteria, separating appropriateness and originality into independent rating steps, or using multiple raters could reduce the within-rating multidimensionality observed in the present study. In this view, simplifying (or decomposing) the rating process may ultimately yield more valid and interpretable scores than attempting to accommodate the multidimensionality through increasingly complex models.

A complementary avenue for addressing rater-related limitations is to reduce reliance on human judgment altogether by redesigning tasks or scoring procedures. One possibility is to rely more heavily on task formats in which appropriateness is less central or less ambiguous, such as divergent thinking tasks that emphasize fluency, flexibility, or novelty without explicit usefulness constraints. Another promising direction concerns automated scoring approaches based on large language models (LLMs). Recent work suggests that LLM-based systems can approximate, and in some cases outperform, human raters in scoring originality in divergent thinking tasks, while offering greater consistency and scalability (Organisciak et al., 2023). However, these approaches are not guaranteed to fully resolve the issues highlighted here. If trained on human ratings, LLMs may inherit the same ambiguities and construct conflations present in the original scoring process. Moreover, unless explicitly constrained, automated systems may blur distinctions between appropriateness and originality in ways that parallel human rater behavior. Future research should therefore examine whether automated scoring reproduces within-rating multidimensionality or instead imposes a different structure on creative performance, and whether such systems can be designed to align more closely with theoretically grounded distinctions between creative attributes.

14. Conclusion

One of the core objectives of item response theory is to formulate the relationship between a latent construct and its manifestation in item responses. Because the response process is inherently constrained by the structure of the rating scale, a sound model must be consistent with the way responses are elicited. In the case of creative thinking measurement in PISA, we questioned this consistency. Although the GPCM, currently used, is not entirely incompatible with the rating scale, it does not fully reflect the sequential nature of the judgment process described in the scoring instructions. Indeed, respondents are first asked to evaluate the appropriateness of a production and, only if it is deemed appropriate, to judge its originality. The present study provides empirical evidence that the two dimensions (originality/diversity and appropriateness) can indeed be distinguished in PISA responses, and that using a unidimensional model is a simplification that, while perhaps defensible from a pragmatic perspective, comes at a cost. In sum, our findings highlight the importance of aligning measurement models of ratings with the decision processes that they imply. We hope that these results will help researchers and decision-makers improve the scoring of PISA's creative thinking assessment in future work.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to assist in code development and text editing (grammar, syntax, proofreading, and style improvements). After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Nils Myszkowski: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Martin Storme:** Writing – review & editing, Methodology, Conceptualization.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997, 1983-20083-001.

Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (*Re*)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 233–240. https://doi.org/10.1037/aca0000233, 2019-20312-014.

Barbot, B., & Kaufman, J. C. (2025). PISA 2022 creative thinking assessment: Opportunities, challenges, and cautions. *The Journal of Creative Behavior, 59*(1), Article e70003. https://doi.org/10.1002/jocb.70003

Benedek, M., & Beaty, R. E. (2025). Envisioning the future of creative thinking assessment. *The Journal of Creative Behavior, 59*(2), Article e70036. https://doi.org/10.1002/jocb.70036

Celik, P., Storme, M., & Myszkowski, N. (2022). Individual differences in within-person variability in personality positively predict economic gains and satisfaction in negotiations. *Group Decision and Negotiation, 31*(3), 683–702. https://doi.org/10.1007/s10726-022-09778-x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software, 48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Cuesta-Hincapie, C., & Camargo Salamanca, S. L. (2025). Evaluating the alignment between PISA 2022 creative thinking scoring rubric and creativity theory: A validity framework perspective. *The Journal of Creative Behavior, 59*(4), Article e70065. https://doi.org/10.1002/jocb.70065

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM Family. *Journal of Statistical Software, 48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 35–40. https://doi.org/10.1037/a0038688

Dumas, D., Kim, Y., Carrera-Flores, M., Kagan, S., Acar, S., & Organisciak, P. (2025). Understanding elementary students' creativity as a trade-off between originality and task appropriateness: A Pareto optimization study. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000982

Engelhard Jr, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling, 60*(1), 33–52.

Forthmann, B. (2026). The PISA 2022 creative thinking assessment: A welcome opportunity to explore the mechanics of flexibility scoring. *The Journal of Creative Behavior, 60*(1), Article e70103. https://doi.org/10.1002/jocb.70103

Forthmann, B., Bürkner, P.-C., Szardenings, C., Benedek, M., & Holling, H. (2019). A new perspective on the multidimensionality of divergent thinking tasks. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00985

Hannan, E. J., & Quinn, B. G. (1979). The Determination of the Order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x

Hernández-Ramos, J., & Araya, R. (2025). Do school activities foster creative thinking? An analysis of PISA results. *Education Sciences, 15*(2), 133. https://doi.org/10.3390/educsci15020133

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Lang, J. W. B., Lievens, F., De Fruyt, F., Zettler, I., & Tackett, J. L. (2019). Assessing meaningful within-person variability in Likert-scale rated personality descriptions: An IRT tree approach. *Psychological Assessment, 31*(4), 474–487. https://doi.org/10.1037/pas0000600

Lievens, F., Lang, J. W. B., De Fruyt, F., Corstjens, J., Van de Vijver, M., & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. *Journal of Applied Psychology, 103*(7), 753–771. https://doi.org/10.1037/apl0000280

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Myszkowski, N. (2021). Development of the R library "jrt": Automated item response theory procedures for judgment data and their application with the consensual assessment technique. *Psychology of Aesthetics, Creativity, and the Arts, 15*(3), 426–438. https://doi.org/10.1037/aca0000287

Myszkowski, N. (2024). *Item response theory for creativity measurement.* Cambridge University Press.

Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence, 68*, 109–116. https://doi.org/10.1016/j.intell.2018.03.010

Myszkowski, N., & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts, Creativity Assessment: Pitfalls, Solutions, and Standards, 13*(2), 167–175. https://doi.org/10.1037/aca0000225, 2019-20312-006.

Myszkowski, N., & Storme, M. (2024). Modeling sequential dependencies in progressive matrices: An auto-regressive item response theory (AR-IRT) approach. *Journal of Intelligence, 12*(1). https://doi.org/10.3390/jintelligence12010007. Article 1.

Myszkowski, N., & Storme, M. (2025). One score, two components: Disentangling appropriateness and originality in PISA creative thinking judgments using generalized item response tree models. *The Journal of Creative Behavior, 59*(2), Article e70033. https://doi.org/10.1002/jocb.70033

Myszkowski, N., Storme, M., & Çelik, P. (2024). Unscrambling creativity measurement: An invitation to better formalize the domain generality-specificity of creativity with psychometric modeling. *Learning and Individual Differences, 109*, Article 102401. https://doi.org/10.1016/j.lindif.2023.102401

OECD. (2024a). *PISA 2022 results (Volume III): Creative minds, creative schools*. OECD. https://doi.org/10.1787/765ee8c2-en

OECD. (2024b). *PISA 2022 technical report*. OECD. https://doi.org/10.1787/01820d6d-en

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*, Article 101356. https://doi.org/10.1016/j.tsc.2023.101356

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*(1), 23–32. https://doi.org/10.1016/j.intell.2011.11.002

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences, 15*(5), 537–546. https://doi.org/10.1016/0191-8869(93)90337-3

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of creativity. *Creativity Research Journal, 24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360

Storme, M., Celik, P., & Myszkowski, N. (2020). A forgotten antecedent of career adaptability: A study on the predictive role of within-person variability in personality. *Personality and Individual Differences, 160*, 1–6. https://doi.org/10.1016/j.paid.2020.109936

Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *Journal of Intelligence, 7*(3), 1–17. https://doi.org/10.3390/jintelligence7030017

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*(1), 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Wind, S. A., & Engelhard Jr, G. (2016). Exploring rating quality in rater-mediated assessments using Mokken scale analysis. *Educational and Psychological Measurement, 76*(4), 685–706. https://doi.org/10.1177/0013164415604704