The Journal of Creative Behavior

ORIGINAL ARTICLE

One Score, Two Components: Disentangling Appropriateness and Originality in PISA Creative Thinking Judgments Using Generalized Item Response Tree Models

Nils Myszkowski[1] | Martin Storme[2]

[1]Department of Psychology, Pace University, New York, New York, USA | [2]IESEG School of Management, CNRS, UMR 9221—LEM—Lille Économie Management, Univ. Lille, Lille, France

Correspondence: Nils Myszkowski (nmyszkowski@pace.edu)

Received: 11 January 2025 | **Revised:** 24 March 2025 | **Accepted:** 9 May 2025

Keywords: creativity | item response theory | measurement | psychometrics

ABSTRACT

In the PISA 2022 creative thinking test, students provide a response to a prompt, which is then coded by human raters as no credit, partial credit, or full credit. Like many large-scale educational testing frameworks, PISA uses the generalized partial credit model (GPCM) as a response model for these ordinal ratings. In this paper, we show that the instructions given to the raters violate some assumptions of the GPCM as it is used: Raters are instructed to rate according to steps that involve multiple attributes (appropriateness and diversity/originality), with a different (set of) attribute(s) necessary to pass the different thresholds of the scoring scale. Instead of the GPCM, we propose multidimensional generalized item response tree models that allow us to account for the sequential nature of the ratings and to disentangle the attributes measured from the original scores. We discuss advantages, limitations, as well as recommendations for future research.

Even though creativity is still largely absent from traditional cognitive test batteries (Kaufman 2015), it is increasingly recognized as an important competency to assess. The inclusion of creativity measures in the Programme on International Student Assessment (PISA; OECD 2024a) is an important step in the recognition of creativity as a critical skill to study and nurture in educational contexts. However, programs like PISA, which are highly regarded and can have a strong impact on international benchmarking and policy-making (defining best practices, defining goals, evaluating national/regional educational effectiveness, evaluating reforms, etc.), rely greatly on their measurement device. In the current paper, we will focus on the mismatch between how creative thinking judgment scores are obtained and how they are used, as well as propose a possible solution.

The PISA 2022 results report (OECD 2024a) presents the methods used to collect item responses, as well as how item responses are subsequently used. To summarize, PISA measures creativity using open-ended item prompts, with human raters judging the examinees' responses. The raters use a 3-point ordinal scale, and although there are nuances regarding what is expected for each item, in general, the raters are asked to give no credit if the response is not appropriate, partial credit if the response is appropriate but not original/diverse/elaborate, or full credit if the response is both appropriate and original/diverse/elaborate. We note that this implies that the step between no credit and partial credit is therefore based on appropriateness, whereas the step between partial credit and full credit is based on whether the response is original/diverse/elaborate (assuming the response has been deemed appropriate). Thus, each item response is in fact rated on two dimensions, using a sequential process.

The resulting item scores are then scaled using a generalized partial credit model (GPCM; Muraki 1992), which is common in large-scale assessments. In this paper, we propose a discussion of why the GPCM fails to capture the sequential and

two-dimensional nature of the judgment process as instructed to the raters, leading to potentially biased scores, and we will propose to use a different psychometric model—more specifically, a generalized item response tree (IRTree; De Boeck and Partchev 2012; Jeon and De Boeck 2016) model—which better represents judgment scores obtained with such a process and allows us to disentangle the two underlying attributes. By pinpointing an important incongruence between the measurement device and its statistical treatment, we aim to lay the groundwork for future analyses of PISA 2022 data and for possible refinements in either or both aspects in future PISA efforts in the domain of creativity.

After a description of the rating process used to obtain PISA 2022 creativity item scores, we will discuss the GPCM and its shortcomings in this situation. Afterwards, we will describe how to use IRTree models to more accurately represent the ratings process. Finally, in the discussion, we will examine the implications of this proposition, its limitations, and suggestions for future research.

1 | The Measurement Device

Although we do not aim here to criticize or support the model of creative thinking used in the PISA test, it is important to describe it with enough detail to better understand the aim of the test developers, which allows to better identify whether the measure falls short of these aims. The PISA test identifies 3 facets to creative thinking, which are the ability to "(1) generate diverse ideas; (2) generate creative ideas; and (3) evaluate and improve ideas" (OECD 2024a, 221)—we later refer to these facets as, respectively, GDI, GCI and EII. Items are distributed across 4 domains (written and visual expression, social and scientific problem solving). Each item is open-ended, and the examinee's response is then coded into a score by a human rater.

1.1 | From Responses to Scores

To score responses, different rating instructions are used for the different facets. Importantly, the rating instructions consist of a step-by-step scoring process. For GDI, the rater shall judge whether the response is appropriate (step 1). Then, if appropriate, they shall judge if the responses are sufficiently different (step 2). The response is attributed no credit if it fails on either step, partial credit if it passes step 1 but partially fails step 2 (ideas are appropriate but not all different), and full credit if it succeeds in both steps. In other words, a response going from no credit to partial credit needs both appropriateness and diversity, whereas a response going from partial credit to full credit further requires only more diverse responses (see OECD 2024a, 224 for more details and a flowchart).

For the other two facets (GCI and EII), the process is slightly different. The rater shall first judge whether the response is appropriate (step 1). Then, if appropriate, they shall decide if the response is either original (step 2) or elaborated originally (step 3). The response is attributed no credit if it fails step 1, partial credit if it passes step 1 but fails both step 2 and 3, and full credit if it passes both step 1 and at least one of steps 2 and 3 (see OECD 2024a, 225 for more details and a flowchart). In other words, a response going from no credit to partial credit requires appropriateness, whereas a response going from partial credit to full credit requires originality (in its theme or elaboration).

1.2 | Discussing the Rater's Judgment Process

Some conclusions can be drawn from these two sets of instructions. First, for all facets, raters are instructed to judge responses on (essentially) two attributes: appropriateness and diversity/originality. In other words, each item score is, in fact, conceptually *multidimensional* (more specifically, bidimensional). Although this appears to stem from an effort to implement a definition of creativity as a combination of two attributes (originality and usefulness/effectiveness/appropriateness) that is very commonly adopted by creativity researchers (Runco and Jaeger 2012), and thus may be seen as a suitable decision, it also indicates a mismatch between the aims of the test and how scores are obtained. Indeed, each facet is referred to and scored as a single ability (e.g., the ability to generate diverse ideas), but is in fact measured using several attributes of a person's response combined into one score.

Second, for all facets, the ratings are on 3-point scales, which implies that the responses need to pass two thresholds to go from no credit to full credit. As previously discussed in Myszkowski (2021, 2024), nearly all rater-mediated assessments in creativity research rely on ordinal scales. Although the use of ordinal scales is a common practice and not necessarily problematic (even though continuous scales could provide more information), the issue here lies in a critical nuance: the two thresholds in this scale do *not* reflect the same underlying attribute, but rather two distinct attributes. More specifically, for the GDI facet, passing the first threshold depends on both appropriateness and diversity, whereas passing the second threshold depends only on diversity. For the GCI and EEI facets, passing the first threshold depends on appropriateness, whereas passing the second depends on originality. Therefore, not only is each response conceptually multidimensional, but the different steps of the scoring scale are manifestations of different attributes. We show our representation of the multidimensionality of the judgments for each threshold in Figure 1. In the next section, we will discuss how the resulting scores are dealt with in terms of psychometric modeling.

2 | The Generalized Partial Credit Model and Its Application in PISA

As we previously explained, the ratings obtained for the examinees' responses yield ordinal scores. The PISA creativity report (OECD 2024a), as well as the PISA technical report (OECD 2024b) explain that an item response theory (IRT) model, the GPCM (Muraki 1992), is used for such item scores. The GPCM is used during both the instrument construction process (item properties are notably judged using their item parameters) and with the main survey data. The end-user of the main survey data obtains plausible values, which correspond to observations generated from a (latent regression multiple-group) GPCM model previously estimated on the data. In other words,

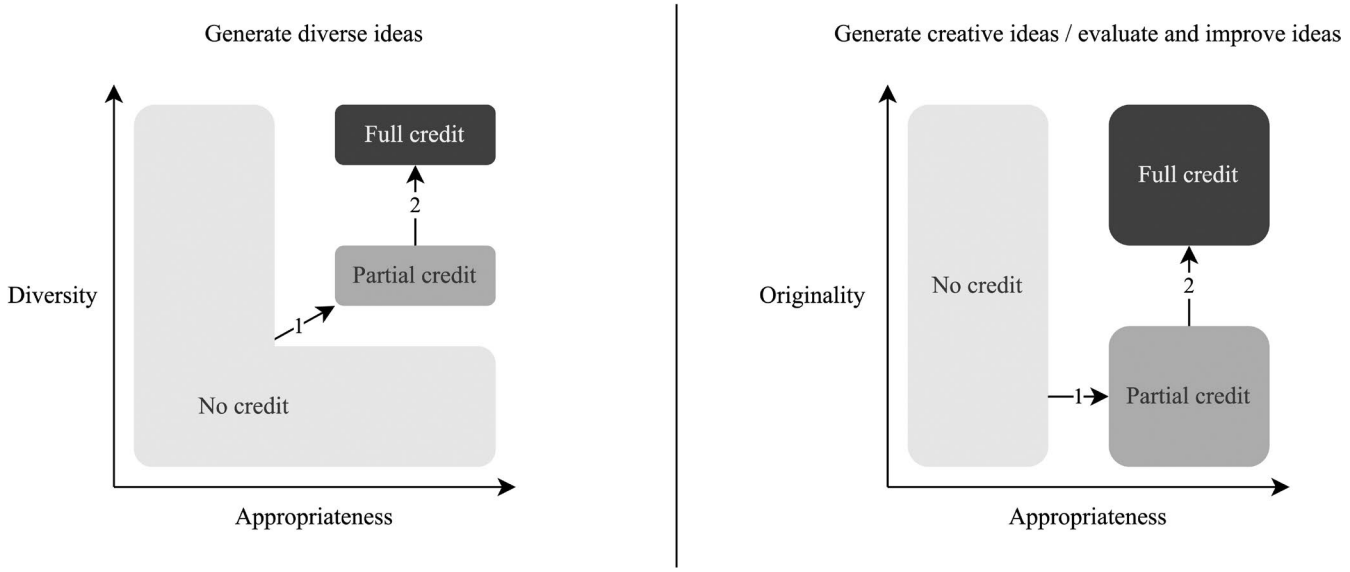

FIGURE 1 | Expectations of ratings as a function of the underlying response attributes.

all claims that are made about these responses, and therefore, all conclusions from PISA creative thinking scores, rely on the GPCM, which we will now (succinctly) describe.

2.1 | A Short in-Context Description of the Generalized Partial Credit Model

Although there are more thorough introductions to this model (e.g., De Ayala 2022), including in the domain of creativity (Myszkowski 2024), it is worth introducing the GPCM in this specific context—that is, ratings with three categories—to better grasp its assumptions. In the GPCM, each item has category thresholds, which represent the points on the latent trait scale where a respondent becomes more likely to receive a higher score. In this case, one threshold represents the point at which the probability of being scored 1 (partial credit) exceeds that of being scored 0, whereas another represents the transition from 1 to 2 (full credit). The probability that examinee i obtains the rating category k ($k \in \{0, 1, 2\}$) at item j, noted $P(X_{ij} = k|\,\theta_i)$, depends on the examinee's latent trait θ_i—assumably corresponding to the targeted competency, either generating diverse ideas (GDI), generating creative ideas (GCI), or evaluating and improving ideas (EII), depending on the item—item discrimination parameter α_j (for readers more familiar with the factor analysis tradition, this is akin to a factor loading), and category threshold parameter β_{jc}, such as:

$$P(X_{ij} = k|\,\theta_i) = \frac{\exp\left(\sum_{c=0}^{k} \alpha_j(\theta_i - \beta_{jc})\right)}{\sum_{m=0}^{2} \exp\left(\sum_{c=0}^{m} \alpha_j(\theta_i - \beta_{jc})\right)}, k \in \{0, 1, 2\}.$$

Although it is not necessary to fully grasp this formula to understand the shortcomings of GPCM that we later highlight, we may note that, on a conceptual level, the numerator represents the cumulative propensity of a respondent to progress through thresholds. For a threshold c, $\theta_i - \beta_{jc}$ represents the location of examinee i's level compared with the threshold's difficulty. The numerator is 1 when $k = 0$ (i.e., no thresholds are required to be passed). The denominator, which is the same for all categories, acts as a normalization constant, which ensures that the sum of probabilities is equal to 1. In general, θ_i is assumed to follow a Gaussian distribution of mean 0 and whose variance is fixed to 1 for identification. In PISA, the use of multiple group models implies that discrimination and threshold parameters are allowed to vary across groups (e.g., countries), and latent regression models implies that the mean of θ_i, instead of being 0 for all examinees, is conditional upon a series of covariates. Finally, it is worth noting that the GPCM assumes unidimensionality at the item level—that is, each item is modeled as a function of a single latent trait. Although this is a standard assumption in IRT, it raises questions about consistency with the multidimensional structure of the Creative Thinking framework, which distinguishes among GDI, GCI, and EII as distinct but related competencies.

2.2 | Problematic Assumptions

The GPCM is a very popular model for polytomous data that provides several flexibilities (items varying in discrimination and category threshold difficulties) that are useful in the context of creativity ratings data (Myszkowski 2021), and is implemented in many IRT packages (e.g., Chalmers 2012). Nevertheless, there are some important inconsistencies between its assumptions and the ratings process earlier described.

First, in the GPCM, a single latent trait predicts the passing of all thresholds. This is in sharp contrast with the rating instructions, in which, for each product, two attributes are involved—diversity and appropriateness for GDI, originality and appropriateness for GCI and EII. Thus, the rating process itself conceptually violates an assumption of the GPCM: The GPCM—as it is used—is unidimensional, but the ratings are, by design, bidimensional. This first issue may be considered minimal, had the raters been trained so that they would use a consistent "blend" of the two rating criteria throughout all responses. For example, had the raters been instructed to rely for 30% on appropriateness and

70% on diversity/originality, then one could advance that the unidimensionality assumptions of the GPCM still hold, and we would simply define the latent trait θ_i as this 30/70 blend. Nevertheless, it was not how the raters were instructed to use both criteria (in all fairness, it may not even be realistic to ask this kind of mental gymnastics from human raters), and it keeps the two underlying attributes entangled (which is ignored by PISA but may be useful to creativity researchers).

A second (and related) issue is that not only are the raters instructed to use several attributes in their ratings, but, as we explained, the attributes rated differ by threshold. In other words, for a single item, the passing of each threshold does not involve the same underlying attribute. This is a result of the sequential nature of the instructions provided to the raters, who are asked to judge on a set of attributes (diversity and appropriateness for GDI, only appropriateness for GCI and EII) for the first threshold, and, if passed, to judge on a different set of attributes for the second threshold (diversity for GDI, originality for GCI and EII). In other words, the GPCM ignores that different thresholds rely on different criteria and thus measure different attributes.

Finally, the GPCM does not represent the sequential nature of the response categories. Indeed, the GPCM provides the probability of rating each category directly but does not model the process of moving through intermediate categories. In other words, in the GPCM, there is no decomposition of the rating process into sequential steps. This is also inconsistent with the instructions given to the raters, which are explicitly a sequence of steps. The lack of sequential structure of the GPCM therefore fails to account for the sequential nature of the instructions provided to the raters.

These concerns echo a broader body of literature questioning the suitability of the GPCM for large-scale assessments. Verhelst and Verstralen (2008) argue that the GPCM is poorly suited to scoring processes that reflect stepwise or conditional judgments, as its thresholds are not interpretable as sequential steps. Similarly, Bürkner et al. (2019) demonstrate that under optimal designs, the model may behave as if it were dichotomous, reducing the informational value of intermediate scoring categories. Finally, Dai et al. (2021) show that the GPCM may perform poorly under common large-scale assessment conditions—such as short scales and substantial missingness—casting doubt on its robustness in applied survey settings like PISA.

2.3 | Possible Consequences and a Missed Opportunity

There are several consequences to the mismatch between the sequential nature of the rating guidelines and the use of the GPCM. On a practical level, there are some possible misinterpretations as to what is being assessed. For example, for GCI and EII, a person receiving no credit for an item has not been measured on originality (and thus, not on creativity as a whole): They have been measured on appropriateness only. Only persons receiving at least partial credit on that item have been measured on both appropriateness and originality. For GDI, because the first step involves both diversity and appropriateness, the problem is more nuanced, but still present: A person who received no credit for an item has been *less accurately* measured on diversity than a person receiving full credit. Consequently, one may conclude that a person (and more generally, a group) has low creativity, where in fact they only (or mainly) have low (or not enough to pass the threshold) capacity for appropriateness. These examples illustrate how the use of the GPCM, in ignoring the sequential and bidimensional nature of the ratings, could be a threat to the validity of the claims made from PISA creativity data.

Beyond being a threat to validity, these examples imply that persons, because of their level on the attribute (or set of attributes) measured for the first threshold, are measured differently (e.g., originality is not measured for persons who did not pass a threshold for appropriateness, whereas it was for others). This can be framed as a problem of measurement invariance: A person or group of people less able to provide responses rated as appropriate is not given credit for the originality (or as much credit for the diversity) of their responses. This could lead to potential biases: For example, the relation between verbal abilities and creativity could be biased if individuals or groups with lower verbal comprehension or expression abilities are less able to provide appropriate responses because they have a less accurate understanding of the instructions but are in fact capable of original answers.

Finally, for researchers, if raters have been instructed to judge not creativity but two distinct attributes, one would expect to have the same degree of detail in the data. In other words, there is a missed opportunity from using a response model that entangles the two attributes being rated (appropriateness and/or diversity) for each person. In the next section, we are going to provide a possible alternative modeling strategy that could address these limitations.

3 | Using Item Response Tree Models in PISA Creativity Ratings

3.1 | A Short Introduction to IRTree Models

Item response tree (IRTree; De Boeck and Partchev 2012; Partchev and De Boeck 2012) and generalized IRTree (DiTrapani et al. 2016; Jeon and De Boeck 2016) models are a class of IRT models for categorical (ordinal or not) responses. Essentially combining IRT modeling with decision trees, (generalized) IRTree models are appropriate when the response process can be conceptualized as the outcome of a sequence of binary decisions—much like the situation at hand. Each binary decision is generally referred to as a pseudo-item (or node). A pseudo-item refers to a latent binary decision that contributes to the observed response score. In the context of PISA, for example, one pseudo-item might represent the decision to award any credit at all, whereas another represents the decision to award full credit (conditional on the first). IRTree models have been used in a variety of situations, such as the understanding of faking (Lee et al. 2022), extreme/midpoint response styles (Ames and Myers 2021) or item skipping (Storme et al. 2024) in personality questionnaires. More closely related to the present topic, IRTree models have been used to explore the

dimensionality of snapshot ratings of divergent thinking test responses (Forthmann et al. 2019).

3.2 | The Proposed Tree Diagram

To understand how IRTree models work, it is probably best to look at a one such model using a tree diagram, which presents the decisions as nodes. In Figure 2, we show the diagram that represents our proposed model and how to restructure the data accordingly. Following the rating instructions, we propose that a first pseudo-item would represent the decision to attribute either no credit or some credit, whereas a second pseudo-item, reached only if the rater decides to attribute some credit, would represent the decision to attribute either partial or full credit.

3.3 | Model Formulation for Each Pseudo-Item

Provided the tree structure, we may now focus on how each pseudo-item response is modeled. As we mentioned before, for a given item, several attributes are measured within a single score. Therefore, we propose a model that is multidimensional—more specifically bidimensional—as it involves two latent attributes. For each pseudo-item, different latent attributes are involved. For clarity, we refer to person i's latent (capacity for) appropriateness as $\theta_{\text{Appropriateness},i}$, diversity as $\theta_{\text{Diversity},i}$, and originality as $\theta_{\text{Originality},i}$. When we write θ_i without a subscript, we refer generically to the person's latent trait(s) involved in the pseudo-item, depending on the context. To each item j are attributed parameters b_j and b'_j to represent pseudo-item difficulties, as well as a discrimination parameters for all latent traits involved in passing each pseudo-item, which are noted as $a_{\text{Trait},j}$ and $a'_{\text{Trait},j}$ ("Trait" is appropriateness, diversity or originality, depending on the pseudo-item). Depending on the pseudo-item, a unidimensional or bidimensional compensatory 2-parameter model is used, depending on whether one or two latent attributes are involved. Finally, a logit or probit link function g is used, yielding models commonly referred to as logistic or normal ogive models, respectively.

3.3.1 | GDI Items

The next step consists of combining our knowledge of the rating instructions with the tree structure and the IRT models. In the case of GDI items, the scoring rubric specifies that a response must demonstrate both appropriateness and diversity to receive partial credit. In other words, moving from no credit to partial credit depends on the presence of both components, which we reflect by modeling the first pseudo-item as bidimensional. We thus obtain an IRTree model where a bidimensional compensatory model predicts the probability to pass the first pseudo-item $P(X_{ij} > 0 | \theta_i)$ as:

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

whereas the probability to pass the second pseudo-item, conditional upon having passed the first pseudo-item $P(X_{ij} = 2 | X_{ij} > 0, \theta_i)$, is given by a unidimensional model:

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

It is important to note here that the $a_{\text{Diversity},j}$ and $a'_{\text{Diversity},j}$ differ, which relates to diversity having a different relevance to the passing of each pseudo-item. Another important note here is that the first pseudo-item model is formulated using a compensatory approach, which implies that the two attributes are allowed to compensate one another (i.e., a sufficiently appropriate response can offset the fact that nondiverse ideas are provided, and vice versa). This does not perfectly represent the instructions provided to the rater (see OECD 2024a, 224), where the response must meet *both* criteria to obtain partial credit. Therefore, we could suggest a conjunctive (noncompensatory) model instead:

$$P(X_{ij} > 0 | \theta_i) = g^{-1}\left[\min\left(a_{\text{Appropriateness},j}\theta_{\text{Appropriateness},i}, a_{\text{Diversity},j}\theta_{\text{Diversity},i}\right) - b_j\right].$$

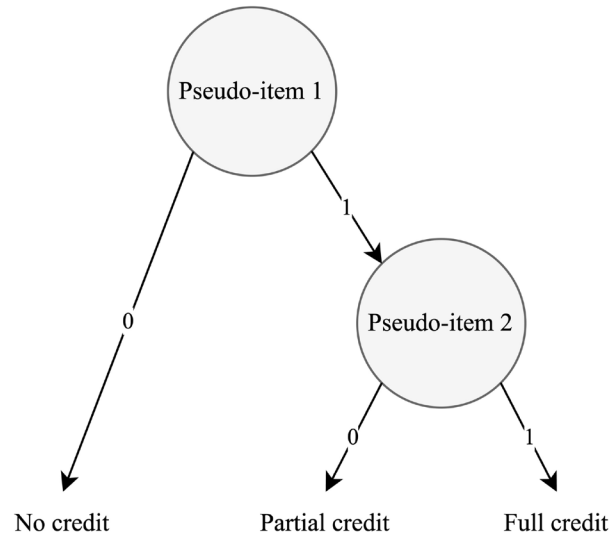

FIGURE 2 | Response tree and corresponding data structure.

Item response	X_{ij}	Pseudo-item 1	Pseudo-item 2
No credit	0	0	(Missing)
Partial credit	1	1	0
Full credit	2	1	1

3.3.2 | GCI and EEI Items

For the GCI and EEI, items, based on the instructions given to the raters, each pseudo-item is assumed unidimensional, and we obtain an IRTree model where a unidimensional model predicts the probability to pass the first pseudo-item $P(X_{ij} > 0 | \theta_i)$ as:

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

Although the probability $P(X_{ij} = 2 | X_{ij} > 0, \theta_i)$ to pass the second pseudo-item, conditional upon having passed the first pseudo-item, is also given by a unidimensional model:

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

These differences in model specification across item types reflect the structure of the PISA Creative Thinking competency framework, which distinguishes between types of creative tasks and their associated response criteria. By allowing different combinations of latent attributes and model structures for GDI, GCI, and EII items, the IRTree approach provides a way to obtain separate estimates of appropriateness and originality/diversity that respect the domain-specific emphases of each item group.

3.3.3 | General Considerations

For the GDI, GCI, and EEI, because we provided conditional probabilities only, we may reformulate the marginal probabilities for partial credit and full credit, respectively as:

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

and:

$$P(X_{ij} = 2 | \theta_i) = P(X_{ij} > 0 | \theta_i) \cdot P(X_{ij} = 2 | X_{ij} > 0, \theta_i).$$

Because of the presence of $P(X_{ij} > 0 | \theta_i)$ in these last two equations, it is important to note that, ultimately, probabilities of all response categories are affected by the latent attribute(s) measured in the first pseudo-item. This is because, even if the second pseudo-item measures a different attribute, it must be reached in the first place, which itself depends on the attributes measured in the first pseudo-item.

Finally, as done using the GPCM in PISA, generalized IRTree models can be extended to multiple-group and latent regression frameworks (e.g., Plieninger 2021; Storme et al. 2024). More broadly, IRTree models can accommodate several features typical of large-scale assessments, including missing data, clustered sampling structures (e.g., students nested within schools or countries), and the use of respondent weights. They have been implemented in multilevel and multiple-group contexts, allowing researchers to account for hierarchical data and to test for measurement invariance across groups. These models can also be estimated using standard software (e.g., Mplus) that supports full information maximum likelihood or Bayesian estimation techniques for handling missingness. Although respondent weights in assessments like PISA are typically applied during the plausible value and secondary analysis stages, some modeling frameworks allow for their inclusion at the estimation stage when needed. In sum, although implementing and calibrating our proposed response model may require considerable work, we do not see any theoretical or practical limitations that would make it unfeasible for the PISA scoring process.

4 | Discussion

In this paper, we propose using generalized item response tree models instead of generalized partial credit models, as they better reflect the creativity rating process as defined in the PISA scoring guidelines. This, in contrast with the GPCM, makes these models more conceptually sound, as they better depict the item score generation mechanism. After having represented the ratings as a tree structure, we provided response models for the pseudo-items.

4.1 | Opportunities

Beyond providing a superior representation of the measurement situation, IRTree models would also provide important advantages. A first obvious advantage is that, from the creativity ratings, the IRTree models proposed involve two latent variables (for each item), which are expected to represent appropriateness and diversity/originality. Consequently, estimating the models could allow obtaining person estimates (point estimates, standard errors, posterior distributions, plausible values, etc.) for the different person attributes separately. In other words, from (apparently unidimensional) creativity ratings, one could disentangle appropriateness and diversity/originality.

This disentanglement opens avenues for future research. First, one could study the correlation between these two attributes, how variable this relation is across groups, or how domain-specific it is. For example, we could speculate that the visual expression domain provides more freedom for originality than other domains (like scientific problem solving), and thus, appropriateness and diversity/originality might be less related in it. Second, one could study how appropriateness and diversity/originality may be differently predicted by various factors (demographic, educational, etc.). We could speculate, for example, that appropriateness, which relies greatly on understanding the instructions, is more strongly related to verbal ability than originality/diversity is. Third, this framework allows for tests of measurement invariance—for instance, whether students with lower verbal or cultural familiarity may be disproportionately rated as inappropriate, and thus less likely to be scored for originality. Fourth, exploring these empirical questions may help explain broader findings such as the unexpectedly high convergence between PISA creative thinking scores and traditional academic performance. We intend to explore these questions in future work building on the present model. Finally, one could also explore whether focusing on originality alone—independent of appropriateness—yields different patterns of group differences or

predictive validity than those observed under models that aggregate both. This could help clarify which aspects of creative thinking are being emphasized or obscured by current scoring practices.

4.2 | Limitations

It is important to note that we do not claim that originality and appropriateness are inherently independent psychological constructs, nor do we offer a substantive argument for modeling them as such. In other words, this work neither contributes to nor challenges existing research on the relationship between originality and appropriateness (e.g., Diedrich et al. 2015). Rather, our argument is that the PISA rating procedure operationalizes them as sequential and distinct, but does not model them accordingly. This implies a conceptual discrepancy between the rating methodology and the psychometric modeling strategy. Thus, our critique is not rooted in a theoretical stance on the separability of creativity's components, but in a methodological mismatch between the structure of the ratings and the assumptions of the model applied to them. Given that the data have already been collected using this rating structure, the only way to reconcile this mismatch is adapting the psychometric model to better reflect the scoring logic.

A first limitation to note is that the implementation of these models may prove challenging. Although the literature on generalized IRTree models and its multiple applications is expanding, the models we propose here are ultimately, after data restructuring, multidimensional IRT models, which implies that they are more heavily parametrized than unidimensional partial credit models. This also means that only packages capable of multidimensional IRT modeling would be suitable, and that there is more risk (in comparison with estimating GPCMs) that the models would fail to converge, at least without guidance (e.g., informative priors, reasonable starting values) from the researcher.

In addition, it is possible that, even though they more accurately represent the ratings instructions, the models we propose may not necessarily fit the data (much) better than a GPCM. This would be especially true if, despite the sequential nature of the rating instructions, the raters only loosely separated appropriateness and diversity/originality. In other words, we perhaps proposed an approach that follows the rating instructions more closely than the raters did. It is also worth emphasizing that our critique is not directed at the PISA rating guidelines themselves, nor do we suggest that raters apply them subjectively. On the contrary, the guidelines are clearly defined, and the technical documentation outlines substantial efforts to train raters and ensure consistency. It is possible that the sequential nature of the scoring procedure—requiring appropriateness to be evaluated before originality—is designed precisely to enhance objectivity and inter-rater agreement. Our argument, then, is not with the quality or intent of the scoring process, but with the need for psychometric models that better reflect its structure.

Another potential limitation is the challenge of combining originality and appropriateness into a single creativity score. PISA may indeed need a global creativity score, whereas the IRTree models we have described produce two scores, and it is unclear how they should be combined (additively, multiplicatively or otherwise). This question highlights a wider conceptual gap: creativity has no formal definition specifying how originality and appropriateness should be integrated into one creativity score. Rather than seeing this as an argument against using an IRTree model, we see it as an opportunity to develop a clearer framework for measuring creativity by promoting a more nuanced understanding of the concept.

5 | Conclusion

As much as the authors of this paper appreciate the elegance of response models meticulously tailored to intricate situations, maybe a more sensible conclusion here is that the ratings process itself should be simplified in future PISA iterations. Instead of contorting the model to gracefully accommodate convoluted rating instructions, we would suggest making the rating process straightforward by providing a more minimal definition of creativity—or none—to apply across all thresholds, or by separately measuring more specific attributes. The possible drop in reliability could be counteracted by more points on the rating scale and/or more raters and/or more items. As pointed out previously (Myszkowski and Storme 2019; Storme et al. 2014), hair-splitting instructions for raters can be a double-edged sword: While maximizing interrater agreement, raters may reliably miss the target—are we really measuring creativity, or merely the researchers' idiosyncratic standards?

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Ames, A. J., and A. J. Myers. 2021. "Explaining Variability in Response Style Traits: A Covariate-Adjusted IRTree." *Educational and Psychological Measurement* 81, no. 4: 756–780. https://doi.org/10.1177/0013164420969780.

Bürkner, P.-C., R. Schwabe, and H. Holling. 2019. "Optimal Designs for the Generalized Partial Credit Model." *British Journal of Mathematical and Statistical Psychology* 72, no. 2: 271–293. https://doi.org/10.1111/bmsp.12148.

Chalmers, R. P. 2012. "Mirt: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software* 48, no. 1: 1–29. https://doi.org/10.18637/jss.v048.i06.

Dai, S., T. T. Vo, O. J. Kehinde, et al. 2021. "Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data." *Frontiers in Education* 6: 1–18. https://doi.org/10.3389/feduc.2021.721963.

De Ayala, R. J. 2022. *The Theory and Practice of Item Response Theory*. Second ed. Guilford Press.

De Boeck, P., and I. Partchev. 2012. "IRTrees: Tree-Based Item Response Models of the GLMM Family." *Journal of Statistical Software* 48, no. 1: 1–28. https://doi.org/10.18637/jss.v048.c01.

Diedrich, J., M. Benedek, E. Jauk, and A. C. Neubauer. 2015. "Are Creative Ideas Novel and Useful?" *Psychology of Aesthetics, Creativity, and the Arts* 9, no. 1: 35–40. https://doi.org/10.1037/a0038688.

DiTrapani, J., M. Jeon, P. De Boeck, and I. Partchev. 2016. "Attempting to Differentiate Fast and Slow Intelligence: Using Generalized Item Response Trees to Examine the Role of Speed on Intelligence Tests." *Intelligence* 56: 82–92. https://doi.org/10.1016/j.intell.2016.02.012.

Forthmann, B., P.-C. Bürkner, C. Szardenings, M. Benedek, and H. Holling. 2019. "A New Perspective on the Multidimensionality of Divergent Thinking Tasks." *Frontiers in Psychology* 10: 985. https://doi.org/10.3389/fpsyg.2019.00985.

Jeon, M., and P. De Boeck. 2016. "A Generalized Item Response Tree Model for Psychological Assessments." *Behavior Research Methods* 48, no. 3: 1070–1085. https://doi.org/10.3758/s13428-015-0631-y.

Kaufman, J. C. 2015. "Why Creativity Isn't in IQ Tests, Why It Matters, and Why It Won't Change Anytime Soon Probably." *Journal of Intelligence* 3, no. 3: 59–72. https://doi.org/10.3390/jintelligence3030059.

Lee, P., S.-H. Joo, and Z. Jia. 2022. "Opening the Black Box of the Response Process to Personality Faking: An Application of Item Response Tree Models." *Journal of Business and Psychology* 37, no. 6: 1199–1214. https://doi.org/10.1007/s10869-022-09791-6.

Muraki, E. 1992. "A Generalized Partial Credit Model: Application of an EM Algorithm." *ETS Research Report Series* 1992, no. 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, no. 3: 426–438. https://doi.org/10.1037/aca0000287.

Myszkowski, N. 2024. *Item Response Theory for Creativity Measurement*. Cambridge University Press.

Myszkowski, N., and M. Storme. 2019. "Judge Response Theory? A Call to Upgrade Our Psychometrical Account of Creativity Judgments." *Psychology of Aesthetics, Creativity, and the Arts* 13, no. 2: 167–175. https://doi.org/10.1037/aca0000225.

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.

Partchev, I., and P. De Boeck. 2012. "Can Fast and Slow Intelligence Be Differentiated?" *Intelligence* 40, no. 1: 23–32. https://doi.org/10.1016/j.intell.2011.11.002.

Plieninger, H. 2021. "Developing and Applying IR-Tree Models: Guidelines, Caveats, and an Extension to Multiple Groups." *Organizational Research Methods* 24, no. 3: 654–670. https://doi.org/10.1177/1094428120911096.

Runco, M. A., and G. J. Jaeger. 2012. "The Standard Definition of Creativity." *Creativity Research Journal* 24, no. 1: 92–96. https://doi.org/10.1080/10400419.2012.650092.

Storme, M., N. Myszkowski, P. Çelik, and T. Lubart. 2014. "Learning to Judge Creativity: The Underlying Mechanisms in Creativity Training for Non-Expert Judges." *Learning and Individual Differences* 32: 19–25. https://doi.org/10.1016/j.lindif.2014.03.002.

Storme, M., N. Myszkowski, E. Kubiak, and S. Baron. 2024. "Personality Traits Leading Respondents to Refuse to Answer a Forced-Choice Personality Item: An Item Response Tree (IRTree) Model." *Psychiatry* 6, no. 1: 100–110. https://doi.org/10.3390/psych6010006.

Verhelst, N. D., and H. H. F. M. Verstralen. 2008. "Some Considerations on the Partial Credit Model." *Psicológica: International Journal of Methodology and Experimental Psychology* 29, no. 2: 229–254.