PISA 2022 Technical Report

18 PISA 2022 Innovative Domain Test Design and Test Development

Introduction

This chapter describes the assessment design framework for the PISA 2022 innovative domain of creative thinking as well as the processes used by the PISA Core B contractor, ACT, the PISA Secretariat, and the international test development team to develop the creative thinking assessment for the PISA 2022 cycle.

Activities undertaken in the context of the innovative domain test design and development included the following:

- The creation of a Creative Thinking Expert Group (CTEG) to guide the assessment framework, test design and test development;
- The development of a creative thinking assessment framework;
- The assessment design and development;
- A series of small-scale validation studies;
- Field trial activities; and
- The main survey administration.

The role of the Creative Thinking Expert Group (CTEG) in the framework and item development

As the contractor for the creative thinking instrument development, Core B was responsible for working with the creative thinking expert group (CTEG) and the PISA Secretariat. Work focused on understanding the CTEG and PISA Secretariat's vision for the creative thinking assessment framework as well as the range and types of items to be developed for the test and questionnaire instruments. The PISA Secretariat and CTEG members began work on the framework in September 2017 finalised the framework in September 2022. Core B's work with the PISA Secretariat and CTEG began in February 2018 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in each domain as defined in the framework;
- reviewing and understanding the proposed assessment design in order to define the number and types of items that were needed for each of the domains;
- defining the testing functionalities that would be desirable to develop for measuring the construct and that would be feasible to implement in the context of the PISA 2022 administration.

Work with the CTEG continued beyond the initial meeting in February 2018 through the entire phase of instrument development and during data analysis. CTEG members played an important role in reviewing

and providing feedback on the assessment tasks as they were developed, providing input into the analysis of the data from multiple small scale validation exercises and the field trial(s), approving the set of items for the main survey administration, and working with instrument development and data analysis staff to develop the described scales and performance level descriptors used for reporting the PISA 2022 creative thinking results.

PISA 2022 creative thinking assessment framework

The PISA Secretariat, together with guidance from the CTEG, developed the PISA 2022 creative thinking assessment framework. The PISA 2022 creative thinking assessment focused on the creative thinking processes that can be reasonably expected from 15-year-old students around the world. It does not aim to single out exceptionally creative individuals but rather to describe the extent to which students can think creatively when searching for and expressing ideas, and to describe how this capacity is related to teaching approaches, school activities and other features of education systems.

The main objective of PISA is to provide internationally comparable data on students' competencies that have clear implications for education policies and pedagogies. In the context of the PISA 2022 assessment, the creative thinking processes in question therefore need to be malleable through education; the different enablers of these thinking processes in the classroom context need to be clearly identified and related to performance in the assessment; the content domains covered in the assessment need to be closely related to subjects taught in common compulsory schooling; and the test tasks should resemble real activities in which students engage, both inside and outside of their classroom, so that the test has some predictive validity of creative achievement and progress in school and beyond.

While closely related to the broader construct of creativity, the PISA 2022 assessment focuses on creative thinking understood as the cognitive processes that are required to engage in creative work. Creative thinking was considered a more appropriate construct to assess in the context of PISA as it is a malleable individual capacity that can be developed through practice, and it refers more to specific cognitive processes than to the subjective quality of an output.

PISA defines creative thinking as:

> *The competence to engage productively in the generation, evaluation, and improvement of ideas, that can result in original and effective solutions, advances in knowledge, and impactful expressions of imagination. (OECD, 2023[1])*

The PISA definition builds on definitions of creativity and creative thinking found in the literature, following a comprehensive review, and it was developed with the guidance of a wider interdisciplinary group of experts in the field (the CTEG). The definition is aligned with the cognitive processes and outcomes associated with "little-c" creativity – in other words, it reflects the types of creative thinking that 15-year-old students around the world can reasonably demonstrate in everyday contexts. It emphasises that students need to learn to engage productively in generating ideas, reflecting upon ideas by valuing their relevance and novelty, and iterating upon ideas before reaching a satisfactory outcome. This definition of creative thinking applies to learning contexts that require imagination and the expression of one's inner world, such as creative writing or the arts, as well as contexts in which generating ideas is functional to the investigation of problems or phenomena.

The competency model

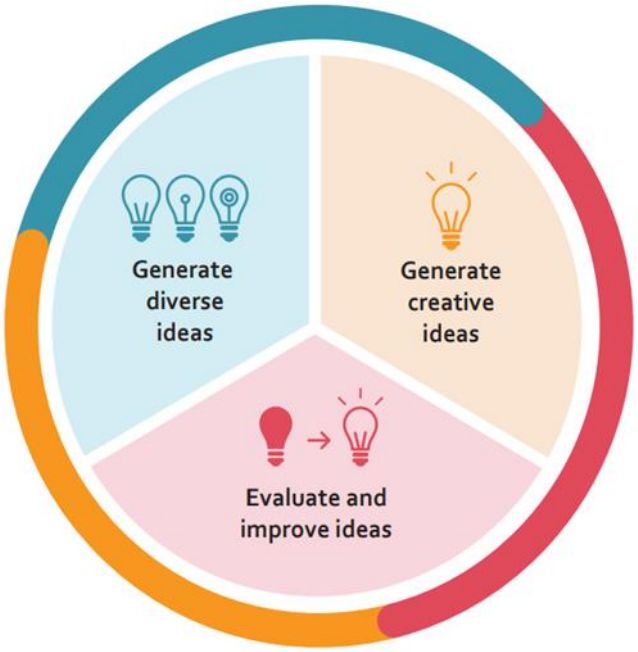
Three cognitive facets support creative thinking and constitute the competency model for the PISA 2022 creative thinking assessment (see Figure 18.1). These three facets are:

- generate diverse ideas;

- generate creative ideas; and
- evaluate and improve ideas.

These three facets reflect the PISA definition of creative thinking and incorporate both divergent cognitive processes (the ability to generate diverse ideas and to generate creative ideas) and convergent cognitive processes (the ability to evaluate other people's ideas and identify improvements to those ideas). “Ideas” in the context of the PISA assessment can take many forms, and the test units provide a meaningful context and sufficiently open tasks in which students can demonstrate their capacity to produce different ideas and think outside of the box.

Figure 18.1. The PISA 2022 competency model for creative thinking



Generate diverse ideas

Typically, attempts to measure creative thinking have focused on the number of ideas that individuals are able to generate – often referred to as ideational fluency. Going one step further is ideational flexibility, or the capacity to generate ideas that are different to each other. When it comes to measuring the quality of ideas that an individual generates, some researchers have argued that fundamentally different ideas should be weighted more than similar ideas (Guilford, 1956[2]). The facet ‘generate diverse ideas’ of the competency model encompasses these notions and refers to a student’s capacity to think flexibly by generating multiple distinct ideas. Test items for this facet present students with a stimulus and ask them to generate two or three appropriate ideas in response that are as different as possible from one another.

Generate creative ideas

The literature generally agrees that creative ideas and outputs are defined as being both novel and useful (Plucker, Beghetto and Dow, 2004[3]). Expecting 15-year-olds around the world to generate ideas that are completely unique or novel is clearly neither a feasible nor appropriate approach for the PISA assessment. Instead, originality represents a useful concept as a proxy for measuring the novelty of ideas. Defined by Guildford (1950[4]) as “statistical infrequency”, originality encompasses the qualities of newness, remoteness, novelty or unusualness, and generally refers to deviance from patterns that are observed

within the population at hand. In the PISA assessment context, originality is therefore a relative measure established with respect to the responses of other students who complete the same task.

The facet 'generate creative ideas' focuses on a student's capacity to generate appropriate and original ideas. This dual criterion ensures the measurement of creative ideas – ideas that are both original *and* of use – rather than ideas that make random associations that are original yet not meaningful. Test items for this facet present students with a stimulus and ask them to develop one original idea in response.

Evaluate and improve ideas

Evaluative cognitive processes help to identify and remediate deficiencies in initial ideas as well as ensure that ideas or solutions are appropriate, adequate, efficient and effective (Cropley, 2006[5]). They often lead to further iterations of idea generation or the reshaping of initial ideas to improve a creative outcome. Evaluation and iteration are thus at the heart of the creative thinking process. The facet 'evaluate and improve ideas' focuses on a student's capacity to evaluate limitations in ideas and improve their originality. To reduce problems of dependency across items in the test, students are not asked to iterate upon their own ideas but rather to modify a provided "idea". Test items for this facet thus present students with a given scenario and idea and ask them to suggest an original improvement in response, defined as a change that preserves the essence of the initial idea but that adds or incorporates original elements.

Task contexts: domains of creative thinking

The literature suggests that the larger the number of domains included in an assessment of creative thinking, the better the coverage of the construct given that creative thinking draws on both domain-general and domain-specific resources. The choice of which domains to include in the PISA test was thus a central design question. Given the age and diversity of PISA test takers (15 years-old in over 60 countries), and the fact that domain knowledge is an important enabler of creative thinking, the domain contexts included in the assessment needed to be familiar and accessible to most students around the world, be relevant to schooling, reflect realistic manifestations of creative thinking that 15-year-olds could achieve in a constrained test context, and represent a sufficiently diverse coverage of different types of "everyday" creative thinking as reflected in the literature. Further practical constraints, including the available testing time (a maximum of one hour for the creative thinking test) and testing technology, also informed design choices.

Taking these main constraints into account, the PISA test of creative thinking includes tasks situated within four distinct domain contexts:

- written expression;
- visual expression;
- social problem solving; and
- scientific problem solving.

In the PISA test, the written and visual expression domains involve communicating one's imagination to others, and creative work in these domains tends to be characterised by originality, aesthetics, imagination, and affective intent and impact. In contrast, the social and scientific problem-solving domains involve investigating and solving open problems. They draw on a more functional employment of creative thinking that is a means to a better end, and creative work in these domains is characterised by ideas or solutions that are original, innovative, effective and efficient.

The inclusion of tasks situated in several domain contexts will allow the PISA 2022 creative thinking test to provide information about students' strengths and weaknesses in creative thinking across countries. The test items were distributed across the three facets and four domain contexts to allow for a range of

opportunities for students to engage and express creative thinking. Table 18.1 sets out the distribution of items across facets and domains for the field trial(s) and the main survey administration.

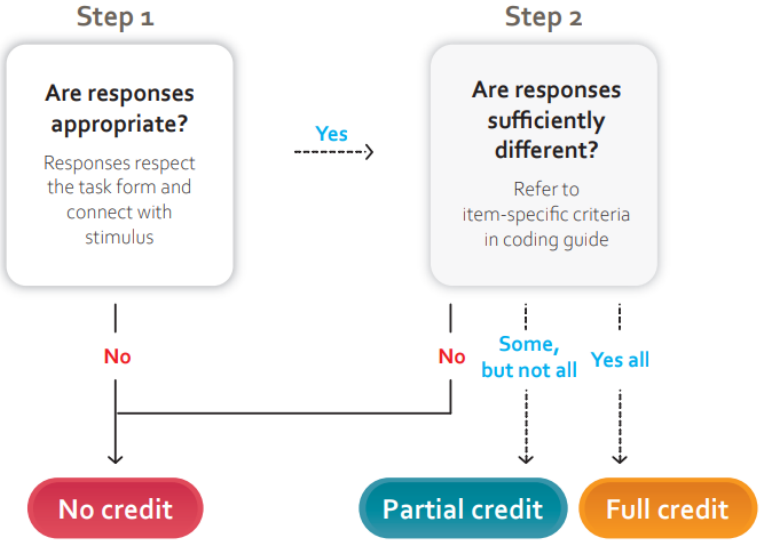
Coding and scoring processes

Every task in the PISA creative thinking test is open-ended and scoring student responses relies on human judgement following detailed scoring rubrics and well-defined coding procedures. All items corresponding to the same facet of the competency model (i.e. 'generate diverse ideas', 'generate creative ideas' and 'evaluate and improve ideas') apply the same general coding procedure. However, as the form of response varies by domain and task (e.g. a title, a solution, a design, etc.), so do the item-specific criteria for evaluating whether an idea is different or original. ACT developed detailed coding guides to describe the item-specific criteria for each item and provide annotated example responses to help human coders score consistently.

Scoring of 'generate diverse ideas' items

All items corresponding to the 'generate diverse ideas' facet of the competency model require students to provide two or three responses. The general coding procedure for these items involves two steps, as summarised in Figure 18.2. First, coders must determine whether responses are appropriate**.** Appropriate in the context of the creative thinking assessment means that students' responses respect the required form and connect (explicitly or implicitly) to the task stimulus. Second, coders must determine whether responses are sufficiently different from one another based on item-specific criteria described in the coding guide**.**

Figure 18.2. General coding process for 'generate diverse ideas' items



The item-specific criteria are as objective and inclusive as possible of the range of different potential responses. For example, for a written expression item, sufficiently different ideas must use words that convey a different meaning (i.e. are not synonyms). For items in the problem-solving domains, the coding guides list pre-defined response categories to help coders distinguish between similar and different ideas. The coding guides provide detailed example responses and explanations for how to code each example.

Full credit is assigned where all the responses required in the task are both appropriate and different from each other. Partial credit is assigned in tasks requiring students to provide three responses and where two

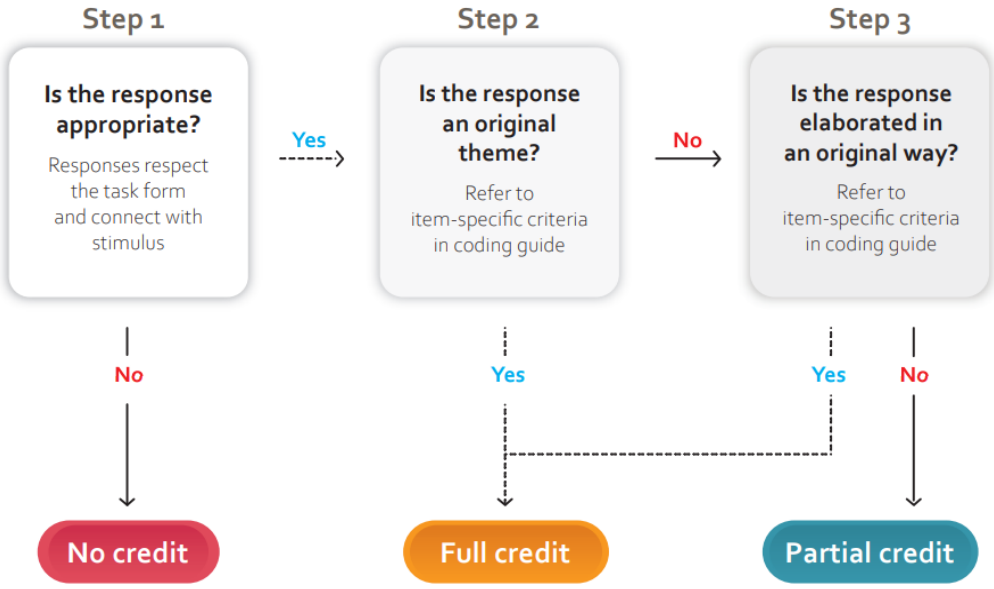
or three responses are appropriate, but only two are different from each other. No credit is assigned in all other cases.

Scoring of 'generate creative ideas' items

All items corresponding to the facet 'generate creative ideas' of the competency model require a single response. The general coding procedure for these items involves two or three steps depending on the content of the response. First, as with all items, coders must determine whether the response is appropriate. Then, coders must determine whether the response is original by considering two criteria (see Figure 18.3).

An original idea is defined as a relatively uncommon idea with respect to the entire pool of student responses. The coding guide identifies conventional themes for each item according to the patterns of genuine student responses revealed in the validation studies. If a response does not correspond to a conventional theme as described in the coding guide, it is directly coded as original; however, if an idea does correspond to a conventional theme, then coders must determine whether it is original based on its elaboration. The coding guide provides item-specific explanations and examples of original ways to elaborate on conventional themes. For example, a student might add an unexpected twist to a story idea that otherwise centres on a conventional theme.

Figure 18.3. General coding process for 'generate creative ideas' and 'evaluate and improve' items



This twofold originality criteria ensures that the scoring model takes into account both the general idea and the details of a response. While this approach does not single out the most original responses in the entire response pool, it does ensure that the coding process is less susceptible to culturally-sensitive grading styles that favour middle points or extremes and it provides some mitigation against potential cultural bias in the identification of conventional themes across countries.

Full credit is assigned where the response is both appropriate and original. Partial credit is assigned where the response is appropriate only, and no credit is assigned in all other cases.

Scoring of 'evaluate and improve ideas' items

All items corresponding to the facet 'evaluate and improve ideas' of the competency model require a single response and generally ask students to adapt a given idea in an original way rather than coming up with an idea from scratch. The general coding procedure for these items involves the same steps as those for the 'generate creative ideas' items. However, appropriate responses for these items must be both relevant and constitute an improvement. The threshold for achieving the appropriateness criteria for these items is thus somewhat strengthened with respect to items measuring the other two facets, as responses must explicitly connect to the task stimulus and attempt to address its deficiencies. The coding guide provides item-specific criteria, examples and explanations to help orient coders. For responses considered appropriate, coders must then establish the originality of the improvement by considering the same two originality criteria as for 'generate creative ideas' items.

Full credit is assigned where the response is both appropriate and an original improvement. Partial credit is assigned where the response is appropriate only, and no credit is assigned in all other cases.

PISA 2022 innovative domain test assembly design

According to the PISA assessment design, about 28% of the sample of PISA students were administered the creative thinking assessment. Students who took the creative thinking assessment spent one hour on creative thinking test items with the remaining hour of testing time assigned to one of the other core domains (mathematics, reading or scientific literacy).

The creative thinking items were organised into test units. The units vary in terms of the facets that are measured (i.e. generate diverse ideas, generate creative ideas, and evaluate and improve ideas), the domain context (i.e. written expression, visual expression, social problem solving, or scientific problem solving) and the duration of the unit (guidelines of between 5 and 15 minutes). Some units are composed of a single item and some units have multiple items. Dependencies between items within units was minimised.

The creative thinking units were then organised into five, mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 3 of this Technical Report.

Constructed-response tasks accounted for 92% of the items in the creative thinking assessment. The tasks typically call for a written response, ranging from a few words (e.g. cartoon caption or scientific hypothesis) to a short text (e.g. creative ending to a story or explanation of a design idea). Some constructed-response items call for a visual design response (e.g. designing a poster combining a set of given shapes and stamps) that is supported by a simple drawing editor tool. The assessment also included 2 items that were part of an interactive simulation-based task and two (possible) multiple-choice items where students are given the option to select a previously suggested idea or to generate a new idea.

PISA 2022 innovative domain assessment design and development

Test development for the PISA 2022 creative thinking assessment cycle began in early-2018 and focused on the development of items for a computer-based assessment. Through a process that included both CTEG contributions, as well as country submission and country review, Core B along with the PISA Secretariat selected an initial set of unit and item scenarios. Core B test developers then further developed the unit and item scenarios. The PISA Secretariat reviewed all unit scenarios and items early in the review process, prior to country reviews, to ensure the items fulfilled the goals of the assessment framework.

The developed units were submitted for translatability review at the same time that they were released for country review. Linguists representing different language groups provided feedback on potential translation, adaptation and cultural issues arising from the initial wording of items. Experts at cApStAn and the translation referee for the 2022 cycle alerted test developers to both general wording patterns and specific item wording that are known to be problematic for some translations and suggested alternatives. This allowed test developers to make wording revisions at an early stage, in some cases simply using the alternatives provided and in others working with cApStAn to explore other possibilities.

To ensure that the creative thinking assessment items were understood the same way across linguistic and cultural groups, participating countries also engaged in several cycles of review of the test material to help identify items that may be likely to suffer from cross-cultural bias. This enabled problematic cultural and linguistic characteristics to be identified during the early stages of the assessment development process. Countries had two weeks to perform reviews and submit feedback on all draft stimuli and items.

Preparation of the French source version for all of the test units provided another opportunity to identify issues with the English source version related to content and expression. The development of the two source versions helped to identify instances where wording may prove problematic for translation and could be modified to simplify translation into other languages, and specified where translation notes would be needed to ensure the required accuracy in translating items to other languages.

Cognitive laboratories

Experienced testing professionals were engaged to conduct cognitive laboratory exercises with students in Australia, Singapore and the United States. A total of 66 students across the three countries participated in these cognitive labs in the period between August and September 2018 on set of eight prototype units across the four domain contexts. In the format of concurrent and retrospective thinking-out-loud exercises, students around the age of the PISA population were asked to explain their thought processes while completing the test items and to point out any difficulties or misunderstandings in the instructions or stimulus material. After students completed all of the tasks within each unit, they were asked to answer a series of probing questions about their experience working through the tasks including specific questions on the comprehensibility of the task prompt and perceived difficulty of the task. Students also went through each task a second time to verbalise any thoughts they had had when working through the task. The cognitive laboratories helped to evaluate whether students could understand what they were asked to do during the test, whether students perceived the tasks as engaging, excessively demanding or frustrating, and whether they needed more clarifications to be added to the task prompts.

The analysis of the information collected during these sessions, as well as from video recordings, identified opportunities for the revision and optimisation of items as well as to correct several identified bugs in the testing platform (ACT, 2018[6]). Insights from the cognitive laboratories included:

- **Refining the number of required responses in ‘generate diverse ideas’ items.** In the prototype units tested in the cognitive labs, students could enter as many responses as they wished on these items. In general, students created up to three responses for these items with relative ease but expended considerable effort to move beyond three responses. Moreover, their fourth or fifth responses were rarely their most creative ones. As a result, in successive revisions of the test material, the test development team decided to ask students for up to three responses only and emphasise in the task prompt that students should aim to provide responses that are as different as possible from each other.
- **Choosing the number of required responses in time-intensive items.** Some of the prototype items required a significantly greater investment of time, elaboration and careful execution than others. It was evident from students’ feedback and actual responses to some tasks that asking them to iterate upon or produce more the one response could easily generate fatigue; it was thus

decided that in these types of time-intensive tasks, students should be asked to generate no more than one response.

- **Providing guidance about time required on the task.** In the cognitive labs, students could spend as much time as they wished on each task, although many expressed the need to have some kind of guidance on timing. Different solutions for providing guidance on time usage were considered; the test development team decided to provide an indication of the maximum amount of time that students should spend on each task in the respective task prompt to help students manage their time.
- **Clarifying task expectations and instructions.** Some students requested further clarification on certain terms used in the tasks prompts (e.g. "original"). The prompts were subsequently revised to reduce subjective interpretations of such terms as much as possible. For example, a clarification was added to explain that "original" refers to a solution that other students might not have thought of (clearly associating originality to statistical frequency).
- **Selecting the right images as task stimuli.** Several tasks use a visual stimulus and ask students to engage in idea association in order to generate a response inspired by the image. Some images used in the prototype units evoked associations that were strongly culturally-mediated (for example, some students thought of the Beatles' song when they saw the image of a yellow submarine). While cultural influences upon student responses cannot be completely eliminated, the development team revised any images that were clearly susceptible to inspiring culture-specific associations.
- **Defining features of the drawing tool.** Most students rapidly understood how to use the drawing tool provided in the platform. However, in some cases students clearly lost precious time trying to complete specific actions that were not immediately intuitive (e.g. deleting an object.) These issues have been addressed by including a tutorial on the use of the drawing tool. The test development team evaluated the potential advantages and disadvantages of including additional features in the drawing tool and decided to keep a relatively simple tool with limited graphical instruments to limit any potential unfair advantage to those students who are more proficient in doing graphical work on a computer. The analysis of responses from the cognitive labs confirmed that it is possible to generate highly creative outputs using only a limited version of the drawing tool.

Six of the eight prototype units were further developed after the cognitive labs for inclusion in subsequent validation studies, while two units were abandoned at this stage due to unsatisfactory performance in the cognitive labs. A further set of units were also developed in accordance with the insights from the cognitive laboratories.

Small scale validation exercises

Further small-scale validation exercises were conducted in parallel to the overall test development process, in an iterative manner, to observe how the then-current test materials functioned under similar test conditions to the field trial and main survey. The purpose of these validation studies was severalfold:

- to provide evidence on the performance of the creative thinking assessment in PISA-like classroom settings;
- to collect sample student responses in multiple countries to inform the development of the coding and scoring guides;
- to assess the inter-rater reliability of human coded items (i.e. the agreement between raters);
- to gain insights into the difficulty of the items;
- to determine the extent to which a creative thinking score or sub-scores could be obtained from the creative thinking assessment; and

- to gain preliminary insights on the essential coder training materials and processes needed for human coders.

A total of 703 15-year-old students from Singapore (n=206), Australia (n=234) and Canada (n=263) participated in the first validation study between October to November 2018. Samples were recruited through the PISA National Project Managers and coordinated with the PISA Secretariat. The validation study instrument included 12 fully functional prototype units delivered in 3 test forms, with 4 units per form. Each form contained one unit per domain.

The coding of the units was carried out according to the preliminary coding guides developed by ACT. Student responses were scored by a team of professional scorers at ACT. As a group, the team reviewed and assigned scores to 5% of the available responses for each task, which enabled scorers to build a common understanding of the coding procedures. Each response was then coded independently by two scorers. Any questions or issues that arose during the scoring of the data were referred to the Scoring Supervisor and the Assessment Design team at ACT.

An analysis of the genuine student data indicated items that did not perform as intended and informed evidence-based improvements to the test material, as well as development of and improvements to coder training material such as the coding guide (ACT, 2019[7]). The validation study also helped to refine the methodology followed for scoring students' responses – in particular, it informed the introduction of a double criteria for coding the originality of responses taking into consideration both the originality of the theme of a students' response and the originality of their approach – and provided genuine responses for the international coder workshops.

A total of 202 15-year-old students from the Republic of South Africa participated in the second validation study from February to March 2019. ACT and the PISA Secretariat partnered with the Care for Education in Republic of South Africa, with support from the LEGO Foundation, to carry out the validation study. It included 16 units, delivered in two test forms. This validation study was delivered on paper to simplify administration procedures (only a limited time was available for the recruitment of schools and it was not possible to condition participation to the availability of computer equipment). Each test form contained the same units, but the order of the units presented to students varied to mitigate potential order effects on performance.

In the second validation study, student responses were scored by a team of non-professional scorers at Care for Education that were trained by the scoring team at ACT following a standard training process and with the support of the international coding guides. Similarly to the first validation study, each response was coded independently by two scorers.

The second validation study provided valuable insights into the success of revisions to the units throughout the test development process. In general, the performance of the creative thinking item pool in the second validation study improved upon the performance of the items included in the first validation study, particularly in terms of inter-rater reliability.

Field trial

The field trial for creative thinking was initially scheduled for 2020; however, this timeline was disrupted by the COVID-19 global pandemic meaning only a limited field trial (LFT) was carried out, with findings further investigated during a second administration of the field trial in 2021. The LFT conducted in 2020 with 11 countries provided preliminary evidence in support of:

1. the psychometric quality of the PISA 2022 creative thinking assessment units in terms of their validity, reliability, and comparability across participating countries;
2. the ability to construct a creative thinking scale and, possibly, subscales;

3. the inclusion of all the creative thinking units and forms in Field Trial 2021; and
4. further enrichment of the coder training materials utilised in coder training for the full field trial in 2021 and the main survey administration in 2022 (ACT, 2020[8]).

In 2021, a further field trial was conducted with 44 countries to provide additional evidence of the validity and reliability of the creative thinking assessment.

Field trial coder training

Among the total 38 items administered, two items were machine-scored (the simulation-based items) and the remaining 36 items were human-scored items. For the human-scored items, all coding processes were performed by each country's coders. The ACT team provided international coder training and supported the national coding teams through a standard PISA query service.

Limited field trial (2020)

The coding guide for the PISA 2022 creative thinking assessment was developed by test developers and performance scoring experts at ACT, with the support of the PISA Secretariat. Coder training procedures and materials were informed by the cognitive labs and validation studies and included examples of genuine student responses.

The English master version of the coding guide was released in a draft form prior to the in-person PISA International Coder Training meeting in January 2020. The training objectives included developing a foundational understanding of the creative thinking construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. Test developers and performance scoring experts from ACT, with the support of the PISA Secretariat, facilitated discussions at that meeting. The coding guide used in the limited field trial was finalised based on these discussions. The updated English version of the coding guide and the French source version were subsequently released to countries in February 2020 prior to the beginning of the limited field trial data collection period.

Field trial (2021)

The International Coder Training meeting for creative thinking ahead of the full field trial was held virtually over 5 days due to the COVID-19 pandemic in February 2021. Performance scoring experts from ACT developed online coding training modules and facilitated an interactive coder training workshop, held with representatives from the participating countries in the 2021 field trial prior to coding. To facilitate the online coder training, ACT's team developed comprehensive exemplar sets consisting primarily of authentic student responses that were selected and intended to demonstrate a typical response for each credit level and theme assignment (i.e. codes 00, 11, 12, 13, 21, 22, 23, etc., with code 29 used to designate an unlisted theme). Discussion was also dedicated to reinforcing understanding and consensus about the coding rules for each item to better ensure consistency of coding within and between countries.

Facilitators reviewed the layout of the coding guide, general coding principles, common problems, and guidelines for applying special codes. Workshop materials were optimised based on feedback from the LFT coder training, LFT coder queries and translation referee updates to the earlier version of the coding guide. Attendees were required to code the workshop materials (i.e. the exemplar sets) "live" during the interactive workshop; where there were disagreements about the coding for an item, those were discussed in detail so that all attendees understood, and would be able to follow, the intent of the coding guides. In some instances, disagreements – particularly those highlighting possible cultural bias – led to modifications of the coding guide and/or workshop materials.

Preparation of the field trial data collection instruments

The process for creating the field trial national student delivery system (SDS) began with the assembly and testing of the master SDS, followed by the process for assembling national versions of the field trial SDS. After all components of the national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first by Core 2. Countries were then asked to check their SDS and identify any remaining content or layout issues. Once countries signed off on their national SDS, their final systems were released for the field trial. The PISA 2022 creative thinking assessment was only administered on computers.

Field trial coding procedures

The field trial design required that two independent coders review and code each student's responses at a credit level of either 0,1 (i.e. no credit or credit), or 0, 1, or 2 (i.e.no credit, partial credit or full credit), thus generating inter-rater reliability at the credit level. In addition, two selected English-fluent bilingual coders from each country reviewed and coded 30 pre-designated anchor responses to verify coder reliability across countries. These anchor responses were selected from earlier validation studies conducted in Australia, Canada, Colombia, Singapore and South Africa, and represented a range of responses at all credit levels (ACT, 2019[7]). Inter-rater reliability (IRR) on the anchor responses across all items and coder pairs was high (0.71). The average quadratic Kappa was also high (0.79).

For the items measuring either the 'generate creative ideas' or 'evaluate and improve ideas' facets, coders were required to use a second digit to indicate the primary theme of each response that earned either partial or full credit. Partial credit responses could only be coded using values of 1-3 as their second digit (i.e. codes 11, 12 or 13), to represent correspondence with the initial conventional themes designated in the coding guide based on an analysis of available student responses in the validation studies; however, responses that received full credit could use up to 9 different values for the second digit (i.e. codes 21 through 29), with the ninth value representing all themes not associated with themes 1-8. The resulting data informed distinctions between "conventionality" and "unconventionality" of themes across a diverse international student cohort.

Field trial coder queries procedures

As was the case during previous cycles, Core A set up and maintained a coder query service for the 2020 and 2021 field trials. Countries were encouraged to send coder queries to the service so that a common adjudication process was consistently applied to all coder questions about constructed-response items. Core B test developers and performance scoring experts from ACT reviewed and responded to coder queries that were specific to the creative thinking test.

In addition to responses to new queries, Core B curated a selection of queries to include in the Coder Query Log containing accumulated responses from previous cycles of PISA. This helped foster consistent coding of creative thinking items. The query log was regularly updated and posted for National Centres on the PISA portal as new queries were received and processed.

National item review post-Field Trial

The item feedback process began in August 2021 and concluded in October 2021 and was conducted in two phases. Phase 1 occurred before countries received their field trial data and Phase 2 after receipt of their data. This two-phase process was implemented to allow for the most efficient correction of any remaining errors in item content or layout given the extremely short turnaround period between the field trial and main survey.

Phase 1 allowed countries to report any linguistic or layout issues that were noted during the field trial, including errors to the coding guides. All requests were reviewed by Core B. Following the release of the field trial data, countries received their Phase 2 updated item feedback forms that included flags for any items that had been identified as not fitting the international trend parameters. Flagged items were then reviewed by national teams. As was the case in Phase 1, countries were asked to provide comments about these specific items in instances where they could identify serious errors. Requests for corrections were reviewed by Core B and, where approved, implemented.

Field Trial outcomes

The 2021 Field Trial data analyses addressed the issue of construct and score validity and reliability, within and across countries, in addition to differential item functioning. Following the field trial data collection, the items were analysed for inter-rater reliability on anchor responses, inter-rater reliability on all responses, average Quadratic Kappa, item category response functions, item quality, and item omit and not-reached rates. Items that exceeded the omit and not-reached rates were identified and investigated; in some cases, this could be attributed to technical issues with some items during the administration of the test, and cluster placement was also considered to be a contributing factor.

Other analyses of the data included item difficulty, item discrimination, item response time, position effects, IRT scaling, item model fit, IRT parameters and student theta estimates, the evaluation of sub-scores on domain and facet levels, and differential item functioning (DIF) via the item-total score curves from different country-by-language groups. Any flagged items for DIF were further reviewed in terms of their sample size, contents, translations and coding guides (i.e. verified translation vs non-verified translation of coding guides), student responses (indications of misunderstanding), performance in alternative languages for that country, performance on similar items in assessment for that country/language, performance on the other items in that unit, additional item flags for that item, LFT data vs. FT data, and planned optimisations for that item (e.g., theme changes, coding optimisations or cluster placement).

Due to the operational timeline in PISA, it was not possible to include new items in the creative thinking test after this phase and no substantial modifications were made to existing test items, i.e. poorly performing items were removed from the test item pool to ensure a proper coverage of the construct in the main survey. Following the field trial analyses, one unit consisting of two items was removed (see Table 18.1).

In summary, the findings from the field trial analysis supported:

1. the psychometric quality of the PISA 2022 creative thinking assessment units in terms of their validity, reliability, and comparability across participating countries;
2. the ability to construct a creative thinking scale; and
3. the inclusion of 20 of the 21 creative thinking units administered in the field trial for administration in the 2022 main survey.

The field trial(s) also generated insights for the further enrichment of the coder training materials, including the coding guide, prior to the 2022 main survey. Substantial work was undertaken including reviewing large amounts of genuine student responses, conducting an additional frequency analysis of response themes, and identifying instructions that caused coding issues by being absent, too vague or too restrictive. This resulted in substantial modifications of the coding guide, including updates to the designation of conventional and unconventional themes, the refinement of theme descriptions, the increased representation of exemplar responses, and edits to the item-specific instructions to facilitate effective and consistent coding.

PISA 2022 main survey

The PISA 2022 main survey was conducted between March and December 2022. The majority of countries completed the main survey data collection by August 2022. In preparation for the main survey, countries reviewed items based on their performance in the field trial and were asked to identify any serious errors still in need of correction. The Core B contractors worked with countries to resolve any remaining issues and prepare the national instruments for the main survey.

Item review and selection

The PISA 2022 field trial provided evidence in support of the psychometric quality of the PISA 2022 creative thinking assessment units in terms of validity, reliability, and comparability across participating countries. Maintaining the same range of contexts from the field trial to the main survey provided good continuity and kept a consistent representation of skills and domains. Clusters were created following the final item selection and balanced based on the coverage of cognitive processes, the discrimination and difficulty of the items, and the total number of units and items. The duration of each unit was between 5 and 15 minutes. The units were organised into five mutually exclusive 30-minute blocks or clusters, and the clusters were rotated according to the integrated design presented in Chapter 3 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking within the two thematic content areas (creative expression, and knowledge creation and problem solving) and the four domains.

The CTEG reviewed the field trial data and outcomes, the approach to item selection, the content and balance of the proposed main survey clusters, and signed off on the selection.

Main survey coder training

The main survey International Coder Training for creative thinking was held in February 2022. Analysis of student responses and coder queries during the field trial administration helped performance scoring experts from ACT improve upon the online coding training modules and other coder training and workshop materials. Additional sample responses were included in the coding guide to better illustrate different types of student responses. Workshop materials were also enhanced to include additional authentic student responses that better illustrated the boundaries between full credit, partial credit (where appropriate) and no credit.

The main survey coder training process was similar to that ahead of the 2021 field trial in that self-guided online training modules were completed before full-group discussions. The training objectives again included developing a foundational understanding of the construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. Facilitators again reviewed the layout of the coding guide, general coding principles, common problems, guidelines for applying special codes, and workshop materials for each item. Following the international coder training, additional and final revisions were made to the coding guide in response to discussions that took place at the meeting.

Preparation of data collection instruments

The process for creating the main survey national student delivery system (SDS) followed the approach used during the field trial, beginning with the assembly and testing of the master SDS followed by the process for assembling national versions of the main survey SDS. After all components of national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first by Core 2. Countries were then asked to check their SDS and identify any remaining content or layout issues. Once countries signed off on their national SDS, their final systems

were released for the main study. The PISA 2022 creative thinking assessment was only administered on computers.

Main survey coder queries

The coder query service was again used in the main survey as it was in the field trial to assist countries in clarifying any uncertainty around the coding process or students' responses. Queries were reviewed and responses were provided by domain-specific teams including test developers and coding experts. Core B test developers and performance scoring experts from ACT reviewed and responded to queries specific to the creative thinking test. Relevant queries were included in the Coder Query Log, a resource maintained by Core A and accessible by all participant NPMs in the PISA Portal.

Data adjudication and approach to scaling the data for reporting

In June 2023, Core A presented the Technical Advisory Group (TAG) with the PISA 2022 creative thinking data and preliminary psychometric analyses for data adjudication. Following the initial feedback of the TAG on the scalability of the data given the relatively low inter-item correlations and the creation of plausible values, the PISA Secretariat conducted further analyses of the creative thinking data including modifying some of the scoring rules with the goal of increasing the validity of inferences drawn from the creative thinking data, and improving the scalability and comparability across countries.

Following a thorough review of the data, the following changes were implemented:

- **Four items were dropped from the scaling.** The four items identified for exclusion were drawn from two units (one visual expression, and one scientific problem solving) and were all in the same test cluster. These four items showed poor discrimination and high omit rates, likely due to their position within the cluster.
- **The scoring rules for 14 items were modified.** All 'generate creative ideas' and 'evaluate and improve' items were reviewed following the main survey in terms of the distribution of double-digit codes across countries. The scoring process for these items required coders to use a second digit to indicate the primary theme of each response, and those coded using values of 1-3 as their second digit (i.e. 11, 21, 12, 22, 31 or 32) represented correspondence with the initial conventional themes designated in the coding guide. The double-digit codes were intended to serve as a mechanism through which to review the distribution of codes across countries and adjust the themes designated as conventional following the field trial and main survey. The number of conventional themes were modified for 14 of the 18 items corresponding to 'generate creative ideas' and 'evaluate and improve ideas' based on the results of the main survey to improve the validity of the scoring rules for these items and to align the scoring with the framework (i.e. originality as statistical infrequency, with respect to the responses of other students who completed the same task).
- **Responses submitted in fewer than 15 seconds were invalidated (i.e. converted to missing responses).** For most items in the creative thinking test, students must generate a written or visual artefact in response to a written or visual stimulus (i.e. task prompt with instructions and material for inspiration). The construct of creative thinking also aims to measure the cognitive processes associated with idea generation, evaluation and improvement, which are considered to be slow and thoughtful processes rather than reflective of opportunistic or rapid processes. For most items in the test, responses submitted within 15 seconds of viewing the item cannot be considered reflective of creative thinking processes. A review of the timing data for the items also showed a clear bimodal distribution of response submission, with one peak prior to 15 seconds and another peak a significant time afterwards. This modification was applied to all items, with the exception of

three: in two cases, students were able to select a response to a previous question akin to a multiple-choice mechanism; and in the other item, students were asked to generate a very short written artefact. In these three cases, it was judged that students could submit a response that reflected creative thinking processes within 15 seconds and thus no minimum response time was imposed.

In October 2023, the PISA Secretariat, Core A and the TAG reconvened for the data adjudication of the creative thinking data following the further analyses conducted by the PISA Secretariat and to finalise the reporting approach. The TAG recommended to report the creative thinking data according to a non-linear transformation of the "theta" scale, using the test-characteristic curve for a hypothetical test using the final pool of 32 creative thinking items and based on international item parameters. The advantages of this approach include:

- Reporting student performance according to a bounded scale (between 0-60, reflecting the maximum sum-score of all items) that is the same for all countries. This solution maintains the possibility to report performance on a scale, but signals a clear difference to the PISA scales used for the other domains and the broader "grain" size of the creative thinking scale signals its relative lower reliability compared to the other PISA scales (a 1 point change in the creative thinking scale reflects about 10% of a standard deviation).
- Scores can be easily interpreted in terms of the number of items correct on this specific test (rather than a more general reflection of students' creative thinking ability applied to other performance tests), drawing attention to the actual test content and the framework that guided its development and facilitating the interpretation of the relatively high frequency of low scores in this test (i.e. students scored 0 on the test, rather than not having any creative thinking skill).
- Test scores differ more where the test has more information about students.
- The international database still includes 10 "plausible scores" per student.

Performance level descriptors

Following the data adjudication process and the finalisation of the scale for reporting the creative thinking data, the PISA Secretariat, in collaboration with Core B and the CTEG, defined performance level descriptors. Performance on the creative thinking scale was split into 6 performance levels.

Level 1

At level 1, students can generate very simple visual designs using isolated shapes or existing visual elements, and in some cases very short written artefacts (e.g. a few words), that require them to engage their imagination. In general, students at this level rely on obvious themes or idea associations as the basis for their response and struggle to generate more than one appropriate idea even for open and simple imagination tasks. These students typically generate simple visual or written artefacts with few details that reflect a minimal level of engagement with the task.

Level 2

At level 2, students can generate appropriate ideas for simple visual and written expression tasks as well as those that focus on solving familiar, everyday social problems. With respect to students at level 1, students in level 2 can develop simple written ideas in the form of longer captions or short dialogues. Students at level 2 typically suggest ideas that rely on obvious idea associations for expressive tasks or that refer to existing solutions for problems in social problem-solving tasks. Students can generate more than one appropriate idea for some written expression and social problem-solving tasks, but these ideas are not qualitatively different to one another.

Level 3

At level 3, students can generate one or several appropriate ideas for simple to moderately complex expressive and problem-solving tasks, including extended written ideas that require them to engage and express their imagination and coherently build upon others' ideas. Students at level 3 still typically suggest ideas that rely on obvious idea associations or common themes with respect to their peers, but they begin to demonstrate the ability to recognise and generate original solutions for familiar, everyday problems with a social focus. They may suggest solution ideas that not many other students think of or add an innovative or different twist to more conventional solution ideas.

Level 4

At level 4, students can productively engage in idea generation across a range of expressive and problem-solving tasks. Students at level 4 can also generate original and diverse ideas for simple tasks in more familiar domain contexts. With respect to students at level 3, students at this level can generate an appropriate idea for most types of idea generation task, including more complex or unfamiliar problem-solving tasks and tasks in a scientific context. They can also build on others' ideas for solutions in social and scientific contexts, although they tend to provide an obvious or common iteration with respect to their peers. Students at level 4 can generate their own original ideas in written expression tasks and sometimes when iterating on others' ideas. They can express their imagination in unexpected ways, making unconventional idea associations between elements of the stimulus and their written artefact, or they can add atypical details to elaborate creatively on more common ideas. Students at this level can often suggest two or three qualitatively different ideas in open written expression and social problem contexts but are less successful in more complex or constrained social and scientific problem contexts.

Level 5

At level 5, students can productively engage in creative idea generation, generating both original and diverse ideas for a range of expressive and problem-solving tasks. Students at level 5 can think of qualitatively different ways to express their imagination and to address familiar social and scientific problems. They can make several different idea associations, considering different interpretations and perspectives on the same issue or stimulus. For both simple and more abstract written expression tasks, they can use their imagination to create original written artefacts that make unconventional associations between ideas or that add atypical details to elaborate creatively on common themes. With respect to students at level 4, students can create original visual artefacts that combine elements in an unusual or unexpected way for open visual design tasks. Students at this level can also generate unconventional solution ideas that integrate innovative approaches in familiar social, and sometimes scientific, problem contexts. This includes when tasked to iterate on and improve an existing solution idea in more open, familiar problem contexts.

Level 6

At level 6, students can productively engage in creative idea generation, generating both original and diverse ideas for a wide range of expressive and problem-solving tasks including those in more complex, abstract and unfamiliar contexts. With respect to students at level 5, students at this level can identify weaknesses in existing solutions to social or scientific problems, including those that are in less familiar contexts, and build on this understanding to suggest original and innovative ways to improve solutions. They can also generate several appropriate solution ideas for complex social and scientific problems that require more specific knowledge of the domain context and that have a more restricted solution space. For expressive tasks, students at level 6 can create and improve more abstract visual designs, combining visual elements and representations in unexpected ways and conveying an original interpretation or iteration of an existing representation.

References

ACT (2020), *PISA 2022 Creative Thinking Limited Field Trial Research Report*, ACT, Iowa City, IA. [8]

ACT (2019), *PISA 2021 Creative Thinking Validation Study Research Report*, ACT, Iowa City, IA. [7]

ACT (2018), *PISA 2021 Creative Thinking Cognitive Lab Research Report*, ACT, Iowa City, IA. [6]

Cropley, A. (2006), "In Praise of Convergent Thinking", *Creativity Research Journal*, Vol. 18/3, pp. 391-404. [5]

Guilford, J. (1956), "The structure of intellect", *Psychological Bulletin*, Vol. 53/4, pp. 267-293, https://doi.org/10.1037/h0040755. [2]

Guilford, J. (1950), "Creativity", *American Psychologist*, Vol. 5/9, pp. 444-454, https://doi.org/10.1037/h0063487. [4]

OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/dfe0bf9c-en. [1]

Plucker, J., R. Beghetto and G. Dow (2004), "Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research", *Educational Psychologist*, Vol. 39/2, pp. 83-96. [3]

Chapter 18 tables

Table 18.1. Distribution of items across facets and domains for the PISA 2022 creative thinking test

Domain	Facet					
	Field trial administration			Main survey administration		
	Generate diverse ideas	Generate creative ideas	Evaluate and improve ideas	Generate diverse ideas	Generate creative ideas	Evaluate and improve ideas
Written expression	4	6	2	4	6	2
Visual expression	2	2	4	2	1	1
Social problem solving	4	3	3	4	3	3
Scientific problem solving	4	1	3	4	1	3
Total	14	12	12	14	11	11