PISA 2022 Technical Report

4 Creative Thinking Test Design and Test Development

Introduction

This chapter describes the assessment design for the PISA 2022 Innovative Domain: Creative Thinking (CT) as well as the processes used by the PISA Core B3 contractors, ACT and Cito, and the international test development team to develop the innovative domain assessment for the PISA 2022 cycle.

Activities for the innovative domain test design and test development included the following:

- The creation of a CT Expert Group to guide test design and test development
- Development of a CT assessment framework
- Assessment development
- Creative thinking validation studies
- Field Trial
- Main Survey

The Role of the Creative Thinking Expert Group in Item Development

As the Core B3 contractor in charge of Creative Thinking instrument development, ACT was responsible for working with the Creative Thinking Expert Group (CTEG) as applicable. Work focused on understanding the CTEG's vision for the Creative Thinking framework as well as the range and types of items to be developed for PISA 2022 Creative Thinking assessment. CTEG members began work on the framework in September 2017 and finalized the framework September 2022. Core B3's work with the 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 (OECD, 2019[1]).
- 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 (e.g., drawing tool, simulation, innovative item types) that would be desirable to develop for measuring the construct and would be feasible to implement in the context of PISA.

Work with the CTEG continued beyond the initial meeting through instrument development and data analysis. CTEG members played an important role in reviewing assessment tasks as they were developed, providing input into the analysis of the Field Trial (FT) data, approving the set of items for the Main Survey, and working with development and analysis staff to develop the described scales and performance level descriptors used for reporting the PISA 2022 CT results.

PISA 2022 Creative Thinking Assessment Framework

The PISA 2022 CT assessment focused on the creative thinking processes that one can reasonably expect from 15-year-old students. It does not aim to single out exceptionally creative individuals, but rather to describe the extent to which students are capable of thinking creatively when searching for and expressing ideas, and 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' creative thinking competence that have clear implications for education policies and pedagogies. 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.

As the innovative domain for the PISA 2022 cycle, the creative thinking assessment focused on the skills that twenty-first century students need as organizations and societies around the world increasingly depend on innovation and knowledge creation to address emerging challenges, giving urgency to innovation and creative thinking as collective enterprises. The domain is defined as follows:

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, 2019[1]).

Three cognitive facets that support creative idea generation and evaluation were further defined and included:

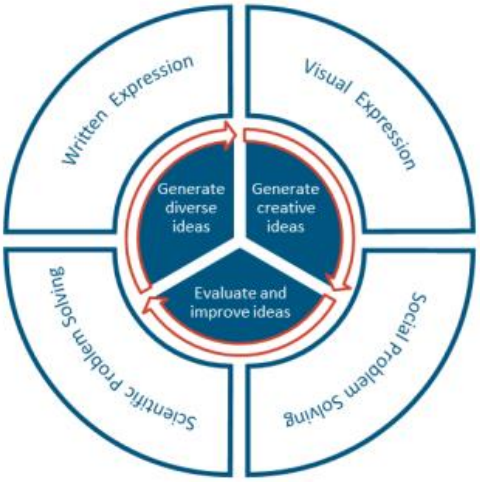
- **Generate Diverse Ideas (GDI)**: students are asked to develop two or three ideas and are assessed on the appropriateness of these ideas (their alignment with the task requirements) as well as whether the two or three ideas are sufficiently different from one another.
- **Generate Creative Ideas (GCI)**: students are asked to provide creative ideas and are assessed on the appropriateness of these ideas as well as whether the ideas occur with thematic infrequency.
- **Evaluate & Improve Ideas (EII)**: students are asked to improve on the creativity of an idea that is provided to them and are assessed on whether the idea occurs with thematic infrequency.

As creative thinking can be expressed in a large number of possible applications, and the nature of these applications influence the knowledge and skills that are required to produce a creative output four domains were chosen for the PISA 2022 CT assessment:

- Written Expression
- Visual Expression
- Social Problem Solving
- Scientific Problem Solving

The resulting competency model allows students the opportunity to demonstrate their capacity to generate, evaluate, and improve ideas across four distinct domains of applications. This design is expected to provide information about students' strengths and weaknesses across countries.

Figure 4.1. Competency model for the PISA test of creative thinking



Items were distributed across facets and domains to allow for a range of opportunities for expression. The distribution for the field trial included 14 generate diverse ideas items, 12 generate creative ideas items, and 12 evaluate & improve items These are shown in Table 4.1.

PISA 2022 Innovative Domain Assessment Design

According to the 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 items with the remaining hour assigned to one of the other core domains (mathematics, reading, or science).

Creative thinking items were organized into test units. The units vary in terms of the facets that are measured, the domain and duration. Items were distributed within the units with some units having a single item and some units having multiple items.

Dependencies between items within units was minimized. The duration of each unit was between 5 and 15 minutes. The units were then organized into five mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 2 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking and the four domains.

The items used to assess the creative thinking facets required of three different types of responses. Constructed-response tasks accounted for 92% of the items in the assessment. These 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 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 two items that were part of an interactive simulation-based task which employs an interactive simulation environment and two items that consist of a task which calls for answers that are based on the choice of selecting a previously suggested idea or generating a new idea.

PISA 2022 Innovative Domain Assessment 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 B3 along with the OECD selected a final set of item scenarios. Core B3 test developers further developed the scenarios. The OECD reviewed all scenarios and items early in the review process, prior to country reviews to ensure the items fulfilled the goals of the revised framework.

Newly developed units were submitted for translatability review at the same time 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 PISA 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 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 new units provided another opportunity to identify issues with the English source version related to content and expression. Development of the two source versions helped identify instances where wording 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.

Experienced testing professionals were engaged to conduct cognitive laboratory exercises with students in Australia, Singapore, and the United States. In the format of thinking-out-loud exercises, students around the age of the PISA population were asked to explain their thought processes in answering, and point out any difficulties or misunderstandings in the instructions or stimulus material. Information from these sessions was used to identify opportunities for revision and optimization of items as well as to correct several identified bugs (ACT, 2018[1]).

Validation Studies were conducted in parallel to the overall test development process, in an iterative manner, in order to observe how the then-current test materials functioned under similar test conditions. The purpose of each Validation Study was to provide evidence on the performance of creative thinking assessment in PISA-like classroom settings, collect sample student responses in multiple countries, assess the inter-rater reliability of human coded items (i.e., agreement between raters); determine the extent to which a creative thinking score or sub-scores can be obtained from the creative thinking assessment; and gain preliminary insights on the essential training materials needed for human coders.

A total of 703 15-year-old students from Singapore (206), Australia (234), and Canada (263) participated in the Validation Study between October to November 2018. Samples were recruited through PISA National Project Managers and coordinated with the OECD Secretariat.

The Validation Study instrument included 12 fully functional prototype units delivered in three forms, four units per form. Each form contained one unit per domain. Each unit included between 4-6 items (tasks). An analysis of the genuine student data indicated items that did not perform as intended (e.g., inter-rater scoring agreement, item difficulty, credit distribution), 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[3]). The validation study also helped refine the methodology followed for scoring students' responses and provided genuine responses for the international coder workshops.

Field Trial

The Field Trial for creative thinking was initially scheduled for 2020; however, this timeline was disrupted by the COVID-19 pandemic, with findings to be further investigated during a second administration of the Field Trial in 2021. The limited field trial (LFT) conducted in 2020 with 11 countries provided preliminary evidence in support of: (a) the psychometric quality of PISA creative thinking assessment units in terms of validity, reliability, and comparability across participating countries; (b) the ability to construct a Creative Thinking scale and, possibly, subscales; (c) the inclusion of all the creative thinking units and forms in Field Trial 2021. It also generated (d) insights for further enrichment of the coder training materials utilized in coder training for Field Trial 2021 and the Main Survey 2022 (ACT, 2020[4]).

In 2021 a further Field Trial (FT) was conducted with 44 countries to provide additional evidence of the validity and reliability of the creative thinking assessment. Among the total of 38 CT items, two items were machine-scored, and the remaining 36 items were human-scored items. For the human-scored items in the 2020 LFT and the 2021 FT, all coding processes were performed by each country's coders. The ACT team provided national coder training and supported the national coding teams through a standard PISA query service. Items were initially reviewed for appropriateness (e.g., on task and on topic). Items determined to be appropriate were then scored using a single-digit or double-digit rubric. Scoring of Generate Creative Ideas and Evaluate and Improve Ideas items was conducted using a double-digit scoring rubric which captured data on the primary focus of a student response in addition to reflecting its credit level. Students demonstrated creativity in these facets by utilizing unconventional foci or employing innovative approaches. Scoring of the Generate Diverse Ideas items was conducted using a single-digit scoring rubric. Students demonstrated creativity in this facet by generating multiple, different ideas (see Table 4.2).

2020 Limited Field Trial Coder Training

The coding guide for creative thinking was developed by test developers and performance scoring experts at ACT for the Field Trial with the support of the OECD. 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 Creative Thinking Coding Guide was released in draft form prior to the in-person PISA International Coder Training meeting in January 2020.

Test developers and performance scoring experts from ACT, with the support of the OECD, facilitated discussions at that meeting. The coding guide used in the limited field trial was finalized based on these discussions. The updated English version of the coding guide and the French source version were released to countries in February 2020, prior to the beginning of the limited field trial data collection period.

2021 Field Trial Coder Training

The 2021 field trial International Coder Training for creative thinking was held over five days, virtually, due to the pandemic, in February 2021. Performance scoring experts from ACT developed online coding training modules and facilitated an interactive coder training webinar, held with representatives from participating 2021 Field Trial countries prior to coding. The training objectives 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. In order to facilitate coder training, ACT's team developed comprehensive exemplar sets consisting primarily of selected authentic student responses intended to demonstrate a typical response for each

credit level and theme assignment (i.e., codes 00, 11, 12, 13, 21, 22, 23, and so on, with code 29 used to designate an unlisted theme, as explained in Section X). Discussion was also dedicated to reaching 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 optimized based on feedback from the LFT coder training, LFT coder queries, and translation referee updates to the 2021 coding guide. Workshop materials comprised primarily sample student responses that were provided for each item, and attendees were required to code them during the interactive workshop. Where there were disagreements about 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 data collection instruments

Preparing the Field Trial national student delivery systems (SDS)

The process for creating the field trial national SDS followed the approach used during the field trial, beginning with assembly and testing of the master SDS followed by the process for assembling national versions of the field trial 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 A. 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. PISA 2022 CT was only administered on computer-based participants.

Field Trial Coding Procedures

The FT design required that two independent coders review and code each student's responses at a credit level of either 0,1 (no credit or credit) or 0, 1, or 2, (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 the countries. These anchor responses were selected from earlier pilot studies conducted in Australia, Canada, Colombia, Singapore, and South Africa, and represented a range of responses at all credit levels (ACT, 2019[3]).

For the items measuring either the Generate Creative Ideas or the Evaluate and Improve Ideas facets, coders were required to use a second digit to indicate the primary theme of each response that earned partial or full credit.

Responses that received partial credit could only use values of 1-3 as the second digit to represent the preliminary conventional themes chosen based on available student responses (11, 12, or 13); however, responses that received full credit could use up to 9 different values for the second digit, with the ninth value representing all themes not associated with themes 1-8 (i.e., 21 through 29). The resulting data informed distinctions between "conventionality" and "unconventionality" of themes across a diverse international student cohort.[1]

Inter-rater reliability (IRR) on anchor responses across all items and coder pairs was 0.71. the average quadratic Kappa was also relatively high (0.79). Items were reviewed for the item category response functions, item quality. Items that exhibited high omit and not-reached rates were reviewed to rule out technical issues with the platform. Cluster placement was also considered to be a contributing factor when exploring reasons for high rates of omission or not-reached coding. Items were further analysed for item difficulty, item discrimination, response time, position effect, IRT scaling, Item model fit, IRT parameters

and student theta estimates, evaluation of subscores on domain and facet levels, and differential item functioning via the item-total score curves from different country-by-language groups. The findings supported (a) the psychometric quality of PISA Creative Thinking assessment units in terms of validity, reliability, and comparability across participating countries; (b) the ability to construct a Creative Thinking scale; and (c) the inclusion of 20 of the 21 the Creative Thinking units in the 2022 Main Survey. For details of the findings please refer to the PISA 2022 CT Field Trial Research Report (ACT, 2021[5]).

Field Trial Coder Queries

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 queries to the service so that a common adjudication process was consistently applied to all coder questions about constructed-response items. Core B3 test developers and performance scoring experts from ACT reviewed and responded to queries specific to the Creative Thinking test developers.

In addition to responses to new queries, Core B3 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 following the 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 the 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 B3. Following 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 reviewed by national teams. As was the case in Phase 1, countries were asked to provide comments about these specific items where they could identify serious errors. Requests for corrections were reviewed by Core B3 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. 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, item omit and not-reached rates, item difficulty, item discriminations, Item response time, position effect, IRT scaling, item model fit, IRT parameters and student theta estimates, evaluation of sub-scores on domain and facet levels and differential item functioning (DIF).

Flagged items were further reviewed in terms of their sample size, contents, translations, and coding guides (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, planned optimizations for that item (e.g., theme changes, coding optimizations, cluster placement). Due to the operational timeline in PISA 2022, it was not possible to include new items in the test after this phase, and no substantial modifications were made to existing test items. Poorly performing items were removed from the test item pool provided coverage of the domain was

not affected significantly. For the Creative Thinking test, one unit consisting of two items was removed. The PISA 2021 Field Trial also generated insights for further enrichment of the coder training materials, including the coding guide, towards the 2022 Main Survey. Substantial work was undertaken, including reviewing large amounts of student responses, additional frequency analysis of themes, and identification of 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 conventional and unconventional themes, refinement of theme descriptions, increased representation of exemplar responses, and edits to item-specific instructions to facilitate effective and consistent coding (see Table 4.3).

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. 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 B3 contractors worked with countries to resolve any remaining issues and prepare the national instruments for the main survey.

Item selection

The PISA 2022 Field Trial provided evidence in support of the psychometric quality of PISA CT assessment units in terms of validity, reliability, and comparability across participating countries. Improvements in performance for the 20 units included in the Main Survey are anticipated based on optimizations to the coding guide, coder trainings, and cluster arrangements. 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 organized into five mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 2 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking within the two thematic content areas and the four domains.

Review by the Creative Thinking Expert Group

The Creative Thinking Expert Group reviewed the pilot study data, the approach to item selection, the content and balance of the clusters, and signed off on the selection.

Preparation of data collection instruments

Preparing the main survey national SDS

The process for creating the main survey national SDS followed the approach used during the field trial, beginning with 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 A. 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. PISA 2022 CT was only administered on computers.

Main survey coding

Main Survey Coder Training

The Main Study International Coder Training for Creative Thinking was held in February 2022. Analysis of Field Trial responses and coder queries helped Performance scoring experts from ACT improve upon online coding training modules and other coder training materials. Additional sample responses were included in the coding guide to better illustrate different types of responses. Workshop materials were also enhanced to include additional authentic student responses that better illustrate the boundaries between full credit, partial credit (where appropriate) and no credit.

The process used for the Main Survey International Coder Training was similar to the 2021 Field Trial International Coder Training in that self-guided 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, and guidelines for applying special codes, and workshop materials for each item. Following the international coder training, additional revisions were made to the Creative Thinking Coding Guide in response to discussions that took place at the meeting.

Main Survey Coder Queries

The coder query service was again used in the Main Survey as it had been 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 B3 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.

References

ACT (2021), *PISA 2022 Creative Thinking Field Trial Research Report*, ACT, Iowa City, IA. [5]

ACT (2020), *PISA 2022 Creative Thinking Limited Field Trial Research Report*, ACT, Iowa City, IA. [4]

ACT (2019), *PISA 2021 Creative Thinking Validation Study Research Report*, ACT, Iowa City, IA. [3]

ACT (2018), *PISA 2021 Creative Thinking Cognitive Lab Research Report*, ACT, Iowa City, IA. [2]

OECD (2019), *PISA 2021 Creative Thinking Framework (Third Draft)*, OECD, Paris, https://www.oecd.org/pisa/publications/PISA-2021-creative-thinking-framework.pdf. [1]

Notes

1. The conventionality or unconventionality of responses was determined by the originality of the response amongst those in the entire pool of responses (OECD, 2019[1]).

Chapter 4 tables

Table 4.1. Distribution of items by Facet and Domain

	Facet		
Domain	Generate Diverse Ideas	Generate Creative Ideas	Evaluate & Improve Ideas
Visual Expression	2	2	4
Written Expressions	4	6	2
Social Problem Solving	4	3	3
Science Problem Solving	4	1	3

Table 4.2. Creative Thinking Assessment Field Trial item distribution by facet, unit, and domain

		Facet		
Domain	Unit	Generate Diverse Ideas	Generate Creative Ideas	Evaluate and Improve Ideas
Visual	Unit 1		X	X
	Unit 2	X		X
	Unit 3		X	X
	Unit 4	X		X
Written	Unit 5	X	X	
	Unit 6	X	X	
	Unit 7	X	X	X
	Unit 8	X	X	X
	Unit 9		X	
	Unit 10		X	
Social	Unit 11	X	X	X
	Unit 12	X	X	
	Unit 13	X		X
	Unit 14	X		
	Unit 15			X
	Unit 16		X	
Science	Unit 17	X		
	Unit 18			X
	Unit 19	X	X	
	Unit 20	X		X
	Unit 21	X		X

Table 4.3. Creative Thinking Assessment Main Study item distribution by facet, unit, and domain

		Facet		
Domain	Unit	Generate Diverse Ideas	Generate Creative Ideas	Evaluate and Improve Ideas
Visual	Unit 1		X	X
	Unit 2	X		X
	Unit 4	X		X
Written	Unit 5	X	X	
	Unit 6	X	X	
	Unit 7	X	X	X
	Unit 8	X	X	X
	Unit 9		X	
	Unit 10		X	
Social	Unit 11	X	X	X
	Unit 12	X	X	
	Unit 13	X		X
	Unit 14	X		
	Unit 15			X
	Unit 16		X	
Science	Unit 17	X		
	Unit 18			X
	Unit 19	X	X	
	Unit 20	X		X
	Unit 21	X		X