Measuring Scientific Creativity with Subjective Creativity Assessments: Psychometric Challenges and Suggestions[*]

Sujie Yang[1] and Jue Wang[1]

[1] University of Science and Technology of China, Hefei, Anhui, 230026, P.R. China
juewang01@ustc.edu.cn

Abstract. The research on scientific creativity plays an increasingly important role in education. Scientific creativity can be viewed as a type of domain-specific creativity. Subjective creativity assessments (SCA) are particularly useful for measuring domain-specific creativity and earned wide popularity in evaluating product-based creativity. However, under the psychometric framework of examining rating scores, the application of SCA has several challenges, including (a) a lack of clarity and consistency among raters on the understanding and rating criteria of scientific creativity, (b) varying levels of reliability and validity across studies, and (c) insufficient evidence for supporting fairness and comparability arguments. In this study, we address several issues related to raters and rating scores with the use of SCA, and provide a psychometric framework for evaluating the quality of ratings and assessment design of SCA. Suggestions for psychometric analysis for creating objective and fair measurement of scientific creativity are discussed.

Keywords: psychometric theory, scientific creativity, subjective creativity assessment, Rasch measurement theory, unfolding theory

1 Introduction

Scientific creativity can be defined as using domain-relevant knowledge and skills to generate novel products with scientific usefulness and significance, as well as scientific values to society, for example, see [1], [2], [3], [4]. It can be viewed as domain-specific creativity in solving science-based tasks, e.g., generating hypotheses and experiments, analyzing and evaluating scientific work [5]. Previous studies showed that scientific creativity is related to scientific reasoning and general intelligence, which are particularly important for developing talents in science, technology, engineering, and mathematics fields [6]. Scientific creativity empowers students to pursue science as a potential career path and encourages them to creatively apply scientific methods to

[*] Correspondence should be addressed to Jue Wang, Department of Psychology, University of Science and Technology of China, 1129 Huizhou Blvd, Baohe District, Hefei, Anhui, 230026 China. Email: juewang01@ustc.edu.cn

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*,
Atlantis Highlights in Social Sciences, Education and Humanities 23,
https://doi.org/10.2991/978-94-6463-494-5_4

solve real-world social and environmental problems [7]. The future of our society partly depends on how we identify and educate scientifically talented students [5]. Therefore, assessing scientific creativity and developing objective measurement tools are crucially important.

The measurement of scientific creativity is different from measuring domain-general creativity. The domain-specific and scientific knowledge plays significant roles in producing creative ideas and products [3]. Meanwhile, the science tasks may also require creativity to generate solutions, such as designing chemical experiments, solving math problems, and communicating science ideas [8], [9]. Traditional creativity assessment tools (e.g., divergent and convergent thinking tests) may not be efficient and suitable for evaluating scientific creativity. There is a call for developing creative problem-solving tasks in knowledge-rich contexts to assess domain-specific creativity [10]. These types of assessments often contain open-ended questions that do not have standard solutions.

The consensual assessment technique (CAT) developed by Amabile becomes the gold standard for assessing product-based creativity [11], which requires human raters to provide their judgments toward the creativity levels reflected by the products or ideas [12], [13]. CAT has earned wide popularity in assessing creativity in various domains [14], [15], [16], [17], and it is particularly useful for assessing scientific creativity. With the involvement of rater judgment, it has also been called the subjective creativity assessments (SCA). Within the broader areas of educational assessments, SCA can be viewed as a type of performance assessments or rater-mediated assessments [18]. The psychometric scoring of rater-mediated assessments requires that raters possess sufficient expertise in relevant domains and share a common cognitive framework and process to produce valid and consistent scores [19].

The development and application of SCA for assessing scientific creativity come with psychometric challenges related to raters. In this study, we summarize existing issues related to the scoring procedure in SCA and present a psychometric framework for examining the use of rating scores. At last, we discuss the future direction on the measurement of scientific creativity and suggest different measurement approaches for generating valid, reliable, and fair scores in SCA. Specifically, this paper is guided by the following three research questions.

1. How has scientific creativity been measured in the literature using SCA?
2. Are there any psychometric issues with the use of SCA?
3. What are the suggestions for conducting psychometric analysis of SCA scores for measuring scientific creativity?

2 The Measurement of Scientific Creativity

Over the past four decades, SCA has been widely used in measuring scientific creativity. Table 1 presents a few examples of science tasks used in previous SCA studies. For instance, Hu and Adey composed items specifically for measuring scientific creativity based on the structure of Torrance Tests of Creative Thinking that mainly examines the fluency, flexibility, and originality in thinking through designing, improving, or using science products [4], [20]. Similarly, a few other studies created tests consisting of hypothesis generation, experimental design, and improvements in science-related tasks [21], [22], [23], [24]. Kaufman, Evans, and Baer created a life science task for 4th-grade students that asks students to imagine a new animal based on the given information and describe the habitat it may live in [25]. Long created scenario-based science tasks for 6th-grade students, e.g., one task asks for solutions to get drinkable water on an unknown island, and the other asks for a survival plan given a change in the tilt of the earth [17]. These types of tasks evaluate the scientific ideas and solutions that students produce and judge their creative thinking skills through the products.

The consensual assessment technique is widely used for scoring these types of tasks in SCA. A panel of experts in the related domain produce a consensual judgment toward the creativity of a product. Once the consensus among raters is achieved, the creativity of the product can be determined in comparison to others given specific samples and context [15]. The procedure of using SCA to assess scientific creativity can be summarized in the following three steps. First, a set of creative products are collected based on one or more science tasks. Second, a group of experts with domain-relevant expertise provide their judgment independently toward the creativity level of the products. Third, all rating scores provided by experts are analysed to create objective scores and indices for evaluating the quality of scores.

Raters play a crucial role that mediates the scoring process of SCA and determines the quality of assessment results. SCA can be viewed as a type of rater-mediated assessment [26]. Rater-mediated assessment refers to those assessments that involve a group of raters to judge test-takers' performances by using a rating scale in one or more domains [27]. The scoring procedure of rater-mediated assessment can be described using a lens model approach [28]. Lens model was originally proposed for understanding human perception and judgment. Engelhard created a general lens model for rater-mediated assessment, and a more specific lens model for writing assessments with influencing factors (i.e., cues) relevant to the scoring of writing tasks [18]. Wang and Long proposed a lens model for depicting the judgmental process of raters in scoring creativity assessments, in which the latent trait of student creativity is mediated through a set of cues that may influence raters' understanding and judgment of student work [26]. Researchers found that rater expertise, domain-specific knowledge, number of raters, and task domains, may affect the final scores, for example, see [15].

3 Psychometric Framework for Subjective Creativity Assessment

The Standards for Educational and Psychological Testing [29, abbreviated Standards] provides criteria for developing and evaluating tests and testing procedures, and it establishes guidelines for assessing the validity, reliability, and fairness of test scores for the intended uses. The Standards applies to the test which is "a device or procedure in which a sample of an examinee's behavior on a specified domain is obtained and subsequently evaluated and scored using a standardized process" [29, p. 2]. The SCA for examining scientific creativity describes a procedure for obtaining rating scores toward creativity products based on a creativity test. Three fundamental issues – validity, reliability, and fairness of the score uses should be addressed for SCA. Engelhard and Wind [27] specifically discuss these three foundational areas for rater-mediated assessments and the uses of raters and rating scales. In this chapter, we briefly address each foundational area within the context of creativity assessments. A comprehensive psychometric framework is presented in Figure 1.

3.1 Validity

The standards for validity issues are regarding samples and settings used in validation. As a type of rater-mediated assessments, raters play a key role in SCA. The following standard provides a detailed guideline for describing the raters' role in the validation procedure: "Standard 1.9: When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth." [29, p.25].

More specifically, the Standards encourages integrating various sources of validity evidence into one sound argument to support the interpretations of test scores for proposed uses of tests. It emphasizes five forms of validity evidence that are based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. The evidence regarding test content is constructive as it sets the construct of the test for other forms of validity to follow. The evidence based on response processes is concerning "the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" [29, p.15]. The consensual assessment technique requires experts to judge creative products [11], so that the judgmental processes of expert raters in scoring creativity products should also be considered. The Standards elicits the need of "relevant validity evidence [that] includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores" [29, p.15]. The evidence concerning internal structure is often viewed as of paramount importance in validating the construct that is described in the test content evidence. The relations to other variables are also important to assess convergent (with the same or similar constructs) and discriminant (different constructs) evidence, test-criterion evidence (e.g., predictive and concurrent validity), as well as the generalizability of validity. Lastly, the evidence based on the consequences of testing emphasizes the importance of considering intended and unintended test score interpretations for a given use. For instance, creativity assessments are often used in selecting the gifted students, which makes it crucial to include consequential validity in the argument.

3.2 Reliability/Precision

The reliability or precision of the scores is defined "in terms of consistency over replications of the testing procedure" [29, p.35] in the Standards. Consistent scores lead to high reliability. Different sources of variations across replications can affect test-takers' scores, such as test-taker's responses, task's difficulty, and rater's scoring performance. When any of these sources of variations becomes non-negligible, the amount of variability should be examined. In classical test theory (CTT) [30], [31], reliability coefficients are defined as the ratio of true score variance to observed score variance, and these are mainly computed as the correlation between scores over replications or testing conditions. The Standards defines three broad categories of reliability coefficients: (a) alternative-form coefficients, (b) test-retest coefficients, and (c) internal-consistency coefficients. Furthermore, for rater-mediated assessments that involve rater judgments, rater consistency needs to be estimated. In generalizability theory (GT) [32], a general framework for partitioning error variances due to different error sources (e.g., tasks, occasions, and raters) is proposed, and a generalizability coefficient defines the ratio of "true" (i.e., universe) score variance to observed score variance. In Rasch measurement theory [33] and item response theory, the information function indicates the precision of measurement at each level of latent ability. In contrast to CTT- and GT-based reliability coefficients, the information function emphasizes precision and accuracy of measurement of latent abilities. In addition, Rasch measurement theory provides a reliability of separation index that can indicate the reproducibility of the latent scale. The interpretation is similar to Cronbach's alpha reflecting the degree of internal consistency, but the computation is based on the latent measures. Different reliability coefficients convey different messages and should be used and interpreted appropriately.

The Standards requires that when raters are involved in scoring test-takers' performances, "reliability/precision data should be gathered and reported for the local scoring" [29, p. 44]. For instance, reliability analysis may indicate if raters need additional training, and the examination of raters should be part of the assessment system. To properly document the reliability data, the rating designs used for collecting the scores should also be reported. Furthermore, the Standards requires multiple reliability coefficients to be estimated, as stated, "Standard 2.7: When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products." [29, p.44].

Variation may arise from task to task, rater to rater, and performance to performance. The Standards elicits the needs of estimating and reporting different error sources affecting the reliability of scores. Considering the requirements of the Standards, the generalizability model and many-facet Rasch model (MFRM) can be particularly useful for gathering the relevant reliability information. GT partitions the error variances due to different sources, while MFRM produces a reliability index for each source (or facet) based on the latent measures. Furthermore, the Standards empirically stresses that the interrater agreement does not necessarily lead to the high reliability of test-takers' scores, so that the evidence should be gathered on both.

3.3 Fairness

The latest Standards lists fairness as a separate foundational area, because "fairness to all individuals in the intended population of test takers is an overriding, foundational concern, and that common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretations" [29, p. 49]. Fairness can be viewed as a lack of measurement bias or the possession of measurement invariance across different identifiable subpopulation groups. The techniques for detecting differential item functioning (DIF) are widely used to examine if individuals at the same level differ in their probabilities of achieving a certain score, and the probabilities are found to vary as a function of their group membership. Meanwhile, differential person function can be detected across different tasks with similar features and the same difficulty level [34]. For SCA, differential rater functioning (DRF) may also occur. DRF appears when a rater judges the test-takers with the same level of creativity differentially as a function of their group memberships (e.g., gender, ethnicity, and enrolment in gifted class). The analysis of DRF may reveal potential rater bias against a particular subgroup as well as the quality of ratings toward examinee scores. Wang and Long provided an illustrative analysis of the DRF analysis for SCA [26].

3.4 Comparability

Mislevy proposed a fourth concept called comparability [35]. It is viewed as an extension of issues related to reliability that link the scores across different test forms, different subgroups of students, or different sets of raters. The discussion on comparability is particularly useful for developing an item bank with creativity tasks and supporting the research on computer adaptive testing as well as machine scoring for SCA.

4 Psychometric Challenges of Assessing Scientific Creativity Using SCA

In this chapter, we identify two major challenges in applying SCA for assessing scientific creativity under the above-mentioned framework.

The first challenge is that there is no explicit definition or detailed rating rubric to guide the scoring of creativity tasks. As indicated by Amabile, raters are encouraged to judge the creativity of a product based on their definition and criteria of creativity [11]. This procedure places more emphasis on the selection of experts who should possess domain-relevant expertise. However, many studies recruited quasi-experts (e.g., undergraduate and graduate students) for scoring the tasks and did not explicitly demonstrate their expertise in scoring a certain creativity task. Rater judgment may vary as a function of their educational background, years of scoring experience, as well as their creativity levels [17, 36, 37].

On one hand, there is a need to describe (a) the selection procedure of experts, (b) domain-relevant expertise, e.g., educational background, knowledge, and skill levels in a relevant domain area, as well as previous scoring experience on similar tasks, and (c) the scoring procedure that whether raters make their judgments independently and/or through a panel discussion. As the Standards 1.9 suggests: "When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described." [29, p. 25]. On the other hand, in order to produce consistent ratings or measures for a common latent trait, expert raters also need proper guidance, training, and monitoring before and during the scoring activities.

Secondly, the use of psychometric indices for examining the quality of creativity ratings was limited, compared to other rater-mediated assessments (e.g., language testing). Long and Wang conducted a systematic review of existing literature using SCA and indicated the following findings: (a) the most frequently obtained validity evidence is based on the relations to other variables (e.g., other creativity measures) using correlation and regression techniques, and the internal structure of the creativity scale through factor analyses; (b) there was an over-reliance on Cronbach's alpha for measuring interrater consistency; (c) empirical values of validity and reliability indices varied across different studies, especially by task domains and rater expertise levels [38]. When providing validity arguments, existing research mostly provides one or two forms of validity evidence. As elicited by the Standards, evidence of the rater's response processes is an integral part of the validity argument for rater-mediated assessments. However, very few studies examined the response process of rater scoring procedure. Meanwhile, the Standards suggests the use of reliability indices that can differentiate error sources and show consistency for each facet (e.g., raters and test-takers). Furthermore, the fairness and comparability of scores were not often discussed in SCA studies. There is a need for more research on the development and adaptation of psychometric methods for evaluating ratings of SCA.

5 Suggestions on Future Psychometric Analysis

Wang and Engelhard suggested a framework for evaluating rater-mediated assessments, and illustrated the use of this framework within the context of writing assessments [39]. We depicted the procedure in Figure 2, where it starts with a theoretical model (e.g., Lens model) for the rater scoring process, followed by the selection of an appropriate psychometric model depending on the judgmental process of rater scoring. The psychometric analysis provides quantitative evidence for calibrating and evaluating the creativity ratings. Along with qualitative methods, we can obtain a substantive interpretation of rater judgments and identify influencing factors (e.g., cues). This information can in turn improve our understanding of the theoretical model of rater judgmental process. Overall, the entire process ensures the quality of ratings and improves the assessment design for SCA.

In terms of the selection of an appropriate psychometric model for analyzing creativity ratings. Wang and Engelhard suggested two types of models, that are based on Rasch measurement theory and unfolding theory, respectively [39], [40]. They describe different rater judgmental processes. Rasch models analyze a cumulative response process that raters should all share a common mental model and possess a consistent understanding of rating categories. Unfolding models reflect an unfolding response process, where raters may have their own unique rating pattern and provide higher scores to different creative products. In rater-mediated assessments, Rasch models are suitable for analyzing impersonal judgments among raters, while unfolding models are suggested when ratings reflect individual rater preferences.

There is a wide application of Rasch models in rater-mediated assessments, especially language testing and writing assessments. Within the context of SCA, Rasch models earn the attention in recent years. Based on a search of "Rasch model" and "creativity assessment" as keywords in the Scopus database (up to 2023), it returned 12 studies, and only 8 of them are in fact relevant to SCA. Despite the limited application of Rasch models, researchers have recognized its potential and advantage for evaluating SCA, for example, see [15], [26], [41], [42].

Rasch measurement model calibrates the observed scores and places all latent measures along a common latent scale. Meanwhile, it produces reliability (e.g., reliability of separation), validity (e.g., Infit and Outfit mean square errors), fairness (interaction and bias analysis), and comparability indices like Rasch's Equating Invention in [43]. Rasch models are also useful for detecting various rater effects [44], [45].

In rater-mediated assessments, impersonal judgments are often expected in scoring activities. Raters are asked to provide ratings of student performances based on the instructions in scoring activities and the set of rubrics used to guide the assessment system. However, empirical findings indicate that despite training, human raters may still be influenced by their own characteristics and unique prior experiences.

Unfolding models can be used to quantify personal preferences and detect potential biases by raters. Unfolding models, also called ideal-point item response models, are in general used less frequently than other item response models. In writing assessments, we found 19 studies that examined rater issues using unfolding models. However, in creativity research, there was only one study that applied an unfolding model for analyzing raters' ranking data. None of the studies empirically examined ratings of SCA with a focus on rater judgmental process yet, based on unfolding models.

When raters possess their own perceptions, understandings, and preferred usages of rating categories in judging the products, the responses reflect their personal preferences. Based on an unfolding model, each rater has a unique ordering of student work depending on the distance between the rater and student locations on the underlying scale. The probability function of unfolding models is single-peaked and reflects an ideal-point location (e.g., preferred creativity level) of raters. Modeling personal preferences with an unfolding model can be a useful addition to other approaches for evaluating the quality of ratings. On one hand, unfolding measures reflect unique rater response patterns, which can better depict the scoring process with no explicit rating rubric. On the other hand, unfolding models can examine different sources of rater biases and detect raters with aberrant scoring behaviors.

6 Final Words

In this chapter, we introduced a comprehensive psychometric framework for examining rating scores within the context of creativity assessments, and identified psychometric issues related to raters and rating scores in measuring scientific creativity. Lastly, we made several suggestions on the psychometric methods to promote objective and fair measurement of scientific creativity.

In 2021, Science published an updated list of 125 questions that deserve exploration and discovery by researchers in different fields. One of the important questions is “Can robots or AIs have human creativity?”. This question was raised partly due to an astounding event that AlphaGo, an AI program created by Google, defeated a human master of the ancient game of Go. A year later, a major large language model -- ChatGPT was publicly released by OpenAI, and it has been constantly evolving. Just before the submission of this chapter, we observed the release of Sora (i.e., an artificial intelligence model that can create a minute-long realistic and imaginative scenes based on text instructions).

In the era of generative artificial intelligence, creativity researchers have been re-considering the definition of creativity [46] and asking for new assessment tools for evaluating creativity [47], [48]. Psychometricians are responsible for developing innovative measurement models that can support the evaluation of new types of creativity assessments (e.g., interactive creativity tasks) and considering different types of psychometric issues that we may encounter when assessing human and artificial creativity.

Table 1. Example science tasks used in subjective creativity assessments.

Task Description	Grade Level	Source Article
1. Please write down as many as possible scientific uses as you can for a piece of glass. For example, make a test tube. 2. If you can take a spaceship to travel in the outer space and go to a planet, what scientific questions do you want to research? Please list as many as you can. For example, are there any living things on the planet? 3. Please think up as many possible improvements as you can to a regular bicycle, making it more interesting, more useful and more beautiful. For example, make the types reflective, so they can be seen in the dark. 4. Suppose there was no gravity, describe what the world would be like? For example, human beings would be floating. 5. Please use as many possible methods as you can to divide a square into four equal pieces (same shape). Draw it on the answer sheet. 6. There are two kinds of napkins. How can you test which is better? Please write down as many possible methods as you can and the instruments, principles, and simple procedure. *7.* Please design an apple picking machine. Draw a picture, point out the name and function of each part.	Secondary school	Hu & Adey, (2002, pp. 394-395)
1. There is an animal named Zook that lives here on earth. A Zook is light in color, has big sharp teeth, and a tail. What type of animal do you think a Zook is? Why? Where do you think Zooks live? How would the habitat meet the needs of the Zook? What living and nonliving things would be in this habitat? What do you think Zooks eat? How do they get their food? *2. Now pretend that all the Zooks in the world were moved to a tropical location.* How will the Zooks' lives change? Now that the Zooks have lived in a tropical place for twenty years, what do you think that the new Zook babies will look like?	Fourth grade	Kaufman, Evans, & Baer (2010, pp.13-14)

Note. The example studies are sorted based on their publication years.

Table 1. (Continued).

Task Description	Grade Level	Source Article
1. Fly experiment. This problem presents a figure of an experiment designed by a researcher. Students are required to generate as many hypotheses as they can think of that the researcher might want to test by this experiment. 2. Change graph. This problem presents a graph of reverse changes in the amounts of two variables and an effect that starts these changes. Students are asked to think of as many pairs of variables as they can that fit the graph. 3. Sugar experiment. A figure of an experiment designed by a researcher and a graph showing the researcher's hypothesis are presented in this problem. Students are required to think of as many changes as they can that should be made in the experiment in order for the researcher to prove the hypothesis. 4. String experiment. A figure of an experiment is presented in this problem. Students are asked to think of as many changes as they can that should be made in the experiment to achieve a goal. 5. Food chain. This problem presents a figure of a food chain and a graph of the change in this food chain. Students are asked to think of as many causes as they can of the change. This problem measures fluency, flexibility, and creativity in evidence evaluation in the area of ecology.	Middle school	Sak & Ayas (2013, pp. 320-321)
In the year 2050, a meteorite narrowly brushes the earth. While a major explosion is avoided, it results in tipping the earth 75◦ more than its current 22◦ tilt. How will this change in tilt affect climate of North America and the lives of the people who live there? How will people need to adjust in order to survive (for example, food, agriculture, clothing, etc.)? Use what you have learned about season and climate to explain your ideas. Make your ideas as creative as possible.	Sixth grade	Long (2014, p.192)

Note. The example studies are sorted based on their publication years.

Table 1. (Continued).

Task Description	Grade Level	Source Article
Technical Product (Item 1) Suggest as many scientific improvements to a pen (Form A) /whiteboard pen (Form B) to make it look interesting, unusual and no need to be practical. You can show your idea using a drawing. Science Knowledge (Item 2) Write down as many scientific words as you know about 'magnet' (Form A) /'microorganisms' (Form B) Science Phenomena (Item 3) Write as much as possible in an interesting scientific story to imagine the following topics: 1) The sun is losing its light (Form A); 2) Plants can move like animals (Form B). Science Problem (Item 4) By using as many methods as possible, divide a square into 4 equal parts (same form). Show your answer using a drawing (Form A). By rearranging or removing matchsticks of the following symbols, create as many symbols as possible by using 5 matchsticks (Form B).	Fifth grade	Siew, Chong and Chin (2014, pp. 113-114)
A plastic bottle cap floats on the water surface. How do you make the plastic bottle cap sinks in the water? You found a magnet on the floor. Think as many materials or objects as possible that are not attracted to the magnet. A glass of orange juice spilled on the floor. Show in your drawing on how the orange juice can be dried up. While walking to school. Ali saw shadows around him. Show how shadows are formed. You have mixed some materials and objects in a basin filled with water. Think as many materials or objects as possible that are not soluble in the water after stirring with spoon. You have mixed salt with the sand. How do you separate the sand from the mixture?	Kindergarteners, Preschool	Chin and Siew (2015, pp. 1395)

Note. The example studies are sorted based on their publication years.

Table 1. (Continued).

Task Description	Grade Level	Source Article
1. Hypothesis generation subtests Item 1 shows a child who goes by a swamp. A sudden question comes to his mind about the life of flies in the swamp. Children are asked to generate many ideas (hypotheses) related to the question. Item 2 shows that two children are drinking water from their water bottles after they get tired. They realize that water in the two bottles have different temperatures. Students are asked to generate many ideas (hypotheses) as causes of the difference in water temperature. Item 3 shows a mother presents a problem situation related to a toy ship to her daughter. Students are asked to generate as many ideas (hypotheses) as they can think of that the girl can think of. 2. Experiment design subtests Item 1 shows that a child and his father are preparing a living area for hamsters in an animation. The father indicates some problems in the hamsters' living area and asks his son to make changes in the living area so that the hamsters can live there. Students are asked to generate as many changes as they can think of that the child could do. Item 2 shows that a child is playing with a ball on a sand pool in an animation. He wants to make some changes in the sand pool to achieve a goal. Students are asked to find as many changes as they can think of that the child can do. Item 3 shows that a child and her aunt are making a setup with a toy car and a tunnel. They cannot achieve their goal with the setup, and must make some changes in the setup to accomplish their goal. Students are asked to find as many changes as they can think of that the child and the aunt can do to achieve their goal.	Kindergarten to second-grade students	Atesgoz & Sak (2021, pp.3)

Note. The example studies are sorted based on their publication years.

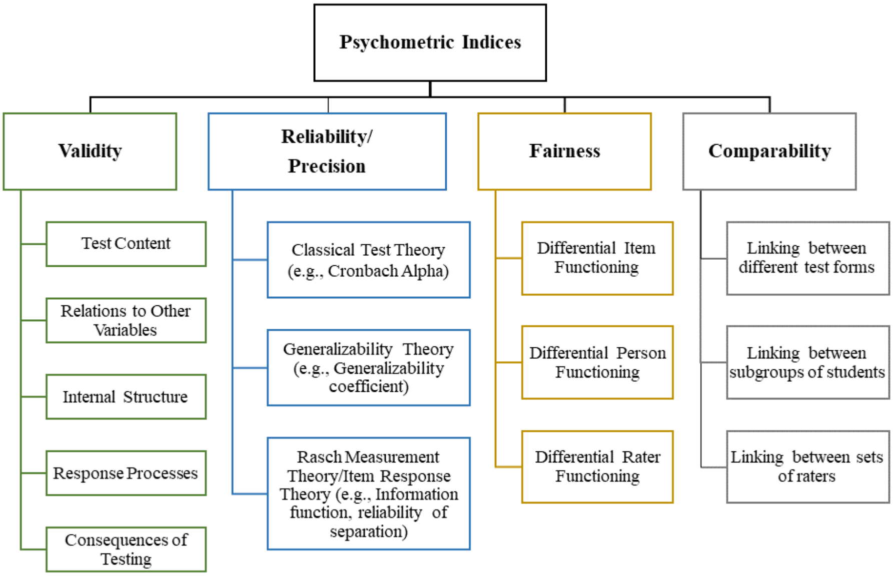
Psychometric Indices
Validity
Test Content
Relations to Other Variables
Internal Structure
Response Processes
Consequences of Testing
Reliability/ Precision
Classical Test Theory (e.g., Cronbach Alpha)
Generalizability Theory (e.g., Generalizability coefficient)
Rasch Measurement Theory/Item Response Theory (e.g., Information function, reliability of separation)
Fairness
Differential Item Functioning
Differential Person Functioning
Differential Rater Functioning
Comparability
Linking between different test forms
Linking between subgroups of students
Linking between sets of raters

Fig. 1. Psychometric framework for evaluating subjective creativity assessments.

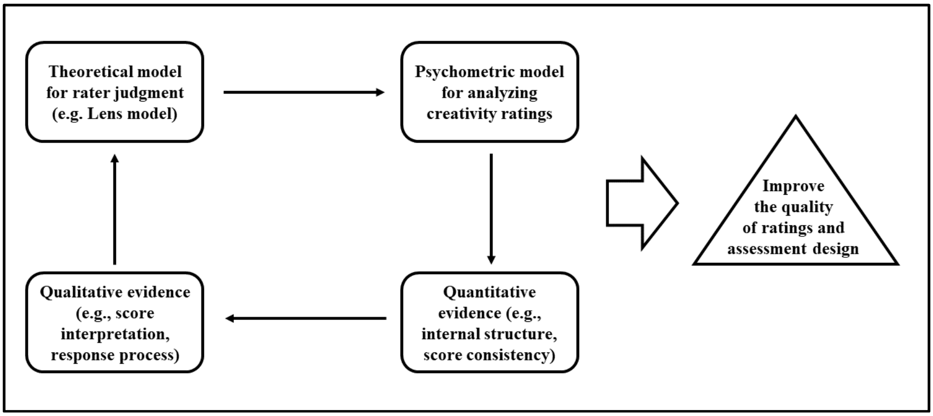
Theoretical model for rater judgment (e.g. Lens model)
Psychometric model for analyzing creativity ratings
Quantitative evidence (e.g., internal structure, score consistency)
Qualitative evidence (e.g., score interpretation, response process)
Improve the quality of ratings and assessment design

Fig. 2. Psychometric framework for evaluating subjective creativity assessments.

Note. Adapted from Wang and Engelhard (2019, p.583).

Acknowledgements

This work was supported by National Key R&D Program of China (2023YFC3341302).

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Runco, M. A. (2003). Education for creative potential. Scandinavian Journal of Educational Research, 47(3), 317-324.
2. Simonton, D.K. (1990). History, chemistry psychology, and genius: An intellectual autobiography of historiometry. In M.A. Runco & R.S. Albert (Eds.), Theories of creativity (pp. 92–115). Newbury Park, CA: Sage.
3. Ayas, M. B., & Sak, U. (2014). Objective measure of scientific creativity: Psychometric validity of the Creative Scientific Ability Test. Thinking Skills and Creativity, 13, 195-205.
4. Hu, W., & Adey, P. (2002). A scientific creativity test for secondary school students. International Journal of Science Education, 24(4), 389-403.
5. Sternberg, R. J. (2018). Direct measurement of scientific giftedness. Roeper Review, 40(2), 78-85.
6. Sternberg, R. J., Todhunter, R. J. E., Litvak, A., & Sternberg, K. (2020). The Relation of Scientific Creativity and Evaluation of Scientific Impact to Scientific Reasoning and General Intelligence. Journal of Intelligence, 8(17).
7. Watson, M., & McMahon, M. (2020). Career assessment and creativity: potential complementarity or a contradiction in terms? British Journal of Guidance & Counselling, 48(1), 40-51.
8. Hernández-Torrano, D., & Ibrayeva, L. (2020). Creativity and education: A bibliometric mapping of the research literature (1975–2019). Thinking skills and creativity, 35, 100625.
9. Huang, C. F., & Wang, K. C. (2019). Comparative analysis of different creativity tests for the prediction of students' scientific creativity. Creativity Research Journal, 31(4), 443-447.
10. Yang, W., Green, A., Chen, Q., Kenett, Y., Sun, J., Wei, D., & Qiu, J. (2022). Creative problem solving in knowledge-rich contexts. Trends in Cognitive Sciences. 26.
11. Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. Journal of personality and social psychology, 43(5), 997.
12. Baer, J. (2010). Is creativity domain specific? The Cambridge handbook of creativity, 321-341.
13. Carson, S. H. (2006, April). Creativity and mental illness. Invitational panel discussion hosted by. Yale's Mind Matters Consortium.
14. Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research:(Re) setting the standards. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 233.
15. Cseh, G. M., & Jeffries, K. K. (2019). A scattered CAT: A critical evaluation of the Consensual Assessment Technique for creativity research. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 159–166.
16. Long, H. (2014a). An empirical review of research methodologies and methods in creativity studies (2003-2012). Creativity Research Journal, 26(4), 427–438.
17. Long, H. (2014b). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. Thinking Skills and Creativity, 13, 183–194.

18. Engelhard Jr, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. Routledge.
19. Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system, Assessing Writing, 4(1), pp. 83-106.
20. Torrance, E. P. (1966). Torrance tests of creative thinking. Personnel Press, Princeton, N.J.
21. Sak, U., & Ayas, M. B. (2013). Creative Scientific Ability Test (C-SAT): A new measure of scientific creativity. Psychological Test and Assessment Modeling, 55(3), 316-329.
22. Atesgoz, N. N., & Sak, U. (2021). Test of scientific creativity animations for children: Development and validity study. Thinking Skills and Creativity, 40, 100818.
23. Siew, N. M., Chong, C. L., & Chin, K. O. (2014). Developing a scientific creativity test for fifth graders. Problems of Education in the 21st Century, 62, 109.
24. Chin, M. K., & Siew, N. M. (2015). The development and validation of a figural scientific creativity test for preschool pupils. Creative Education, 6(12), 1391.
25. Kaufman, J. C., Evans, M. L., & Baer, J. (2010). The American Idol Effect: Are students good judges of their creativity across domains? Empirical studies of the arts, 28(1), 3-17.
26. Wang, J., & Long, H. (2022). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. Psychology of Aesthetics, Creativity, and the Arts.
27. Engelhard, G., & Wind, S. A. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge.
28. Brunswick, E. (1952). The conceptual framework of psychology. University of Chicago Press.
29. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). The standards for educational and psychological testing. AERA, APA, and NCME.
30. Spearman, C. (1904a). The proof and measurement of association between two things. American Journal of Psychology, 15, 72–101.
31. Spearman, C. (1904b). General intelligence, objectively determined and measured. American Journal of Psychology, 15, 201–292.
32. Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley, New York.
33. Rasch, G. (1960). Probabilistic models for intelligence and attainment tests. University of Chicago Press.
34. Engelhard, G. (2009). Using Item Response Theory and Model—Data Fit to Conceptualize Differential Item and Person Functioning for Students With Disabilities. Educational and Psychological Measurement, 69(4), 585-602.
35. Mislevy, R. J. (2018). Sociocognitive foundations of educational measurement. Routledge.
36. Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. Thinking Skills and Creativity, 15, 13–25.
37. White, A., Shen, F., & Smith, B. L. (2002). Judging advertising creativity using the creative product semantic scale. The Journal of Creative Behavior, 36(4), 241-253.
38. Long, H., & Wang, J. (2022). Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. Educational psychology review, 34(3), 1399-1443.
39. Wang, J. & Engelhard, G. (2019a). Conceptualizing rater judgments and rating processes for rater-mediated assessments. Journal of Educational Measurement, 56(3), 582-609.
40. Coombs, C. H. (1964). A theory of data. New York, NY: Wiley.

41. Hung, S. P., Chen, P. H., & Chen, H. C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. Creativity Research Journal, 24(4), 345-357.
42. Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 176.
43. Wright, B. D., & Grosse, M. (1993). How to set standards. Rasch Measurement Transactions, 7(3), 315-316.
44. Engelhard Jr, G. (1996). Clarification to “Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model”. Journal of Educational Measurement, 33(1), 115-116.
45. Wolfe, E. W. (2014). Methods for Monitoring Rating Quality: Current Practices and Suggested Changes. Iowa City, IA: Pearson Inc.
46. Runco, M. (2023). Updating the Standard Definition of Creativity to Account for the Artificial Creativity of AI. Creativity Research Journal. 1-5.
47. Acar, S. (2023). Creativity Assessment, Research, and Practice in the Age of Artificial Intelligence. Creativity Research Journal. 1-7.
48. Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. Nature Human Behavior, 7, 1836–1838.