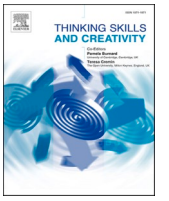
Thinking Skills and Creativity 43 (2022) 100994

Contents lists available at ScienceDirect

Thinking Skills and Creativity

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



Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks☆

Mare Van Hooijdonk [a,*], Tim Mainhard [b], Evelyn H. Kroesbergen [a], Jan Van Tartwijk [c]

[a] Radboud University, Behavioural Science Institute, PO Box 9104, 6500 HE Nijmegen, The Netherlands
[b] Leiden University, Educational Sciences, PO Box 9555, 2300 RB Leiden, The Netherlands
[c] Utrecht University, Department of Education & Pedagogy, PO Box 80140, 3508 TC Utrecht, The Netherlands

ARTICLE INFO

Key words:
Assessment
Creative problem solving
Creativity
Generalizability theory
G-study
D-study

ABSTRACT

The assessment of creative problem solving (CPS) is challenging. Elements of an assessment procedure, such as the tasks that are used and the raters who assess those tasks, introduce variation in student scores that do not necessarily reflect actual differences in students' creative problem solving abilities. When creativity researchers evaluate assessment procedures, they often inspect these elements such as tasks and raters separately. We show the use of Generalizability Theory allows researchers to investigate creativity assessment procedures - and CPS assessments in particular - in a comprehensive and integrated way. In this paper, we first introduce this statistical framework and the choices creativity researchers need to make before applying Generalizability Theory to their data. Then, Generalizability Theory is applied in an analysis of CPS assessment tasks. We highlight how alterations in the nature of the assessment procedure, such as changing the number of tasks or raters, may affect the quality of CPS scores. Furthermore, we present implications for the assessment of CPS and for creativity research in general.

1. Introduction

The assessment of creativity is a challenge (Cropley, 2000; Piffer, 2012; Sternberg, 2020). Researchers develop creativity assessment procedures to measure creativity, for instance to determine students' creative abilities in educational settings. A creativity assessment procedure includes the full range of chosen methods for evaluating a student's creative performance (Gipps, 1994). This could for instance mean a variation of tasks, different raters or rating procedures (see e.g., Bouwer, Béguin, Sanders & van den Bergh, 2015; Crossley, Johnson, Booth & Wade, 2011).

However, these elements of an assessment procedure usually also introduce variation in scores which does not necessarily reflect actual differences in students' creativity. Specifically for the assessment of creativity, creativity task characteristics, rater biases in, for example, originality scores, and the chosen rating procedures are all examples of potential sources of variability in creativity scores, which are not due to actual differences between students' creativity (Amabile & Pillemer, 2012; Conti, Coon & Amabile, 1996; Guo, Ge & Pang, 2019; Kaufman, Baer, Cole & Sexton, 2008; Reiter-Palmon, Illies, Cross, Buboltz & Nimps, 2009). For creative problem solving

This research was made possible by funding from the Dutch Ministry of Education, Culture and Science, OCW/PromoDoc/1065001. The data, syntax, and output that support the findings are available online via https://doi.org/10.34894/GIUAOI.

* Corresponding author.
E-mail address: mare.vanhooijdonk@ru.nl (M. Van Hooijdonk).

https://doi.org/10.1016/j.tsc.2021.100994
Received 7 April 2021; Received in revised form 9 December 2021; Accepted 12 December 2021
Available online 15 December 2021

(CPS) in particular, determining these potential sources of variability might be of interest. CPS is a measure of little-c, that is, everyday creative achievement, in which students use subject knowledge and creativity to solve everyday problems creatively (Craft, 2001; Isaksen & Treffinger, 2004; Okuda, Runco & Berger, 1991). The problems students solve may stem from different domains and students may as such have different subject knowledge or previous experiences with the problem, impacting the generalizability of scores on CPS tasks (Reiter-Palmon et al., 2009). Raters on the other hand might have more affinity with a certain problem or might have preliminary ideas about what solutions could or could not work (Long, 2014). Again, this may impact CPS ratings and consequently the generalizability of students' CPS scores. In divergent thinking studies, ideas are most often assessed on fluency, flexibility, and originality (Reiter-Palmon, Forthmann & Barbot, 2019). In open-ended tasks like CPS tasks, raters often apply more criteria such as how appropriate (or practical) and thoughtful (or complete) ideas are (Long, 2014). Reviewing how generalizable CPS scores for these criteria are, may give us insight in specific biases in CPS assessments and how these differ from biases in divergent thinking tasks (see e.g., Hass, Rivera & Silvia, 2018; Silvia et al., 2008). Knowing to what extent elements of a CPS assessment procedure and their interactions cause variance in addition to actual differences in students' CPS ability helps interpreting students' scores. Furthermore, it points to elements of the CPS assessment procedure that potentially require improvement.

Besides determining the influence of the separate elements of the CPS assessment procedure, assessment developers also want to determine the quality of the assessment procedure as a whole. Calculating to what degree the assessment procedure produces reliable scores for the constructs of interest assists in determining this psychometric quality. When there is a lot of "noise" in student scores, adaptations in the assessment procedure, such as more problem tasks or more raters, might be needed to decrease its impact on the reliability of the results. However, more problem tasks and raters do also implicate using more time and resources. Especially when the aim is to embed the assessment of CPS in the curriculum, procedures need the be as efficient as possible. Thus, neither too many nor too few problem tasks and raters should be used. *Generalizability Theory* (Brennan, 2001) provides a statistical framework which assists in making these informed decisions about assessment procedures. Several authors presented a general overview of this theory and how it can be applied (see e.g., Brennan, 2010; Mushquash & O'Conner, 2006). In this paper, we want to complement this work by offering a non-technical manual specifically in the CPS context. After introducing Generalizability Theory, including an explanation of the decisions researchers need to make, we will demonstrate how Generalizability Theory can be applied on CPS data with a step-by-step manual. By providing information on the statistical formulae, the data structure, the syntax, and the output, we invite creativity researchers to try Generalizability Theory themselves and experience how it may enhance their development and evaluation of creativity assessment procedures and CPS assessments in particular.

1.1. Generalizability theory

Generalizability Theory provides a statistical framework for investigating and developing reliable assessment procedures. Generalizability is regarded as a form of reliability or reproducibility that is important for estimating to what extent observations can be used to make claims about a student's true ability (Brennan, 2010). In classical test theory, as opposed to Generalizability Theory, measurement error caused by multiple sources is regarded as random and intertwined; Classical test theory does not allow for one to disentangle the sources of error. Generalizability Theory, however, determines the sources of error variation, disentangles them, and estimates the error for each source. By reviewing this error, Generalizability Theory inspects how consistent elements of an assessment procedure, such as tasks and raters, 'behave' and transforms this behavior into a coefficient. These coefficients serve three purposes: (1) they can be reviewed to determine the impact of one of the elements in an assessment procedure, (2) they can be used to calculate the generalizability of scores produced by the assessment procedure as a whole, and (3) they can be used to calculate how potential changes in the assessment procedure would impact the generalizability of the scores. Generalizability Theory proposes two stages of analysis: the Generalizability study (G-study) and the Decision study (D-study). Where the G-study estimates the error attributed to the elements in the assessment procedure and calculates current generalizability, the D-study uses the estimated errors to predict how alterations in the assessment design (e.g., more or less tasks or raters) would influence generalizability.

1.1.1. G-study

In the G-study stage, error (i.e., variability in student scores which does not originate from differences in student creative ability) attributed to the separate elements of the assessment procedure, is identified and the relative importance of all elements is estimated. The magnitude of the variance each aspect (and the interaction of aspects) introduced to the test score is reflected in a variance component. The amount of variance that can be rightly attributed to students' true ability can be reviewed here by reviewing the variance component of the students. Besides this, the impact of the other elements, such as tasks, raters or rating procedures and their interactions, can be reviewed and compared as well. Within a G-study, the generalizability of a certain score (e.g., a creativity score) is described in a coefficient that is calculated based on the variance components. This generalizability coefficient reflects the accuracy of generalizations made from the *observed* scores compared with the *universe* score. A universe score is the expected value of a student's observed score over all tasks to which an assessment procedure wants to generalize. By reviewing this generalizability coefficient, the researcher can gain insight in the psychometric qualities of the procedure. Often, a generalizability of 0.70 or higher is regarded as acceptable when we aim to measure abilities such as creativity (Brennan, 2010).

1.1.2. D-study

In the second stage, multiple Decision-studies (D-studies) can be carried out. The ultimate goal of a D-study is to increase the psychometric quality and efficiency of an assessment procedure by choosing the optimal number of elements (e.g., tasks and raters) for an assessment procedure to produce generalizable results in a specific context. This can be done by calculating the generalizability for

hypothetical numbers of elements using the variance components obtained in a G-study (see e.g., Hass et al., 2018). In this way, it is explained how alterations in the assessment procedure (e.g., reducing or extending the number of creativity tasks involved) will affect the score generalizability. Multiple D-studies can be used to estimate the number of tasks and/or raters that is needed for a desired level of generalizability.

1.2. Choices to consider

Before applying Generalizability Theory, researchers have some choices to make. First they need to determine what the universe is they aim to generalize to (Brennan, 2010). In other words, researchers need to describe what specific ability or trait they aim to measure. Especially in light of the discussion on the task or domain specificity of creativity, this is an important aspect to consider (see e.g., Baer, 2012; Barbot, Besançon & Lubart, 2016; Reiter-Palmon et al., 2009). Researchers could for instance decide to generalize to a specific creative process across domains (e.g., problem construction or idea selection; Arreola & Reiter-Palmon, 2016; Zhu, Ritter & Dijksterhuis, 2020) or to a creative ability within a more specific domain such as musical or mathematical creativity (Barbot & Lubart, 2012; Schoevers, Kroesbergen & Kattou, 2020). The chosen universe determines the focus of the assessment procedure and the nature of the elements involved.

Next, researchers need to decide what elements of the assessment procedure to include into their G-study. In Generalizability Theory, the elements of the assessment procedure are usually called *facets.* Facets are similar to factors used in an Analysis of Variance (ANOVA; Brennan, 2010). Fig. 1 shows a conceptual model of possible facets of a creativity assessment procedure and the interactions between these facets.

Facets may for instance include tasks, domains, raters, or measurement moments. The researcher selects those facets that may impact test scores the most. For instance, what rating procedure is used may impact scores more than test settings such as whether the tasks are administrated in or outside of class. Here, it is also important to consider the aspect of time and resources. Researchers need to consider issues like how many tasks would be feasible in a given timeframe and how many raters can be involved, both now and in future assessment. In an analysis based on Generalizability Theory, researchers can choose to treat facets as fixed or random. Cronbach, Gleser, Nanda and Rajaratnam (1972) advise to apply a fixed facet when (1) the test developer deliberately includes certain facets and is not interested in generalizing beyond them or beyond the observed levels or (2) when the number of potential levels or variations of a factor is that small that all levels are already included in assessment procedure. In other words, when facets are treated as fixed, the derived inferences are limited to the facets in an assessment procedure, such as specific tasks or types of scoring (Silvia et al., 2008). Researchers may treat facets as random if they, for instance, aim to apply many similarly qualified raters or when they regard the tasks as a selection of many possible tasks. This may for instance be the case when (trained) teachers assign scores in the assessment procedure or when classroom projects are rated to obtain creativity scores. In other words, facets are regarded as random when there is a theoretically infinite pool of tasks and raters. If we would randomly select a set of tasks or raters, they should produce roughly the same observed scores as another randomly selected set (Silvia et al., 2008).

Besides deciding what facets to include, researchers ideally also decide beforehand if they want to make *absolute* or *relative* decisions in the assessment procedure. For absolute decisions, such as assigning IQ scores, the aim is to index the individual student's

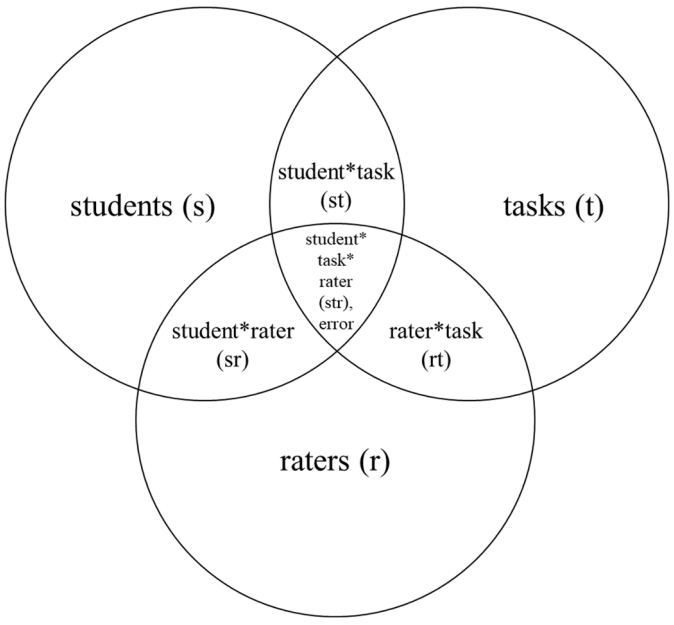


Fig. 1. Example of the facets of an assessment procedure and the interactions.

absolute level of ability, independent of the performance of others. Test developers usually aim for absolute decisions when they develop, for example, a standardized creativity test. A student's performance has improved when this student gets a higher absolute score based on the results of the creativity test, compared with the score this student got based on the results of an earlier test. When researchers decide to make relative decisions, they are merely interested in how a student is positioned compared with other members of a group. These creativity tests aim to measure, for example, whether a student has become more creative over time compared with the rest of this student's class. This may for instance mean a student receives a lower absolute rating on a second test, but because on average the group performed even worse, the student's performance might still have improved. Generalizability Theory provides generalizability coefficients for absolute and relative decisions separately and calculates these with the use of the variance components.

When researchers aim to make absolute decisions, all variance components, including the main variance components of the raters, tasks, and the rater-task interaction, are included in the calculation of the absolute generalizability (see formula 1). This makes the absolute D-coefficient more stringent than the relative g coefficient. With the relative g coefficient, we are interested in the relative standing of students. As such, the g-coefficient only considers variance components (partially) attributed to the students (formula 2). Because the absolute ratings and scores on tasks do not matter here, the main variance components and the interaction of the tasks and raters are not included in the calculation of the relative g coefficient. The formulae below need to be adopted to the specific assessment design (e.g., Fig. 1). For instance, when besides tasks and raters different rating procedures (p) are applied, this is added to the equation as a fourth facet (e.g., sp/n for relative generalizability).

$$\text{Absolute generalizability (D)} = \frac{s}{\mathrm{s} + \mathrm{t/n} + \mathrm{r/n} + \mathrm{st/n} + \mathrm{sr/n} + \mathrm{rt/n} + \mathrm{str,e/n}} \quad (1)$$

$$\text{Relative generalizability (g)} = \frac{s}{\mathrm{s} + \mathrm{st/n} + \mathrm{sr/n} + \mathrm{str,e/n}} \quad (2)$$

where

s = variance component of students
t = variance component of tasks (not included when making relative decisions)
r = variance component of raters (not included when making relative decisions)
st = variance component of student x task
sr = variance component of student x rater
rt = variance component of rater x task (not included when making relative decisions)
str,e = student x rater x task interaction and error (confounded)
n = amount of tasks and/or raters.

As a last decision, researchers need to determine what software to use to perform the statistical analyses. Programs that allow researchers to apply Generalizability Theory are inter alia IBM SPSS Statistics, SAS and RStudio (Jiang, 2018; Mushquash & O'Conner, 2006). Because our analysis of CPS tasks includes a relatively straightforward statistical model, IBM SPSS Statistics was used to execute the Generalizability study. SPSS is a widely used program to manage data and the provided syntax (see Appendix A) allows developers of creativity assessments to quickly perform a generalizability study, for instance in the piloting phases of their research. Especially for more complex models such as multivariate models, RStudio is considered to be a suitable alternative (see e.g., Jiang, Raymond, Shi & DiStefano, 2020).

1.3. Generalizability theory in creativity research

Most researchers investigate the elements from a creativity assessment procedure separately (e.g., Benedek et al., 2016; Ceh, Edelmann, Hofer & Benedek, 2021; Reiter-Palmon et al., 2009), for instance, by focusing on what kind of tasks should be included or how many raters would be necessary to obtain a reliable indication of one's creativity. However, in such research, it is not taken into account how one element (e.g., raters) may interact with other elements (e.g., tasks or students) and how this affects the generalizability of creativity scores. In a variety of educational domains other than creativity, researchers have used Generalizability Theory to investigate assessment procedures. For instance, Bouwer et al. (2015) added genre as a facet and studied the generalizability of writing scores. Crossley et al. (2011) and Moonen-van Loon, Overeem, Donkers, Van der Vleuten and Driessen (2013) compared different workplace-based assessment scales to determine which scales produces generalizable results in a limited number of assessments. Hill, Charalambous and Kraft (2012) applied Generalizability Theory to data generated with a teaching observation tool to illustrate how one can determine the number of raters or lessons that need to be included to determine a teacher's teaching quality.

In creativity research, only a few researchers have applied Generalizability Theory to their data. Or as Myszkowski and Storme (2019) put it: "in creativity research Generalizability Theory is certainly underused, as the decisions relative to the examination of reliability are rarely put in relation with the intended interpretation of the score" (p. 169). The few studies that were published mainly focused on divergent thinking scoring procedures. For instance, Silvia et al. (2008) asked undergraduates to complete three types of divergent thinking tasks: alternative uses items (e.g., "think of alternative uses of a brick"), instances items (e.g., "think of unusual things that are round"), and consequences items (e.g., "think of consequences when human no longer need sleep"). With the use of Generalizability Theory, they compared average scoring (i.e., every single idea is evaluated separately, scores are averaged across

ideas) with Top 2 scoring (i.e., only two ideas selected by the participant are scored). For both methods, only two or three raters rating a single alternative uses or instances item were required to obtain reliable scores to make relative decisions. For the consequences item, four or five raters would be needed. Hass et al. (2018) applied layperson ratings on the same dataset and compared two types of scoring. The authors concluded that with average scoring three layperson raters and four alternative uses items could produce reliable scores to make relative decisions. For the consequences item, three layperson raters and eight items would be needed. For both the alternative uses items and the consequences items, the alternative snapshot scoring system (i.e. assigning a single score to the entire set of responses generated) could not produce reliable scores, even with an increase of raters.

In the domain of creative writing, Kaufman, Lee, Baer and Lee (2007) applied the Consensual Assessment Technique (Amabile, 1996) and concluded 5 raters and 15 captions of every student would be needed to reliably assess students' caption writing ability and make absolute decisions. Long and Pang (2015) focused on science and studied the generalizability of student, teacher, and researcher creativity ratings of two open-ended tasks. They concluded (too) many raters were necessary to obtain reliable results for both relative and absolute decisions. The rater inconsistencies tended to be different across tasks.

The current paper adds to the existing generalizability studies by illustrating how Generalizability Theory may impact the development of a CPS assessment procedure. For CPS in particular, little research is conducted on the generalizability of scores, especially in a young age group. Within this study, primary school students are asked to diverge and give multiple solutions to each problem (see e.g., Reiter-Palmon & Arreola, 2015; Van Hooijdonk, Mainhard, Kroesbergen, & Van Tartwijk, 2020) and a form of average scoring is applied. This allows us to compare findings of this study with the generalizability studies on divergent thinking data. Alongside the generalizability of originality scores, completeness and practicality scores will be evaluated as well in this study. Reviewing how generalizable CPS scores for these criteria are, may give us insight in specific biases in CPS assessments and how these differ from biases in divergent thinking tasks. Therefore, we conducted a G-study and multiple D-studies with a CPS assessment dataset. To encourage creativity researchers to consider this method too, we describe the study step by step alongside information on the data structure, the syntax, and the output.

2. An analysis of creative problem solving assessment tasks

To face the unknown problems of tomorrow, schools need to foster CPS abilities in their students from an early age on (Craft, 2011). Teachers however, struggle with teaching CPS and assessing students' CPS abilities (Kettler, Lamb & Mullett, 2018). Within this CPS assessment project, we are developing tasks and rating procedures to measure CPS in primary education. Three G-studies and a series of D-studies were carried out with CPS assessment data from this project. The data, syntax and output are available for download via https://doi.org/10.34894/GIUAOI.

2.1. Choices made

The 'universe' we aimed to generalize to is upper primary school children's CPS abilities. This means we aim to measure children's everyday abilities to solve problems creatively across domains. The problem tasks we included in our assessment procedure could therefore be regarded as a selection of many possible problems that children can encounter in daily life. As such, we included the tasks in our generalizability study as a random facet. Eventually we would like teachers to rate the CPS tasks themselves. Therefore, we wanted the raters to be regarded as interchangeable and as such included the raters as a random facet too. Because the students included in the sample were a selection of many possible students, they were regarded as a random facet as well. For present purposes, we evaluated CPS assessment scores in the light of both absolute and relative decisions.

2.2. Study design

Six classes of 4th and 5th grade students ($n = 137$; mean age = 10.50; 53% girls) from three schools participated voluntarily in the study (all parents/care takers gave informed consent; see also (Van Hooijdonk, Mainhard, Kroesbergen, & Van Tartwijk, 2020 for more information on the tasks and procedure). Two problem tasks were completed by all students. One problem from the science domain and one from the social domain were selected in order to include distinctive creativity domains (Kaufman, 2012, 2009). The science problem described two children, Lisa and Tom, buying ice cream in the supermarket. The problem they encountered was that the ice cream melted on the way home. The social problem described Simon, who gets distracted in class by his friend Julian all the time. The problem Simon needed to solve was how to stop his friend's behavior without this having a negative impact on their friendship. The problems were presented written on paper and were read to the students by the researcher.

After going through two preparatory steps in which students defined their prior knowledge and described the problem at stake, students received 10 min to list as many different and original ideas as they could that would solve the problem. All the ideas were rated on originality, completeness, and practicality by two raters (one graduate and one post-graduate), using a modified version of the Consensual Assessment Technique with a 5-point Likert scale (see e.g., Amabile, 1996; Byrne, Shipman & Mumford, 2010; Corazza, 2016; Okuda et al., 1991; Reiter-Palmon et al., 2009).

2.3. Preparing the data

For every rater, the scores for originality, completeness and practicality were averaged for the separate tasks (Table 1). Next, the data was reshaped from a wide format (all student's scores in a single row, with each rating for every task in a separate column) to a

Table 1
Descriptives for CPS Scores of Two Raters on Two Tasks ($N = 137$).

	Science Problem Task				Social problem task			
	M	SD	Min	Max	M	SD	Min	Max
Originality								
Rater 1	0.74	0.54	0	2.67	0.51	0.57	0	2.83
Rater 2	0.38	0.38	0	2.00	0.26	0.40	0	2.50
Completeness								
Rater 1	2.48	0.51	1.20	4.00	2.44	0.40	1.33	3.71
Rater 2	2.36	0.53	0.75	4.00	2.18	0.43	0.67	3.40
Practicality								
Rater 1	2.75	0.65	0.80	4.00	3.36	0.60	1.28	4.00
Rater 2	2.66	0.54	0.80	3.57	3.11	0.59	0.83	4.00

long format (each row includes one rating on one task; Appendix A).

2.4. Execution of de G-study: preparing the syntax

In the first stage of our analysis, three G-studies were conducted (i.e., for originality, completeness, and practicality) to determine the variance components attributable to the different facets (students, tasks, raters and their two-way interactions; Fig. 1) for every CPS aspect. The G-studies were executed using syntax in SPSS version 24. The syntax for the G-study is provided in Appendix A.

We aimed to decompose the variance of the students, the rater, and the task. Therefore, the first line of the syntax included the VARCOMP command, followed by the dependent variable (originality, completeness, or practicality) and the BY statement followed by our facets: the students, the tasks and the raters. Any covariates authors wish to control for can be included in this first line as well by adding a WITH statement followed by the desired covariate. Because the students, tasks and raters in this study can be regarded as a selection from many possible students, tasks and raters, we included these facets in our analyses as random with the /RANDOM subcommand. By including a /FIXED line in the syntax, researchers may include facets that are not interchangeable and regarded as fixed such as a specific scoring system. The /DESIGN subcommand describes the facets (and their interactions) of which the variance components should be modelled. Because the variance components of both the main facets and their interactions could provide us with relevant information, all facets and their interactions were included in our syntax. In our design, potential three-way interactions (student x task x rater) were confounded with the residual error. Any nested effects in the design can be specified after the /DESIGN command as well using a WITHIN statement (e.g., task WITHIN domain; see also Mushquash & O'Conner, 2006 for more complex nested models). The /METHOD command includes the estimation method. We choose to apply restricted maximum likelihood (REML) because the maximum likelihood (ML) estimator might be negatively biased with smaller sample sizes, resulting in smaller variance components. For moderate deviations from normality, researchers can choose the MINQUE method as well here. Because we needed an intercept to calculate how our facets induce variance around this intercept, it was included with the /INTERCEPT = INCLUDE command.

2.5. Reviewing the output

By running the syntax, three tables were produced as output (Appendix A). The first table included factor level information. This allowed us to inspect how many scores were included per student (4), per rater (2 * 137 = 274) and per task (2 * 137 = 274). Any missing data could be reviewed in this table as well. The second table included the main outcomes from the G-study, i.e., the variance estimates for every facet and the interactions. The last table included the covariance matrix that was used to calculate these estimates. We ran the syntax separately for the originality, completeness and practicality scores, collected the variance components from the second table and transformed them into percentages of the total variance (Table 2).

Table 2
The Results of the G-study. Variation (s^2) Attributed to the Facets (p, t, r) and their Interactions (st, sr, rt, str,e).

	Originality		Completeness		Practicality	
	s^2	%	s^2	%	s^2	%
Student (p)	0.064	22.22	0.053	21.81	0.141	27.54
Task (t)	0.013	4.51	0.004	1.64	0.137	26.76
Rater (r)	0.044	15.28	0.015	6.17	0.011	2.14
Student*Task (st)	0.100	34.72	0.052	21.40	0.147	28.71
Student*Rater (sr)	0.002	0.69	0.000	0.00	0.000	0.00
Rater*Task (rt)	0.002	0.69	0.004	1.65	0.005	0.98
Student*Task*Rater / error (str,e)	0.063	21.88	0.115	47.33	0.071	13.87
Total s^2	0.288	100.00	0.243	100.00	0.512	100

Table 3
Relative and Absolute Error Variances and Generalizability for Two Tasks and Two Raters.

	Originality	Completeness	Practicality
Absolute error variance	0.29	0.24	0.51
Relative error variance	0.17	0.17	0.22
Absolute generalizability (D)	.40	.45	.46
Relative generalizability (g)	.49	.49	.61

2.6. Interpretation of the G-study

By reviewing the variance of the student facet, the amount of variance that can be rightly attributed to students' true ability was examined first. For all three aspects of CPS, somewhat similar amounts of true student score variance were found. For originality and completeness, about 22% of the variance could be attributed to actual differences between students. For practicality about 28% could be attributed to students. This means that despite the effort to focus on students' true CPS ability, a large share of 72 to 78% of the variance was attributed to other sources. Next, the variance components of the task and rater facet were reviewed to get an idea of how these facets impacted the CPS scores. For originality and completeness, the task itself did not seem to introduce a lot of variability (only 4.51% and 1.64%). The much higher percentage of variance due to the task for practicality (about 27%) implies that the influence of our tasks on the CPS scores for practicality was almost as strong as the influence of the students who completed these tasks. This is also apparent in the average differences in practicality scores in Table 1. Raters did not differ that much in their scoring of completeness and practicality, illustrated by the 6% and 2% variance at the rater level. For originality, the variance at the rater level was about 15%. Thus, raters seemed to disagree more in their originality ratings than in their completeness and practicality ratings.

When it comes to the interactions between facets, the student x task interaction of all three CPS aspects accounted for a large part of the variance (about 34% for originality, 21% for completeness, and 29% for practicality). This implies that for the students' CPS scores, it mattered which student completed which tasks. This might for example be due to task-specific or domain-specific CPS abilities. Almost no variance (max 0.69%) was found on the student x rater level, which implies that raters were not biased towards certain students. Because the students were unknown to the raters, this could be expected. There was also very little (0.69% to 1.65%) variance on the rater x task level, which indicated raters did not differ in their judgements depending on the task they rated. This indicates that our rating scheme may be used across various problem tasks. Additionally, there were considerable amounts of variance left at the student x task x rater and error level. Especially for completeness, this percentage of residual variance was quite large (about 47%).

Next, the generalizability of the applied assessment procedure was reviewed. The formula (Formula 1 & 2) was used to calculate the absolute (D) and relative (G) generalizability that reflect the generalizability for the set of tasks and raters we used (two tasks, two raters; Table 3).

Both the absolute ($D = 0.40$ - 0.46; Table 3) and relative ($g = 0.49$ - 0.61; Table 3) coefficients did not reach the desired level of 0.70 (Brennan, 2010). Therefore, the current assessment procedure including two tasks and two raters could be regarded as insufficient to make both absolute and relative decisions. Therefore, multiple D-studies were conducted to get an idea of how many tasks and raters should be included in the assessment procedure to reach the desired generalizability level of 0.70 for originality, completeness, and practicality.

2.7. Execution of the D-study

Multiple D-studies were executed for the three CPS aspects (i.e., originality, completeness, and practicality) and for both absolute (norm-based) and relative decisions. The numbers of tasks and raters in our assessment design were varied and generalizability was recalculated using the percentages of the variance and the formulae (Table 2; Formula 1 & 2). For instance for originality, five tasks and two raters would produce an absolute generalizability of about 0.55 (Absolute Generalizability (D) = 22.22 / (22.22 + 4.51/5 + 15.28/2 + 34.72/5 + 0.69/2 + 0.69/(10) + 21.88/(10)) and a relative generalizability of about 0.70 (Relative generalizability (g) = 22.22/(22.22 + 34.72/5 + 0.69/2 + 21.88/10). The number of tasks and/or raters was varied until an absolute and relative generalizability of at least 0.70 was reached.

2.8. Interpretation of the D-study

According to these predictions, for absolute decisions about a student's ability to produce original ideas, at least ten CPS tasks rated by four raters are necessary to reach a generalizability of 0.70. When aiming for only two raters, at least thirty tasks should be used. For completeness, four tasks and four raters should be used to make absolute decisions, and for practicality eight tasks and two raters. (Fig. 2). Because for relative decisions the absolute ratings on tasks do not matter, a lower number of tasks and raters needs to be included in the assessment procedure. For instance, only five tasks rated by two raters are necessary to provide generalizable originality and completeness ratings. For practicality, this would be even less: only three tasks rated by two raters. In sum, the number of tasks needs to be increased to make absolute and relative discussions for all aspects. Increasing the number of raters would, on the other hand, positively influence the absolute generalizability of the originality and completeness ratings. With these results in mind, more CPS tasks can be developed, and rating procedures can be refined, to enhance the CPS assessment procedure.

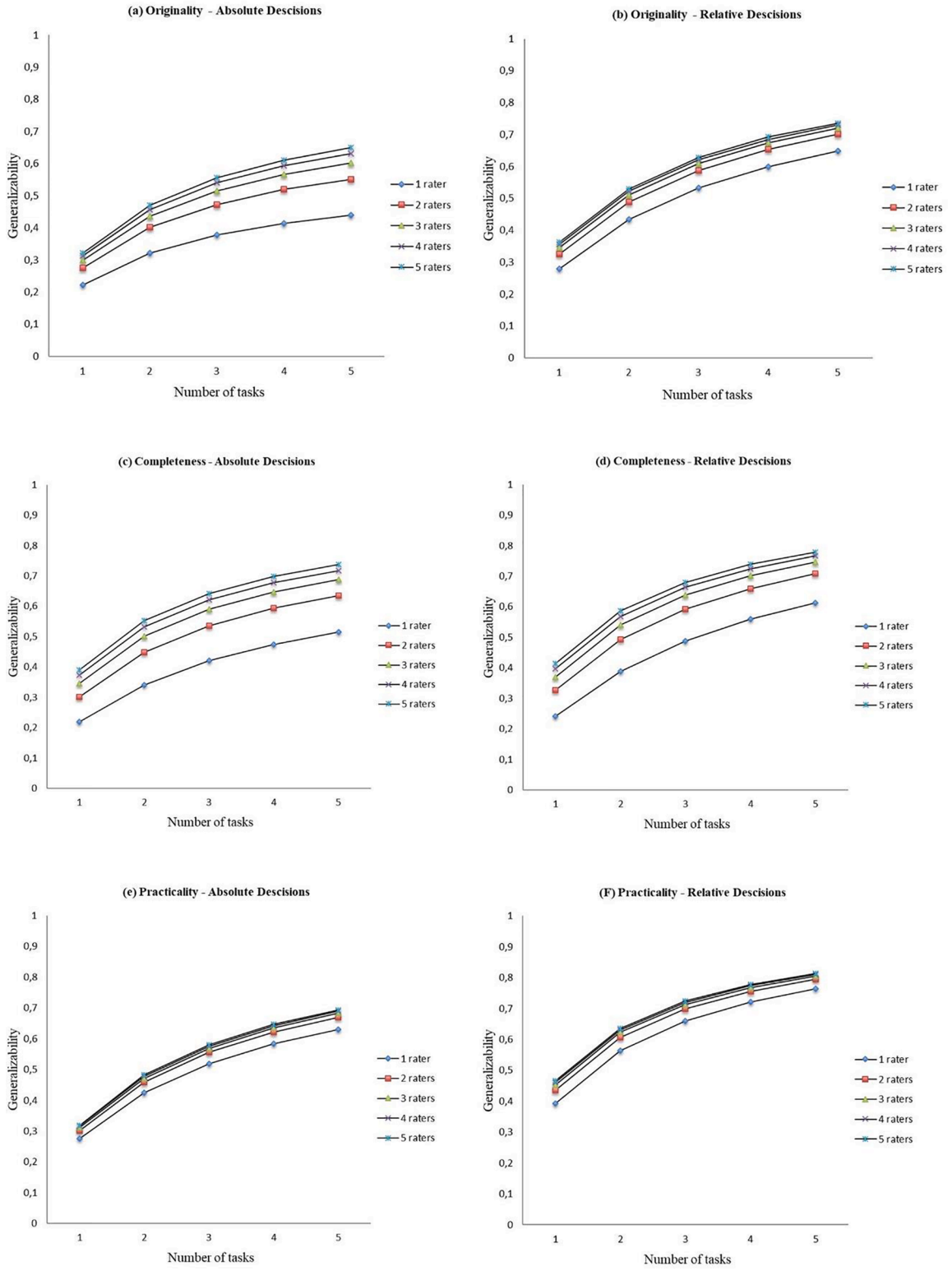


Fig. 2. Graphical representation of the results of the D-studies. The X-axis represents the number of tasks, the Y-axis the generalizability of the scores. The separate lines represent the number of raters.

3. Discussion

Within this report, we aimed to show how creativity researchers can use Generalizability Theory to make informed decisions about creativity assessment procedures and CPS assessments in particular. The G-study on the CPS data not only provided us with valuable information on where to start with improving the CPS assessment procedure but also gave us insight in the assessment of CPS.

3.1. Implications for the assessment of CPS

The assessment procedure under investigation in this study consisted of two CPS tasks and two raters. If only this set of tasks and raters would be used to assess CPS abilities, students' scores would reflect other factors than solely student CPS performance and, in other words, would be biased. The earlier generalizability studies on divergent thinking tasks showed that when scores are averaged across the generated ideas a small increase of the number of tasks (e.g., 4–8; Hass et al., 2018) and raters (e.g., 2 or 3; Silvia et al., 2008) may suffice to obtain reliable scores to make relative decisions. Within our CPS assessment procedure, students also generated multiple solutions to a problem and a similar way of scoring was applied. The series of D-studies indicated that also for the assessment of CPS, three to five tasks rated by two raters may be applied to obtain reliable CPS scores to make relative decisions. This suggests that when an idea generation stage is applied (see also Reiter-Plamon & Arreola, 2015), average scoring works for assessments of CPS as well. For educational practice, this means that a reasonable number of tasks (e.g., five tasks throughout an academic year) and raters (e.g., two teachers) may suffice to get an image of students' CPS abilities. However, more research is necessary to conclude that teacher ratings of students' CPS abilities are aligned with trained rater assessments of these abilities.

Previous studies on divergent thinking showed that it tends to be difficult to get raters in line on what ideas are original and what ideas are not (see e.g., Benedek et al., 2016; Grohman, Wodniecka & Kłusak, 2006; Guo et al., 2019; Silvia et al., 2008). The seemingly large impact of the rater on the originality scores compared to the impact on completeness and practicality scores suggests that for CPS this is the case as well. The applied rating procedure for originality could be reviewed again to minimalize these disagreements in future CPS assessments.

This study also focused on ratings of completeness and practicality, uniquely for CPS. For completeness, a relatively large part of the variance could not be attributed to the students, tasks, and raters. This might be explained by a three-way student x rater x task interaction. It may be that for specific students, the raters scored completeness differently within tasks. This may also mean that a different factor that should be controlled for influenced the scores for completeness, such as perceived problem complexity. More research is necessary to explore both hypotheses. The large share of the variance attributed to the tasks for the practicality scores indicated that the ability to come up with practical ideas when doing CPS highly depends on the type of problem situation or task presented to the students. Differences across tasks were found in many creativity studies (e.g., Hass et al., 2018; Long & Pang, 2015; Reiter-Palmon et al., 2009; Silvia et al., 2008). To assess practicality with the CPS assessment procedure under study, we might need to include more tasks. For the different CPS aspects, the impact of increasing the number of tasks and raters may however vary. For originality and completeness for instance, it seems more worthwhile to increase the number of raters, especially if we intend to make absolute decisions. This does not have to be the case for every creativity assessment procedure using the consensual assessment technique. In other assessment procedures, the number of tasks or raters may even be reduced (e.g., Kaufman et al., 2007). Executing D-studies may help creativity test developers in gaining sufficient data but also helps to prevent administering unnecessary tasks.

In this Generalizability study, the 'universe' we aimed to generalize to was upper primary school children's CPS abilities. In other words, we studied CPS as an ability that we aimed to generalize across tasks or domains. Although more research is necessary to validate the relatively new tasks (see also Van Hooijdonk, Mainhard, Kroesbergen, & Van Tartwijk, 2020), the high variance on the student x task interaction may already indicate that CPS – like many other forms of creativity- is a task-specific or domain-specific ability (Baer, 2012; Barbot et al., 2016; Reiter-Palmon et al., 2009).

3.2. Implications for creativity research

Generalizability Theory could be of assistance to study specific issues in the creativity literature, such as the domain-specificity of creative abilities. By including multiple tasks within and across domains, researchers could study the variance within and between domains. As Reiter-Palmon et al. (2009) suggested, tasks characteristics such as task complexity, involvement, and forms of self-efficacy (e.g., whether a student is convinced he or she can solve the problem) may influence creative outcomes even within domains. Such task characteristics may cause variance in student scores as well and may as such be included in the G-study.

As noted earlier, more elements of a creativity assessment procedure can be added as a facet in a G-study. For instance, the type of rater (e.g., trained expert raters versus teachers or layman raters; Hass et al., 2018, Kaufman, Baer, Cropley, Reiter-Palmon & Sinnett, 2013) or the type of rating procedure (e.g., snapshot scoring, Long & Pang, 2015; Silvia et al., 2008) could be included in the analysis to see how they affect the generalizability of the results. Creativity researchers working with Generalizability Theory should however carefully select the facets they include in their design, because they may miss elements that (strongly) impact test scores. Researchers should be aware that they could overlook something that is called a "hidden facet" (Shavelson & Webb, 1991). A hidden facet can for example be revealed by two facets that cause error in similar ways: when the levels of one facet change, so do the levels from another. This may for instance happen when tasks are administrated in the same order every time. Here, the variance found between the different task scores might be interpreted as task variance, while in fact it could also be an order or learning effect. It is therefore important to consider possible hidden facets in the assessment design.

Further, Generalizability Theory relies on increasing (or decreasing) the number of observations on facets with high (or low)

variance (e.g., adding more tasks and raters) to reach the level of desired generalizability. This could however result in unpractical designs. This is illustrated with our analysis of CPS tasks: with only two raters at least 30 tasks are estimated to be necessary to make absolute decisions about a students' ability to produce original ideas. As mentioned before, creativity test developers should therefore ideally define their aims beforehand (Myszkowski & Storme, 2019). Do they aim to make absolute decisions and develop a standardized creativity test, or do they want to take a more relative approach? Often, creativity test developers may be merely interested in a student's relative performance. Also, as some argue, it is almost impossible to develop a standardized creativity test for education (Harris, 2016). As Silvia et al. (2008) state, it would be valuable to do a Generalizability study in the piloting stage of the development of an assessment. This will give researchers insight in the sources of error and could assist in determining how many facets such as tasks and rater would be necessary to generalize the scores to a specific ability.

Within this report, we discussed in detail how Generalizability Theory may be applied in creativity studies. The information on the analysis including the data structure, the syntax, and the output (Appendix A) may help creativity researchers to apply this theory to their data. We choose this method because it is relatively easy to apply and interpretation is not that difficult for researchers familiar with ML methods like ANOVA or linear regression analysis. Besides this, Generalizability Theory is very useful in providing information on overall decisions about the design of an assessment procedure (Linacre, 1996). The calculation of both absolute and relative generalizability and subsequent D-studies assist in making these decisions. Researchers interested in more specific information at the individual (e.g., student) level might be interested in Many Facet Rasch Modeling (MFRM; Linacre, 1996; Primi, Silvia, Jauk & Benedek, 2019). Although studies applying both techniques in general report a comparable impact of facets (e.g., Smith & Kulikowich, 2004; Sudweeks, Reeve & Bradshaw, 2004), MFRM can provide specific suggestions regarding the characteristics of the added facets, e.g., like adding more lenient raters or more difficult tasks (Linacre, 1996).

3.3. Limitations of the theory

As with every theory, Generalizability Theory comes with some limitations (Brennan, 2010). In balanced designs (when every student completes the same number of tasks and receives the same number of ratings), the calculation of the variance components is straightforward and the estimations are mostly unbiased. When tasks are rated by a different number of raters or when every student completed a different number of tasks, an unbalanced design occurs. Here, the estimation of variance components is more complex and a G-study might produce biased variance components. Consequently, biased D-study results might follow. Brennan (2001) describes G-study and D-study solutions for some frequently encountered unbalanced designs. In the case of an unbalanced design, we recommend researchers to consult this work.

Researchers should also be aware that variances are calculated based on the facets that are included in the design. Adding more facets alters the estimation of all other facets as well. Besides this, the generalizability of an assessment procedure should be reanalyzed and recalculated when the assessment design is altered to make sure adjustments were indeed an improvement. Also, each creativity assessment procedure needs its own G-study and D-study, as it is likely to have its own set of unique tasks and rating procedures. Findings from one generalizability study cannot simply be applied to another assessment procedure. Nevertheless, using more G-studies and D-studies in creativity research may not only enhance the critical evaluation of creativity tests, it may also help to move towards including the assessment of creativity and CPS in particular more efficiently in everyday educational practice.

CRediT authorship contribution statement

Mare Van Hooijdonk: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Tim Mainhard:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision. **Evelyn H. Kroesbergen:** Validation, Writing – original draft, Writing – review & editing, Supervision. **Jan Van Tartwijk:** Validation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tsc.2021.100994.

References

Amabile, T. M. (1996). *Creativity in context.* Boulder, CO: Westview Press.

Amabile, T. M., & Pillemer, J. (2012). Perspectives on the social psychology of creativity. *The Journal of Creative Behavior, 46*(1), 3–15. https://doi.org/10.1002/jocb.001

Arreola, N. J., & Reiter-Palmon, R. (2016). The effect of problem construction creativity on solution creativity across multiple everyday problems. *Psychology of Aesthetics, Creativity, and the Arts,, 10*(3), 287–295. https://doi.org/10.1037/a0040389

Baer, J. (2012). Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior, 46*(1), 16–29. https://doi.org/10.1002/jocb.002

Barbot, B., Besançon, M., & Lubart, T. (2016). The generality-specificity of creativity: Exploring the structure of creative potential with EPoC. *Learning and Individual Differences, 52*, 178–187. https://doi.org/10.1016/j.lindif.2016.06.005

Barbot, B., & Lubart, T. (2012). Creative thinking in music: Its nature and assessment through musical exploratory behaviors. *Psychology of Aesthetics, Creativity, and the Arts, 6*(3), 231–242. https://doi.org/10.1037/a0027307

Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., et al. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity, 21*, 75–84. https://doi.org/10.1016/j.tsc.2016.05.007

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2010). Generalizability Theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Byrne, C. L., Shipman, A. S., & Mumford, M. D. (2010). The effects of forecasting on creative problem-solving: An experimental study. *Creativity Research Journal, 22*(2), 119–138. https://doi.org/10.1080/10400419.2010.481482

Ceh, S. M., Edelmann, C., Hofer, G., & Benedek, M. (2021). Assessing Raters: What Factors Predict Discernment in Novice Creativity Raters? *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.515

Conti, R., Coon, H., & Amabile, T. M. (1996). Evidence to support the componential model of creativity: Secondary analyses of three studies. *Creativity Research Journal, 9*(4), 385–389. https://doi.org/10.1207/s15326934crj0904_9

Corazza, G. E. (2016). Potential originality and effectiveness: The dynamic definition of creativity. *Creativity Research Journal, 28*(3), 258–267. https://doi.org/10.1080/10400419.2016.1195627

Craft, A. (2001). Little c creativity. In A. Craft, B. Jeffrey, & M. Leibling (Eds.), *Creativity in education* (pp. 45–61). London, UK: Continuum.

Craft, A. (2011). *Creativity and education futures*. Trentham Books.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review, 23*(2), 72–79. https://doi.org/10.1080/02783190009554069

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education, 45*(6), 560–569. https://doi.org/10.1111/j.1365-2923.2010.03913.x

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Falmer.

Grohman, M., Wodniecka, Z., & Kłusak, M. (2006). Divergent thinking and evaluation skills: Do they always go together? *The Journal of Creative Behavior, 40*(2), 125–145. https://doi.org/10.1002/j.2162-6057.2006.tb01269.x

Guo, J., Ge, Y., & Pang, W. (2019). The underlying cognitive mechanisms of the rater effect in creativity assessment: The mediating role of perceived semantic distance. *Thinking Skills and Creativity, 33*, Article 100572. https://doi.org/10.1016/j.tsc.2019.100572

Harris, A. (2016). *Creativity and education*. Palgrave Macmillan.

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology, 9*, 1343. https://doi.org/10.3389/fpsyg.2018.01343

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation procedures and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Isaksen, S. G., & Treffinger, D. J. (2004). Celebrating 50 years of reflective practice: Versions of creative problem solving. *The Journal of Creative Behavior, 38*, 75–101. https://doi.org/10.1002/j.2162-6057.2004.tb01234.x

Jiang, Z. (2018). Using linear mixed-effect model framework to estimate generalizability variance component in R: A lme4 package application. *Methodology, 14*(3), 133–142. https://doi.org/10.1027/1614-2241/a000149

Jiang, Z., Raymond, M., Shi, D., & DiStefano, C. (2020). Using a linear mixed-effect model f ramework to estimate multivariate Generalizability Theory parameters in R. *Behavior Research Methods, 52*(6), 2383–2393. https://doi.org/10.3758/s13428-020-01399-z

Kaufman, J. C. (2012). Counting the muses: development of the Kaufman Domains of creativity scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts, 6*(4), 298–308. https://doi.org/10.1037/a0029751

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*(2), 171–178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts, 7*(4), 332–340. https://doi.org/10.1037/a0034809

Kaufman, J. C., Cole, J. C., & Baer, J. (2009). The construct of creativity: Structural model for self-reported creativity ratings. *The Journal of Creative Behavior, 43*(2), 119–134. https://doi.org/10.1002/j.2162-6057.2009.tb01310.x

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*(2), 96–106. https://doi.org/10.1016/j.tsc.2007.04.002

Kettler, T., Lamb, K. N., & Mullett, D. R. (2018). *Developing creativity in the classroom: Learning and innovation for 21st century schools*. Prufrock Press.

Linacre, J. M. (1996). Generalizability Theory and many-facet Rasch measurement. In G. Engelhard, Jr., & M. Wilson (Eds.), *Objective measurement: Theory into practice (Vol.3, pp. 85-98)*. Norwood, NJ: Ablex.

Long, H. (2014). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. *Thinking Skills and Creativity, 13*, 183–194. https://doi.org/10.1016/j.tsc.2014.05.002

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity, 15*, 13–25. https://doi.org/10.1016/j.tsc.2014.10.004

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., Van der Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087–1102. https://doi.org/10.1007/s10459-013-9450-z

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542–547. https://doi.org/10.3758/bf03192810

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

Okuda, S. M., Runco, M. A., & Berger, D. E. (1991). Creativity and the finding and solving of real- world problems. *Journal of Psychoeducational Assessment, 9*(1), 45–53. https://doi.org/10.1177/073428299100900104

Piffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity, 7*, 258–264. https://doi.org/10.1016/j.tsc.2012.04.009

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), 176–186. https://doi.org/10.1037/aca0000230

Reiter-Palmon, R., & Arreola, N. J. (2015). Does generating multiple ideas lead to increased creativity? A comparison of generating one idea vs. many. *Creativity Research Journal, 27*(4), 369–374. https://doi.org/10.1080/10400419.2015.1087274

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144–152. https://doi.org/10.1037/aca0000227

Reiter-Palmon, R., Illies, M. Y., Cross, L. K., Buboltz, C., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts, 3*(2), 73–80. https://doi.org/10.1037/a0013410

Schoevers, E. M., Kroesbergen, E. H., & Kattou, M. (2020). Mathematical creativity: A combination of domain-general creative and domain-specific mathematical skills. *The Journal of Creative Behavior, 54*(2), 242–252. https://doi.org/10.1002/jocb.361

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., & Hess, K. I. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Smith, E. V., J.r., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*(4), 617–639. https://doi.org/10.1177/0013164404263876

Sternberg, R. J. (2020). What's wrong with creativity testing? *The Journal of Creative Behavior, 54*(1), 20–36. https://doi.org/10.1002/jocb.237

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239–261. https://doi.org/10.1016/j.asw.2004.11.001

Van Hooijdonk, M, Mainhard, M. T., Kroesbergen, E. H., & Van Tartwijk, J. (2020). Creative problem Solving in primary Education: Exploring the role of fact finding, problem finding, and solution finding across tasks. *Thinking Skills and Creativity, 37*, Article 100665. https://doi.org/10.1016/j.tsc.2020.100665

Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2020). Creativity: Intrapersonal and interpersonal selection of creative ideas. *The Journal of Creative Behavior, 54*(3), 626–635. https://doi.org/10.1002/jocb.397