TPM®

USING THE GAGE R&R METHOD TO EVALUATE THE RELIABILITY AND ASSESSMENT PROCESS OF THE CREATIVE ENGINEERING DESIGN ASSESSMENT

SOPHIE MORIN
LOUIS-MARC BOURDEAU
JEAN-MARC ROBERT
POLYTECHNIQUE MONTRÉAL

Creativity, being a key competency of engineering, must be taught and assessed. Here, we verify the reliability of the Creative Engineering Design Assessment (CEDA), a psychometric test for engineering students, as well as clarify the assessment process. To test the former, the gage repeatability and reproducibility (R&R) method was applied innovatively. Our findings suggest that the use of the gage R&R method is relevant in a psychometric environment and that the measured total variation is satisfactory. In addition, control charts were used to further analyze the assessment strategy's reliability. We demonstrated that the assessment process for the two qualitative criteria (Originality and Usefulness) of the CEDA was in control, that is, the values' variation was caused by unpredictable but normal and inevitable events. The results demonstrate the test's reliability according to two concepts (repeatability and reproducibility) and allow for the refinement of the assessment process by defining the Likert scales with more precision.

Key words: Creativity; Assessment methodology; Engineering education; Gage R&R method; Reliability.

Correspondence concerning this article should be addressed to Sophie Morin, Polytechnique Montréal, 2900 Edouard-Montpetit Blvd, H3C 3A7 Montréal, QC, Canada. Email: sophie.morin@polymtl.ca

Creative and innovative people are recognized for their contribution to society's well-being. This is particularly the case for engineers who are often called upon to produce innovative ideas and thus participate in the improvement of their organization's products, services, and processes, to keep them competitive. It thus stands to reason that engineering schools must foster creativity among their students as an integral part of their curriculum.

In the context of research and training, the assessment of creativity is a major challenge (Clary, Brzuszek, & Fulford, 2011; Cropley, Kaufman, & Cropley, 2011; Plucker & Runco, 1998; Treffinger, Young, Selby, & Shepardson, 2002). In fact, most of the numerous tools that have been developed in the last 60 years (Clary et al., 2011; Kim, 2014; Treffinger et al., 2002), including the widely used Alternate Uses Test (Guilford, 1968), assess divergent thinking (DT)[1] rather than creativity.

Even though, according to some authors, DT is often confused with creativity, "this can be misleading because convergent thinking[2] is as important for creativity as divergent thinking." (Piffer, 2012, p. 260). As Piffer explains: "The name of the most popular creativity test, the Torrance Test of Creative Thinking, is exemplar. Its name suggests that other cognitive tests (e.g.,

TPM Vol. 24, No. 4, December 2017 – 499-514 – doi:10.4473/TPM24.4.2 – © 2017 Cises

Working memory tests, general knowledge, IQ tests) are not tests of creative thinking." (p. 260). In the same vein, authors like Gabora and Kaufman (2010) have argued that creative production relies just as heavily on knowledge and analytical thinking (which are associated with convergent thinking) as on imagination and divergent thinking.

To address the need for a more holistic evaluation of creativity in engineering, Charyton (2014) recently developed a tool specific to this context: the Creative Engineering Design Assessment (CEDA). The present study seeks to improve the CEDA's evaluation process and determine its reliability using the gage reproducibility and repeatability method (R&R), a statistical method widely used in engineering. Gage R&R is generally used on quantitative data such as length or voltage, but was applied here to the analysis of qualitative data yielded by the CEDA.

The paper is structured as follows. Firstly, we present the CEDA test; secondly, we describe the gage R&R method; thirdly, we explain the methodology and present the results. Finally, we present the discussion and the conclusion with suggestions for future research.

CREATIVE ENGINEERING DESIGN ASSESSMENT (CEDA) TEST

The CEDA is a psychometric test that relies on qualitative criteria requiring assessments made by observers (or judges). Recently developed in the USA by Charyton (2014) to assess engineering students' creative performance, it is based on the Purdue Creativity Test (PCT; Harris, 1960), a well-known, validated test. The CEDA requires participants to conceive concepts using the various geometric shapes presented. As with the PCT, results are assessed according to three criteria (Flexibility, Fluidity, and Originality). The CEDA, however, also includes a "Usefulness" criterion and adopts a more elaborate quantitative scale than the test it derived from. A review of available literature shows the CEDA assesses five important aspects of creativity (Figure 1) as well as confirming both content and face validity of the test (Charyton, 2014; Charyton, Jagacinski, & Merrill, 2008; Charyton, Jagacinski, Merrill, Clifton, & Dedios, 2011; Charyton & Merrill, 2009).

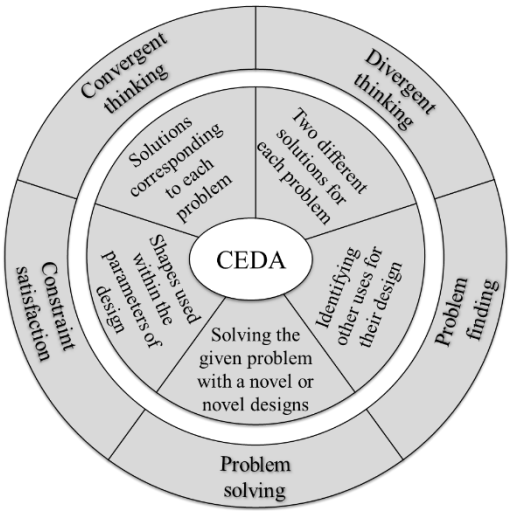

FIGURE 1
Five creativity aspects covered by the CEDA test (figure inspired by Charyton, 2014).

To evaluate creative performance, divided into the five aspects presented in Figure 1, the CEDA uses four criteria: Fluidity, Flexibility, Originality, and Usefulness. Fluidity is the number of different items produced by a person during a creative session; Flexibility is the number of categories covered by these items; Originality is the frequency with which the items are found (or repeated) across the sample; and Usefulness corresponds to how well an item responds to the general goal suggested. The first three (Fluidity, Flexibility, and Originality) are well known and often used to assess creativity (Clary et al., 2011; Kim, 2006; Treffinger et al., 2002). Charyton et al. (2011) added the Usefulness criterion due to its importance, indeed essentiality, in an engineering context. The test provides a numerical score, from zero to 284, with a higher score indicating higher creativity.

Scoring and interpretation of results obtained with the CEDA present certain drawbacks. Indeed, very limited quantitative data have been published, so a range of results is difficult to establish. Moreover, the Fluidity and Flexibility criteria are measured quantitatively, whereas Originality and Usefulness are measured qualitatively using Likert scales. Even with a somewhat detailed description of each level of these scales, subjectivity can still be present. To this day, to our knowledge, no results have been published to show, describe, and analyze links between assessors' work and participants' scores.

A final limitation of the CEDA lies in the actual instructions for assessment of the data collected. From our point of view, this test has several merits but lacks details in the assessing process. "As in the previous study, two judges scored each CEDA: one judge from engineering and one judge from psychology. There was a total of four engineering judges, who scored subsets of the CEDA's, and one psychology judge, who was a CEDA test developer. Two of the CEDA test developers trained the judges. Judges practiced scoring in a team environment; however, each judge evaluated the CEDA's separately" (Charyton et al., 2011, p. 785). From a practical viewpoint, guidelines are vague and difficult to apply. Very limited work has been published by other researchers to further understand the test's operationalization (Carpenter, 2016).

THE GAGE R&R METHOD

The gage R&R is a statistical method used in engineering to measure (or gage) the reliability of a measurement system (Ostle, Turner, Hicks, & McElrath, 1996; Wheeler, 2006). Typically, in engineering these would include the operator (human or not), the work piece, and the tool. A gage R&R study helps determine whether the measurement system's variability is small compared with the process' variability; how much variability is caused by the operators; and whether the measurement system is capable of discriminating between different parts.

Although the number of operators, parts, and trials varies for a particular application of the gage R&R method, the general procedure of measurement with this method consists of forming a part sample; randomly choosing three operators (if the operator is human, they should be trained and familiar with the process but be neither novice nor expert); proceeding with the first assessment (T1); repeating the assessment for the second trial (T2) (randomizing the order of the measurements).

If the total variance associated with the measurement system (repeatability and reproducibility) is less than 10%, the measurement system is judged acceptable. If the total variance is between 10% and 30%, the measurement system is judged acceptable depending on the

application, the cost of the measuring device, the cost of repair, or other factors. Finally, if it is more than 30%, the measurement system is judged unacceptable and should be improved (Automotive Industry Action Group, AIAG, 2010; Ostle et al., 1996; Wheeler, 2006).

Principal Elements

The gage R&R method is used for multiple purposes: to compare the measurement system variability to the process variability, calculate how much variability in the measurement system is caused by differences between operators, and determine if the measurement system is capable of discriminating between different parts.

Accuracy and Precision

The gage R&R method measures accuracy, defined as "hitting the right spot" (i.e., measuring the right thing), and precision, defined as "hitting the same spot every time" (i.e., obtaining the same result every time the measurement is done). In other words, the method can be used to determine if a measuring system consistently measures the right concept. As pertains to psychometric testing, the gage R&R method can thus be used to analyze variability in repeatability and reproducibility and to identify to what extent variance in the results is due to the measurement system. Figure 2 illustrates how repeatability and reproducibility are combined in this statistical method.

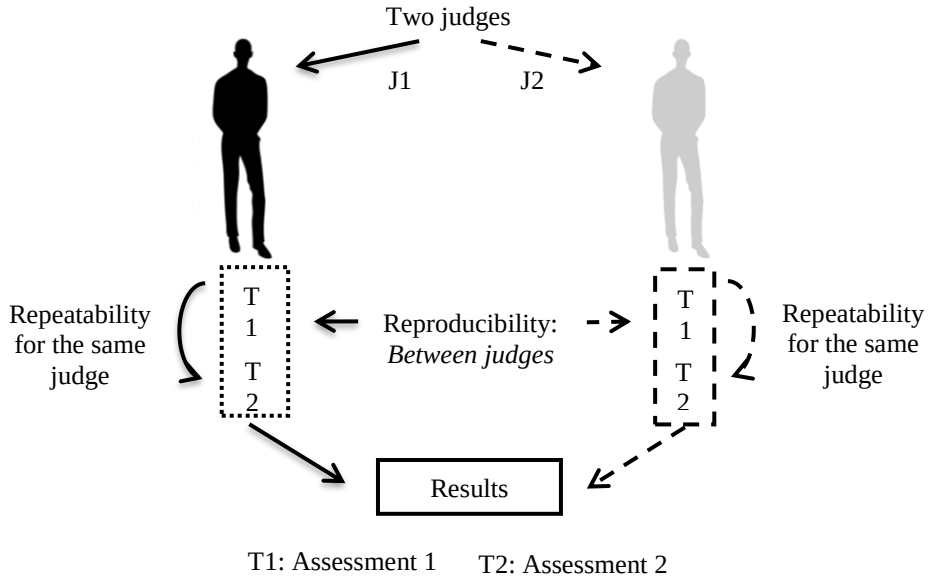

FIGURE 2
Repeatability and reproducibility constructs.

The accuracy aspects of the test, bias, linearity, stability (AIAG, 2010), are not addressed in this study. As presented above, we rely on previous studies by Charyton (2014) to establish that the CEDA has a good validity and indeed assesses creativity.

A measurement system's precision is evaluated by two constructs: repeatability and reproducibility. "The repeatability of a measuring device is the variability observed when repeated measurements are obtained by the same operator on the same unit or part" (Ostle et al., 1996, p. 337). "The reproducibility of the measurement process is estimated by considering the variability among the sample averages for the operators used in the study" (p. 340). The gage R&R method analyzes the two constructs' variability and identifies to what extent variance in the results is due to the measurement system.

Control Charts

One of the most useful features of the methodology of evaluation is the control charts (Figure 3). They allow the professionals in charge of the test (e.g., engineer, researcher, manager) to visually determine if the measured values are between the upper and lower control limits, that is, if a process is "in control." According to Wheeler and Chambers (1992), a process will inevitably include variation. However, two types of variation exist: controlled and uncontrolled. Controlled variation is due to "random" causes and uncontrolled variation is due to "assignable" causes. A process "not in control" is being affected by assignable causes that can be identified and eliminated. Control charts are the tool used to overcome these assignable causes and move beyond the barrier of process improvement. The data of several operators (judges) can be displayed on the same chart to get a global view of the assessments.

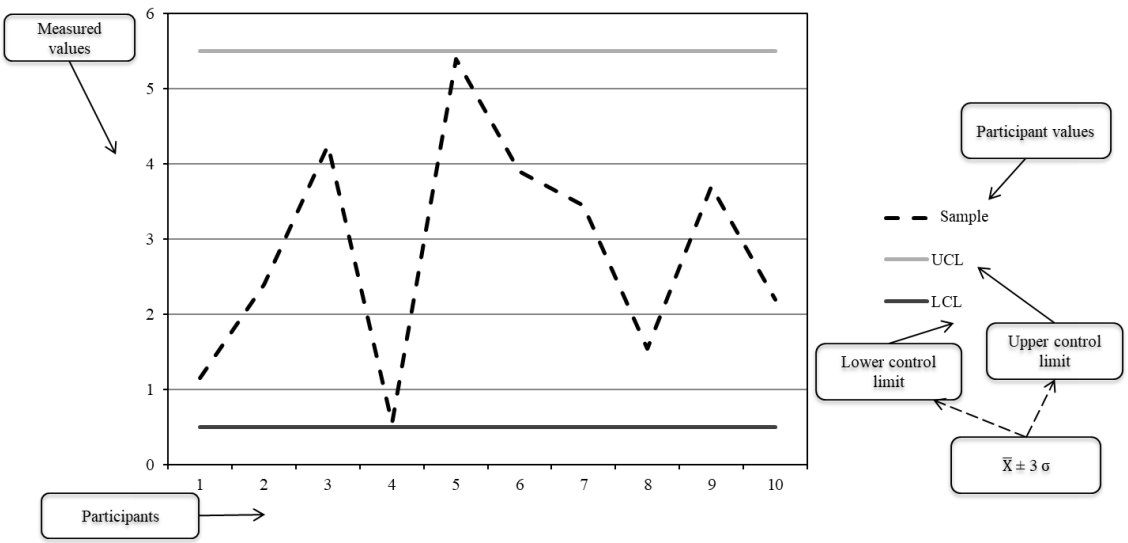

FIGURE 3
Principal elements of the control charts.

METHOD

This study had two objectives: to clarify the CEDA's assessment process and to test the reliability of the CEDA measurement system using the gage R&R method. The method presents

the procedure as a methodological design in two phases, the samples, the judges, and the test used for testing the reliability and the assessment process of the CEDA.

Procedure

Phase 1

In order to optimize participation, we translated the test into French. Hence, a pretest step was conducted to verify language and understanding of the guidelines.

Phase 1 aimed to establish a more detailed assessment scale for the Originality and Usefulness criteria than what was suggested by Charyton as well as to clarify the CEDA's assessment process. Participants were chosen from guests attending a workshop on creativity. They were invited through personal networking and included people interested in learning more about creativity and how it may be developed (e.g., project manager, sound technician, pedagogical consultant, marketing director, industrial engineer, receptionist, professor). The test was presented as an introduction to this conference and workshop. No compensation was given to the participants. Workshops consisted of three sessions of three hours each.

Twenty-two tests were evaluated by three judges, all female engineers aged 25-40 years. All three were educators with an interest in creativity as a competency and a background in art (circus, dance). They assessed the CEDA tests in two phases, two months apart to minimize the memory bias.

Confronted with assessment difficulties and questions, the judges met to clarify certain aspects of their evaluations. To diminish the confusion regarding the levels for Originality and Usefulness, copies of the designs created by the participants were made (sketch only) and classified according to the scale suggested. One of the judges suggested it would be helpful to define with more precision each level in regard to the test itself. To accomplish the task, they looked at each design and used the "think aloud" approach to describe their understanding. This helped them build common knowledge and become more confident with the assessment task.

Judges defined in more detail what each level meant with regard to the CEDA. All assessments were done individually, but a group discussion was arranged between the two assessments of each phase to resolve any remaining confusion or disagreements.

For the first analysis, the three judges used the CEDA's original scoring system. They followed the general instructions given by Charyton (2014). They used objective/quantitative measures for Fluidity and Flexibility, and subjective/qualitative measures (on two Likert scales) for Originality and Usefulness. They made one judgment for each design and a third one (global) for each problem (Figure 4-B; D1, D2, Global). They proceeded with the summation of the four criteria, with the equation suggested by Charyton: Fluidity + Flexibility + 2*Originality + 2*Usefulness (overall creativity score). We found confusing the use of the term "overall" to describe the "global" judgment for design 1 and 2 (D1 and D2) and not the overall creativity score obtained from the previous formula so we changed overall to global.

Moreover, this phase allowed us to conduct a first evaluation of the judges' performance and concordance with the gage R&R method. Two months after a first assessment, a second one was conducted (reproducibility aspect). The tests were randomly distributed to eliminate a possible interaction bias across tests. Control charts were built to visually compare reproducibility and repeatability for each judge.

A B

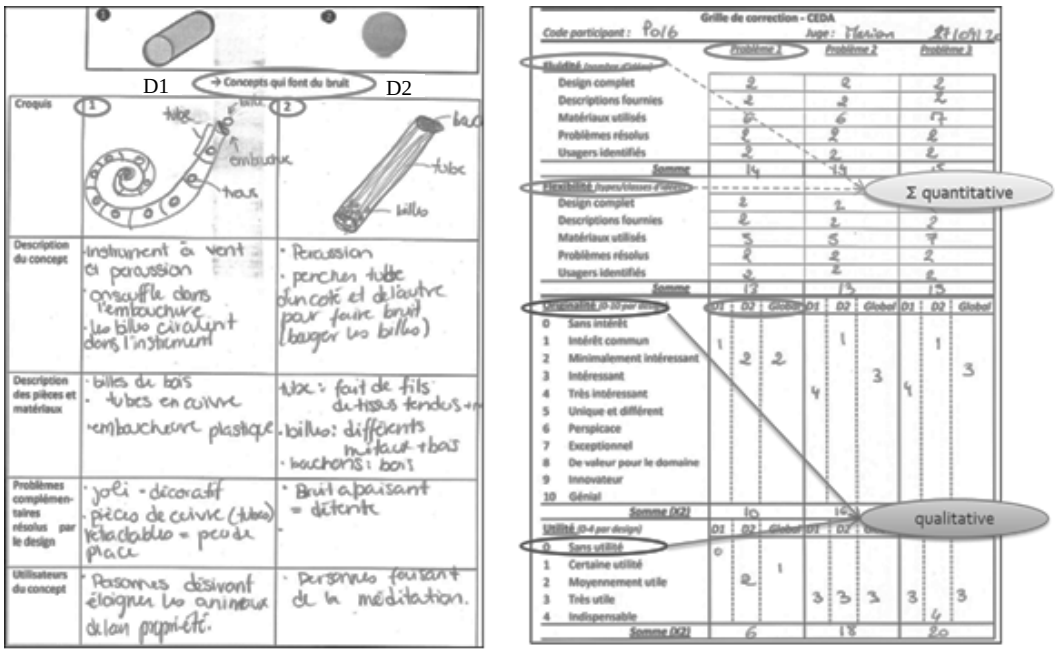

FIGURE 4
CEDA examples (A: answer sheet; B: assessment sheet).
D1 = design 1; D2 = design 2.

Phase 2

Phase 2 aimed to test the reliability of the CEDA measurement system using the gage R&R method. In Phase 2, a somewhat homogenous group (all were industrial engineering students in their second year at Polytechnique Montreal) completed the test. It was administered as a creativity exercise to prepare them for the semester's project. No financial or academic compensation was given, and all participants signed a consent form. The first author of this paper was invited by the professor in charge of the course "Integration project" as a creativity consultant to help students integrate creativity into their project.

Only two judges assessed the tests in Phase 2. The judges counted the items together instead of separately. This method allowed us to obtain faster assessments with greater calculation certainty. At that time, the judges had not discussed or made comments about the Originality or Usefulness of the designs; they kept those opinions for the individual assessments. After the first assessments were completed, judges met to discuss scores that were more than two levels apart (e.g., 2 = *somewhat interesting* and 4 = *very interesting*). They wanted to understand the differences in judgement and adapt the descriptions of the assessment criteria if necessary. Two months later, they proceeded to reassess Originality and Usefulness for the 98 tests, randomly redistributed.

Following the assessment, the gage R&R method was used a second time to evaluate more specifically the variability of the Originality and Usefulness scores. We also built control charts to verify if the scores obtained corresponded to an "in control" assessment process.

Participants

Table 1 describes the two samples of participants to our study. We had a total of 120 participants, 22 in the first phase and 98 in the second. In total, there were 54 females and 66 males, all with engineering backgrounds (94% are undergrads engineering students in different specialties). All but four participants were between 20 and 39 years old with a large majority (N = 102) between 20 and 29.

TABLE 1
Participants description

Sample description	Phase 1	Phase 2
Number of participants	22	98
Gender (female; male)	6; 16	48; 50
Engineering education (undergrads; grads)	16; 6	98; 0
Age (20-29; 30-39; 40-49; 50-59)	8; 11; 2; 1	94; 3; 1; 0

Test

Participants were given the test in one paper document. They could use any of their own crayons, pencils, erasers, and so forth. They had 30 minutes to complete the test. The CEDA is presented on four pages, three with problems and one with guidelines. There is a different general goal for each problem (i.e., design that can produce sound, design that can communicate, design that can travel). Participants had to describe two original designs for each problem (1-2) built around suggested objects (sphere, cube, cylinder, pyramid). In Figure 4-A, two designs can be seen (two columns); the general problem suggested is defined as “a concept that can produce a sound” and the two proposed objects are a sphere and a cylinder. This example represents the first of the three pages comprising the test.

RESULTS

In Phase 1, the assessment process is studied and clarified. Also, control charts allow us to compare judges’ concordance. In Phase 2, a statistical analysis shows how the test is reliable from two aspects, repeatability and reproducibility.

Assessment Clarification

The scale provided by Charyton (2014) to evaluate the two qualitative criteria, Originality and Usefulness, was a starting point but remained difficult to use because of the lack of specificity. With the sorting exercise, each level of the scale was defined more precisely according to the

answers at hand. Even with 22 tests, categories and patterns emerged (e.g., musical instruments, houses, cars, etc.). With these "subgroupings" of designs, it was possible to visualize what could be expected in the specific context of the CEDA and for each level. The same process was applied for Originality and Usefulness.

These discussions and findings were used to clarify and standardize the assessment strategy. Table 2 and Table 3 show what Charyton (2014) provided and what this study added. Another observation concerns the scale itself. No results over 6 were given or obtained.

TABLE 2
Description of the Originality criterion

From Charyton (2014)	Added in our study
0 – Dull	Does not correspond to the general goal suggested; common object (daily use)
1 – Common place	Designs that often reoccur in the tests
2 – Somewhat interesting	Minimal transformation or use of multiple suggested forms, multiple materials, added objects or materials
3 – Interesting	A more perceptive concept, but not developed enough
4 – Very interesting	Combination of two or more simple concepts
5 – Unique and different	Combination of more developed and complex concepts, concepts not existing in this suggested form
6 – Insightful	Well-developed idea, well-described (details), combining multiple concepts (different fields) in a novel way
7 – Exceptional 8 – Valuable to the field 9 – Innovative 10 – Genius	In all of our assessments, no designs obtained higher marks than 6, so we didn't have examples to discuss and compare scores 7 to 10.

TABLE 3
Description of Usefulness criterion

From Charyton (2014)	Added in our study
0 – Useless	Does not respond to the general goal suggested, does not present any possible uses
1 – Somewhat useful	Responds to the general goal suggested but has limited possible uses
2 – Useful	Relevant uses but for very specific cases
3 – Very useful	Existing concepts but not optimal/one solution among others, existing concept that needs elaboration
4 – Indispensable	Existing concepts, indispensable or integrated to modern life

The Gage R&R Method

Phase 1

Figure 5 illustrates the repeatability and reproducibility of creativity measures with the CEDA. Each chart shows the two measures made for each participant by one of the three judges. The first chart (at the top) shows that Judge 1 gave the most consistent judgments since the two lines are very close to each other. Judge 3 (bottom chart) is less consistent, and Judge 2 (middle chart) is even less consistent (see circles).

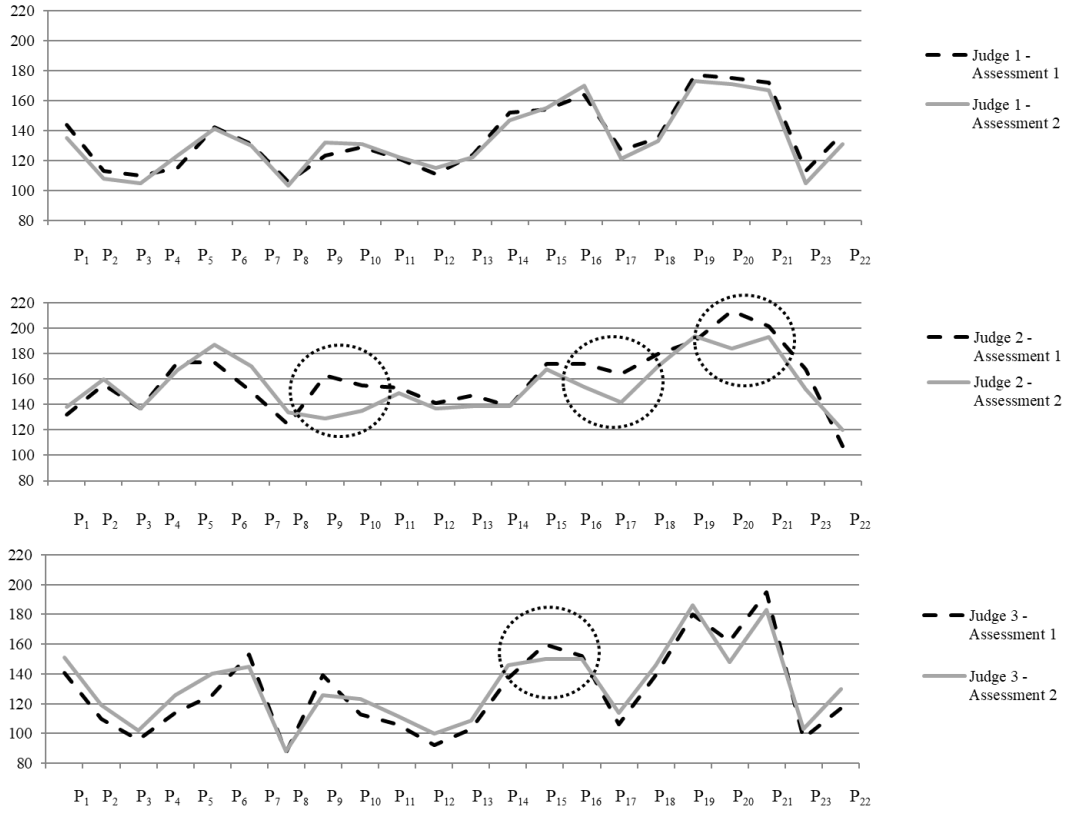

FIGURE 5
Creativity scores on the CEDA attributed by the three judges in Phase 1.
P = participant in ordinate; CEDA score in abscissa = min: 84, max: 214; P23 and P22 were reversed for analysis reasons (Judge 2 did not assess P22 accordingly).

The Pearson concordance coefficient between judges was calculated to show how similarly the three judges assessed the results for each participant (Kline, 2005). As Table 4 shows, the coefficient is higher for Judges 1 and 3. In line with this result and for organizational reasons (availability, time, cost), we felt comfortable proceeding with two judges (1 and 3), as was the case in previous studies by Charyton (Charyton et al., 2008, 2011; Charyton & Merrill, 2009). Furthermore, Charyton proceeded with two judges in all of her studies, so we believed it was a suitable and appropriate decision. Even though we have confidence that with more training Judge 2 could tighten her results, for the reasons mentioned above, we chose to continue the project with only two judges.

TABLE 4
Pearson concordance coefficient

	R^2		
Variable	$\bar{X}$ J1	$\bar{X}$ J2	$\bar{X}$ J3
$\bar{X}$ J1	1	.398	.828
$\bar{X}$ J2		1	.452
$\bar{X}$ J3			1

Note. R^2 = Pearson coefficient; J = judge.

Phase 2

The 98 test results from Phase 2 were divided randomly into four samples (25, 25, 25, 23) because the different constants used to calculate the variations in reproducibility and repeatability are established for small samples (N = 25). Control charts were built to verify if the measures were in control (i.e., predictable from a statistical point of view). Figure 6 shows visually that the values obtained are between the limits so that the assessment process is in control for the two qualitative criteria, Originality and Usefulness, confirming the test's reliability. This means that the values' variation is caused by unpredictable but normal and inevitable events and nothing specific can be done to further control the assessment process.

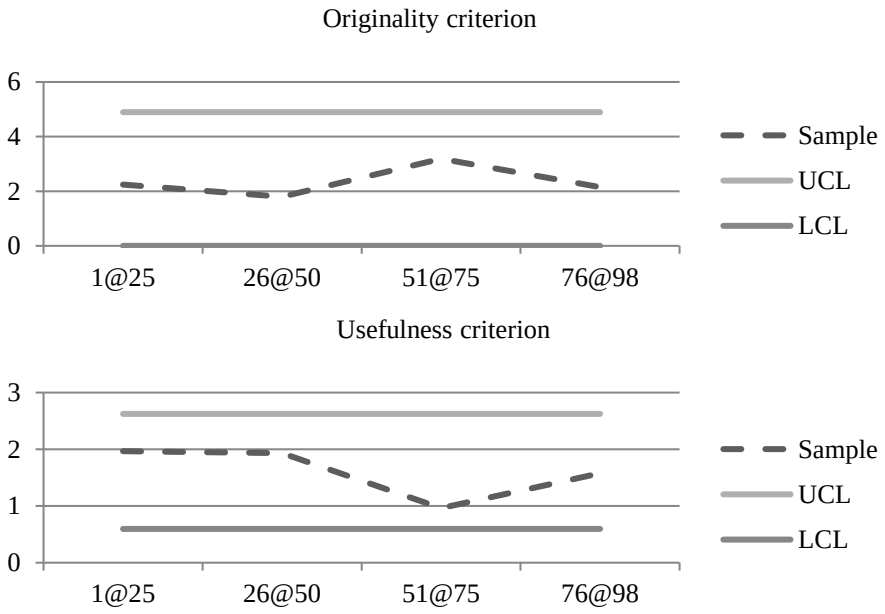

FIGURE 6
Control chart – Originality and Usefulness criteria (four samples).
UCL = upper control limit; LCL = lower control limit.

Using the same four groups of participants, Figure 7 and Figure 8 illustrate the percentages of evaluation variability related to three categories: repeatability, reproducibility, and participants. When using a measuring system, it is essential that the variability observed be due to the participants and not to the instrument itself or its use. This is what is shown in these graphs. For the two criteria Originality and Usefulness, between 85% and 95% of the variability is due to the participants and not the judges or the assessment process (repeatability and reproducibility).

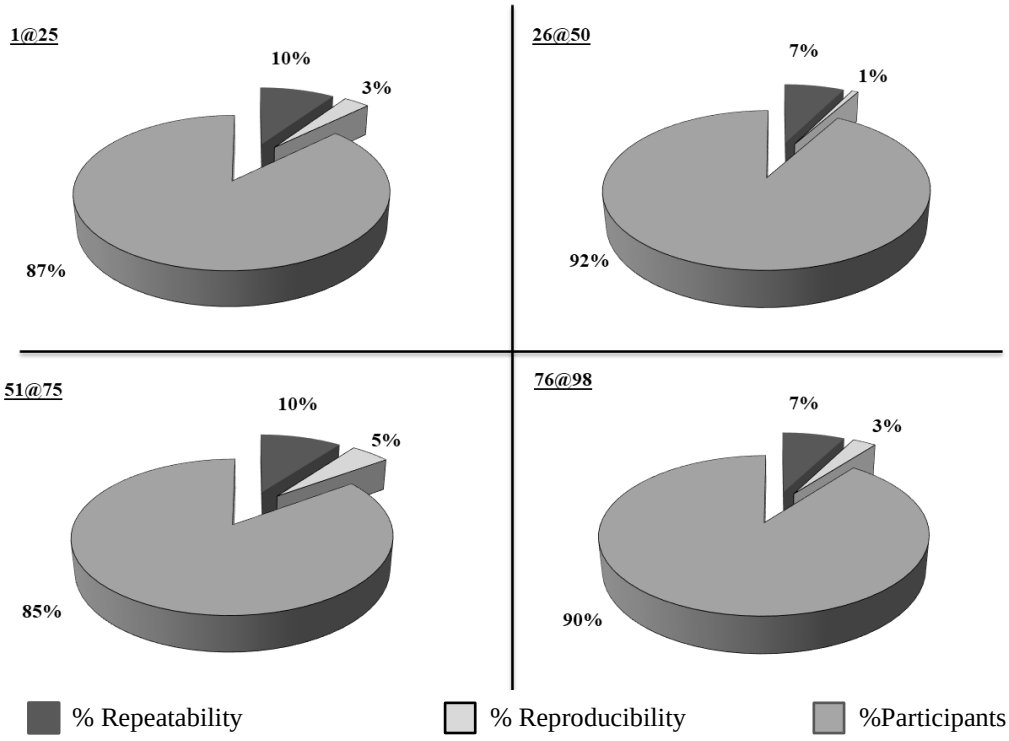

FIGURE 7
Percentages of evaluation variability for Originality (four samples) (Average global score: 88.50%).

In short, we calculated the standard deviation (*SD*), the variance, and their percentages. The measurement system is responsible for 8.50% of the total variance (see Table 5), 6.70% of the variance is caused by repeatability, and 1.80% by reproducibility. For an engineering process, a 10% limit is usually the maximum acceptable (AIAG, 2010; Ostle et al., 1996; Wheeler, 2006). This value is therefore considered acceptable. In other words, 91.50% of the variance is due to the participants' differences.

DISCUSSION AND CONCLUSION

Phase 1

With the assessment guidelines provided by Charyton (2014), the three judges performed a first round of assessments. The guidelines concerning the two criteria Fluidity and Flexibility, and how to count the items and categories, were reasonably easy to follow, but those regarding

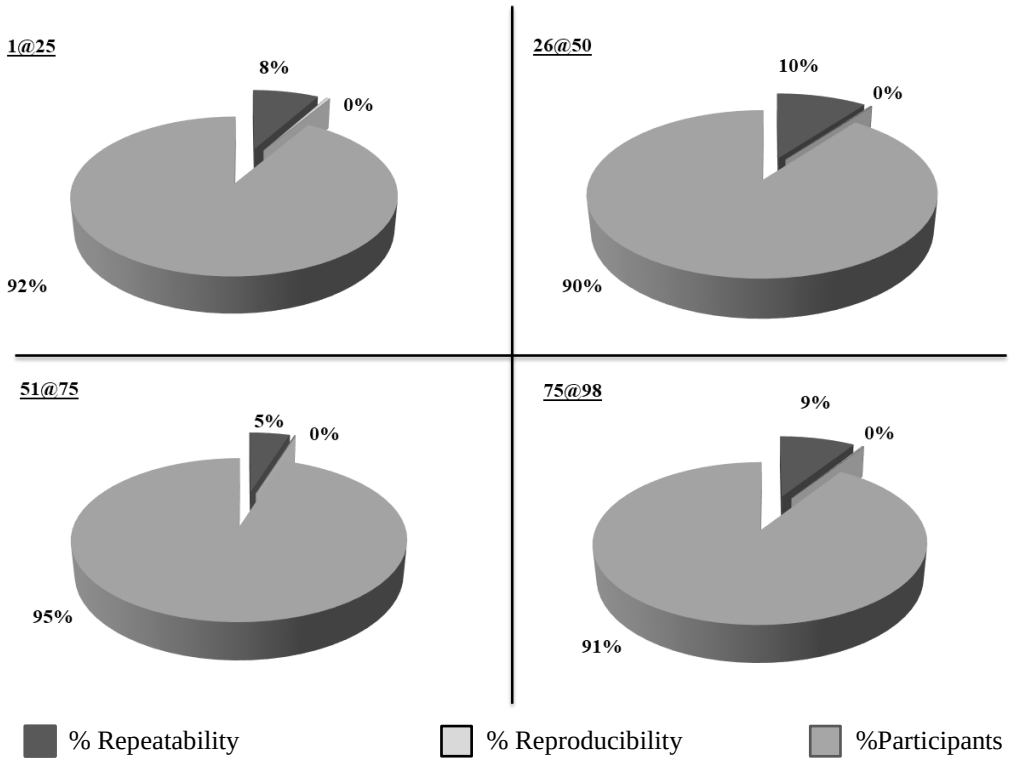

FIGURE 8
Percentages of evaluation variability for Usefulness (four samples) (Average global score: 92%).

TABLE 5
Types of variances

Variations	*SD*	Variance	% (gage R&R)	% (gage R&R total)
Repeatability: measure to measure	5.79	33.55	78.82	6.70
Reproducibility: judge to judge	3.00	9.02	21.18	1.80
Gage R&R total	8.80	42.57	100.00	8.50
Subject to subject	21.40	458.00		91.50
Total	30.20	500.57		100.00

Originality and Usefulness were much less practical. They caused assessment difficulties, requiring the judges to meet and discuss several answers. Even though a few words were provided to describe the Likert scale (Table 2 and Table 3), the judges felt they did not describe in enough detail the possibilities revealed in the tests. What is "somewhat interesting"? What is "moderately useful?" "To score Originality (Uniqueness), rate each design on the scale from 0 to 10. Scorers or judges should think of a word on your own that describes each design and then look on the rubric list to find the word and assign that number to the design" (Charyton, 2014, p. 21).

No completed assessment sheets (examples with designs and scores) have ever been published or made available to understand how the judges on Charyton's team worked. We are aware that the descriptions we came up with still leave room for personal interpretation. However, with training and practice supported by a manual containing examples (designs and scores), judges should be better equipped to make more precise assessments, specifically toward designs produced with the CEDA.

Another difficulty was the assessment of the participant's "global" performance suggested by Charyton. "Each design should be assessed separately (D1, D2). Then, an overall evaluation of the entire problem should be rated. The Originality score for the entire problem (global) will be the score that is analyzed and becomes the overall Originality score for the problem. Although each design score can be inputted and analyzed, we recommend using the overall problem score. It is also important to note that this process of scoring each design is pertinent towards making an assessment of the overall Originality score per problem" (Charyton, 2014, pp. 21-22). This explanation was not convincing and did not sufficiently describe how this overall score could or should be used. Also, according to Charyton, all scores should be added up (D1, D2, global). Given these obstacles, we made changes to this strategy in Phase 2, which we will discuss below.

A gage R&R statistical analysis was performed to see if the three judges were able to adequately repeat their assessments over time (period of two months) as well as come up with similar scores (overall precision of the assessment system). In this phase, the judges followed the aggregation method suggested by Charyton (2014). She proposes a formula adding the four scores (Fluidity, Flexibility, Originality, Usefulness), but the two scores of Originality and Usefulness are multiplied by 2. The explanation provided by the author is the following: "The correlations for the revised formula with Usefulness (2*Usefulness added to the original CEDA formula) illustrates similar findings with the new scoring of the revised CEDA compared with the previous scoring method without Usefulness" (p. 18). When asked in a personal exchange to provide more details, Charyton added "this formula was based on theory in relation to the conceptualization of Originality and Usefulness as integral components of creativity specific to engineering design" (personal communication, May 11, 2013). We still had serious reservations, so we made adjustments in Phase 2.

Phase 2

Charyton (2014) proposed an evaluation of every design but also added a third, more global, one to give an average score for each pair of designs (Figure 4-B). In collaboration with a statistician, we determined this score was unnecessary as it represented an average of two scores we already sum up. Moreover, it accentuated the gap between judges. It was an additional judgment that did not even assess a specific element of design.

An important problem arose when a specific situation occurred: if participants came up with only one of the two designs per problem, they would get a score for the first one (e.g., 3 = somewhat interesting) but would get 0 for the second one. What should the global score be? How does it adequately represent the participant's overall performance? To overcome this hurdle, we decided to eliminate the global score. Statistically it was pulling apart the judges' assessments, and theoretically it was not adding any information to the result.

As mentioned above, we felt uncomfortable using the CEDA's original scoring formula proposed by Charyton et al. (2011) because it lacked specifications. We found limited statistical and theoretical explanations as to why Originality and Usefulness numbers should be doubled as well as for the reference value of 100 for the Fluidity and Flexibility criteria. Therefore, in Phase 2, all four scores were calculated independently and no total scores were tabulated. This would allow us to conduct a more specific analysis of each criterion and keep differences between judges to a minimum. Even with these modifications, the CEDA remains a relevant tool since it provides an overall assessment of creativity (five creativity elements and four evaluation criteria), which is particularly rare in the literature on the subject.

Final Comments

This study had two objectives: to test the reliability of the CEDA measurement system and to clarify the assessment process of the CEDA. For the first objective, we used the gage R&R method to verify reliability regarding two aspects, repeatability and reproducibility. With 91.50% of the evaluation variation caused by the participants (not the judges or the test itself), the test was proven to be highly reliable. The second objective was achieved by organizing discussions between judges about the classification and scoring of the participants' design works. These yielded longer and better descriptions of the qualitative Likert scales for the Originality and Usefulness criteria, as well as a critique and a revision of the scoring process.

It was innovative to use the gage R&R method to analyze data from a psychometric test. To our knowledge, this is not a common application. It has an advantage over other types of analysis commonly used in social sciences (e.g., variance component analysis), as it can be used with a small sample.

For future research, we believe that a global score, a composite indicator of creativity, could be established with more precision and specifications. A single score would be easier to manage and work with (compare, rank, etc.) than four. Charyton (2014; Charyton et al., 2011) did use a global score, but it remains questionable for the reasons evoked above (multiplication by 2, overall problem score). Also, from a statistical point of view, the scores of Fluidity and Flexibility are always very close to each other and seem highly correlated. Should they be merged or should one be eliminated? Interesting research could be done in this direction. Finally, we started to build a visual guide with the different designs collected to facilitate and simplify the assessment process. Sketches with scores could be used to guide future judges in their assessments and help them provide comparable results from one study to another.

NOTES

1. According to Smith & Ward (2012, p. 465), divergent thinking is "The search for many varied and imaginative possible problem solutions."
2. According to Smith & Ward (2012, p. 465), convergent thinking is "Type of problem solving or reasoning in which cognitive operations are intended to converge upon the single correct answer."

FUNDING

We are grateful for funding to the first author from the Fonds de recherche du Québec – Nature et technologies (FRQNT 183790).

REFERENCES

Automotive Industry Action Group. (AIAG). (2010). *Measurement System Analysis (MSA)*. Retrieved from http://www.rubymetrology.com/add_help_doc/MSA_Reference_Manual_4th_Edition.pdf

Carpenter, W. A. (2016). *Engineering creativity: Toward an understanding of the relationship between perceptions and performance in engineering design.* (Unpublished doctoral dissertation). University of Akron, Ohio, OH.

Charyton, C. (2014). *Creative engineering design assessment: Background, directions, manual, scoring guide and uses*: London, UK: Springer.

Charyton, C., Jagacinski, R. J., & Merrill, J. A. (2008). CEDA: A research instrument for creative engineering design assessment. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 147-154. doi:10.1037/1931-3896.2.3.147

Charyton, C., Jagacinski, R. J., Merrill, J. A., Clifton, W., & Dedios, S. (2011). Assessing creativity specific to engineering with the revised creative engineering design assessment. *Journal of Engineering Education*, *100*(4), 778-799.

Charyton, C., & Merrill, J. A. (2009). Assessing general creativity and creative engineering design in first year engineering students. *Journal of Engineering Education, 98*(2),145-156.

Clary, R. M., Brzuszek, R. F., & Fulford, C. T. (2011). Measuring creativity: A case study probing rubric effectiveness for evaluation of project-based learning solutions. *Creative Education*, *2*, 333-340. doi:10.4236/ce.2011.24047

Cropley, D. H., Kaufman, J. C., & Cropley, A. J. (2011). Measuring creativity for innovation management. *Journal of Technology Management & Innovation*, *6*(3), 13-30.

Gabora, L., & Kaufman, S. B. (2010). Evolutionary approaches to creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *Cambridge handbook of creativity* (pp. 279-300). New York, NY: Cambridge University Press.

Guilford, J. P. (1968). *Creativity, intelligence and their educational implications.* San Diego, CA: EDITS-Knapp.

Harris, D. (1960). The development and validation of a creativity test in engineering. *Journal of Applied Psychology*, *44*, 254-257. doi:10.1037/h0047444

Kim, J. (2014, August). *An assessment method to evaluate team project based engineering design*. Paper presented at the 9th International Conference on Computer Science & Education, Vancouver, Canada.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, *18*, 3-14. doi:10.1207/s15326934crj1801_2

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: SAGE Publications Ltd.

Ostle, B., Turner, K. V., Hicks, C. R., & McElrath, G. W. (1996). *Engineering statistics, the industrial experience*. Belmont, CA: Duxbury Presspages.

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. doi:10.1016/j.tsc.2012. 04.009

Plucker, J. A., & Runco, M. A. (1998). The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. *Roeper Review*, *21*, 36-39. doi:10.1080/02783199809553924

Smith, S. M., & Ward, T. B. (2012). Cognition and the creation of ideas. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 456-474). New York, NY: Oxford University Press.

Treffinger, D. J., Young, G. C., Selby, E. C., & Shepardson, C. (2002). *Assessing creativity: A guide for educators.* Retrieved from http://nrcgt.uconn.edu/wp-content/uploads/sites/953/2015/04/rm02170.pdf

Wheeler, D. J. (2006). *EMP III (Evaluating the measurement process III): Using imperfect data*. Knoxville, TN: SPC Press.

Wheeler, D. J., & Chambers, D. S. (1992). *Understanding statistical process control*. Knoxville, TN: SPC Press.