Measuring creativity with planned missing data

Guillaume Fürst[1],

[1]University of Geneva, Switzerland

Correspondence concerning this article should be addressed to

Guillaume Fürst, Université de Genève

Boulevard du Pont d'Arve, 40, CH-1211 Genève 4.

E-mail : fuerstguillaume@gmail.com Phone: +41 22 379 85 58

Abstract

This paper introduces a method for the assessment of creativity that relies on creativity tasks, a subjective evaluation procedure, and a planned missing data design that offers a drastic reduction in the overall implementation costs (administration time and scoring procedure). This method was tested on a sample of 149 people, using three creativity tasks as a basis. Participants were instructed to produce several ideas in each task and then to select what they considered to be their best two ideas (i.e., “Top 2” procedure; Silvia et al. 2008). These ideas were then evaluated by a panel of peers and experts. Creativity ratings were analyzed with structural equations; measurement models were estimated for each task and correlations between factor-scores across the three tasks were investigated. Further insights regarding validity are provided through systematic investigation of the relationship between fluency scores, creativity ratings, intelligence tasks, self-reported idea generation abilities and creative activities and achievements. Overall, the results support the viability of this new approach, providing evidence of convergent and discriminant validity. They are discussed in relation to past research and avenues for further extension are proposed.

Key words

Creativity assessment; divergent thinking; subjective evaluation; peer rating; structural equation models; planned missing data

Measuring creativity with planned missing data

The measurement of creativity is a psychometric challenge. Several techniques have been suggested to approach this complex construct, all with various advantages and drawbacks. One critical, persistent issue is the overall implementation costs, especially the administration time and scoring procedure. In this paper, a brief overview of these methods is provided, with a specific focus on creativity tasks and subjective ratings. Following this review, a new measurement method, based on the planned missing data technique, is proposed and tested.

Classically, the measurement of creativity has been approached through a large array of methods (e.g., Plucker & Renzulli, 1999). These can be roughly classified into two categories: questionnaire-based methods and task-based methods. Questionnaires enable the assessment of many facets of creativity, from creative personality (e.g., Gough, 1979) to creative achievements (e.g., Carson, Peterson, & Higgins, 2005). They are generally simple to administer and their answers are easily processed. However, when answering questionnaires, participants might over- or underestimate their creativity, intentionally or not. Although this is not necessarily a major risk in the context of academic research (Silvia, Wigert, Reiter-Palmon, & Kaufman, 2012), it is clear that questionnaires are not the best choice when it comes to assessing manifest creativity (that is, the creativity of an actual, observable creative product).

Creativity tasks, either in a lab or in real-world settings, followed by an assessment of the creativity of the products, are one of the best complement to questionnaires. However, creativity tasks generally have a very high cost (both in administration time and data preparation). To make things worse, creativity tasks often have a high degree of specificity (e.g., Baer, 1993). This means that *several tasks* should be used – and, ideally, combined in a latent variable – in order to cover an adequate range of manifestations of creativity, thereby approaching a somewhat "general" creative ability. (Although the concept of general creativity remains controversial, this is not a reason to abandon all efforts regarding the conceptualization of some "broad creative ability", e.g., in different domains.) In this context, an ideal creativity protocol should encompass many questionnaires and many tasks – an ideal that is arguably more compromised by the challenges related to the tasks than to the questionnaires. Hence, in the following, we focus on creativity tasks and ways to reduce their cost without compromising their validity and reliability.

Overview of tasks and scoring methods

Creativity tasks come in many different forms. For the purpose of this paper, they are organized into three categories: (1) classic divergent thinking tasks and objective scoring; (2) the consensual assessment technique and expert rating; and (3) similar approaches based on peer or quasi-expert ratings. These techniques are reviewed below, along with an assessment of their respective advantages and drawbacks. Then we consider the various time-saving techniques that have been proposed to date.

Divergent thinking (DT) tasks have a long history in creativity research and have probably been those most widely used for assessing creativity (e.g., Guilford, Christensen, Merrifield, & Wilson, 1978; Torrance, 1966; Wallach & Kogan, 1965). They are open-ended tasks, in which participants are asked to produce a large number of different and original ideas from a basic target situation. Responses are generally scored for fluency (raw number of ideas), flexibility (variety of ideas), and originality (often scored as statistical rarity in a given sample). Although DT tasks have been criticized for being too reductionist, they have some degree of real-life predictive validity (e.g., Plucker, 1999). Nevertheless, some of their limitations are quite serious. For instance, the originality and flexibility scores are often confounded with the fluency score (e.g., Silvia, 2008; Silvia, Winterstein, & Willse, 2008). Moreover, the fact that all classic DT scores involve no direct human evaluation of creativity can be seen as problematic, since virtually all creative endeavors end up being evaluated one way or another – implicitly or explicitly, by peers, experts or an audience at large (e.g., Glăveanu, 2013).

Another popular method in creativity research is the Consensual Assessment Technique (CAT; Amabile, 1982) in which participants complete a creative product (e.g., a short story) which is then evaluated by a panel of experts in the domain. The assessment of creativity relies on the experts' implicit definition of creativity. Although this method has been used successfully in several studies (see, e.g., Baer & McKool, 2009), it can be criticized for its reliance on a single item (i.e., "creativity") and its lack of explicit definition. Moreover, the CAT is quite costly, since it is not easy to find several experts who will agree to assess dozens (or hundreds) of creative products. It could also be argued that the CAT is rather overkill for the assessment of everyday creativity; it may not be necessary to consult expert writers to assess the creativity of a short text written by non-experts in a time-frame of a few minutes (as is often the case in most creativity studies).

In fact, some researchers have suggested that, under certain circumstances, non-expert ratings can be used to assess creative products. Such approaches have been shown to be effective with, for instance, undergraduate students or gifted novices as judges (e.g., Baer, Kaufman, & Gentile, 2004; Kaufman, Lee, Baer, & Lee, 2007). Nevertheless, non-expert raters may use different implicit definitions of creativity, which can lead to validity and reliability issues (e.g., Kaufman, Baer, Cole, & Sexton, 2008). Hence, although the practical advantages are substantial (non-expert ratings cost far less and are much easier to implement in large studies), non-expert ratings should be used with caution. One important aspect is to provide detailed instructions and an explicit definition of creativity (on a related note, see also Besemer & O'Quin, 1999).

Silvia, Winterstein, Willse, et al. (2008) developed an interesting method based on the combination of divergent thinking tasks and non-expert ratings. As in classic divergent thinking tasks, participants are first asked to produce many creative ideas (hence yielding classic fluency scores). Secondly, after the task, people are asked to choose what they believe to be their two most creative responses (i.e., Top 2 ideas). Finally, these Top 2 ideas are assessed by a panel of non-expert raters. To avoid reliability and validity issues, these raters are provided with explicit instructions about how to rate creativity, which emphasize that creative ideas are *uncommon* (rare, original, unique), *remote* (not obvious and far from common ideas), and *clever* (insightful, ironic, humorous, or smart). Silvia, Winterstein, Willse, et al. (2008) have shown that this method can yield reliable scores even with a small number of raters (sometimes only 2 or 3) and that these scores are not confounded with fluency. As regards convergent validity, this study has shown that the scores in different divergent thinking tasks tend to be positively correlated (although quite modestly, mostly in the .10-.40 range) and substantially predicted by personality factors, especially by openness to experience.

Even though this method has been criticized (for a discussion, see Silvia, Winterstein, & Willse, 2008), it has been used quite productively in subsequent research (e.g., Beaty et al., 2014; Jauk, Benedek, Dunst, & Neubauer, 2013). A study dedicated to investigating the impact of variations in instructions (Benedek, Mühlmann, Jauk, & Neubauer, 2013) suggested two ways of improving the reliability of the scores: using a longer time-on-task (3 minutes or more) and more top ideas (3 or more instead of 2). These possible improvements aside, the paper by Benedek et al. essentially confirms the validity of the overall method. It therefore seems that methods mixing divergent thinking tasks and subjective evaluations are promising.

More creativity scores, lower implementation cost

Even with the method proposed by Silvia, Winterstein, Willse, et al. (2008), the problem of time-consuming procedures remains, especially in a research protocol that aims at some degree of generality and consequently requires a procedure involving several creativity tasks. A few alternative scoring techniques of divergent thinking tasks have been suggested, such as rating the creativity of the overall ideational output, instead of single ideas (Runco & Mraz, 1992). This method has been revisited recently by Silvia, Martin and Nusbaum (2009), suggesting that such a "snapshot" assessment of creativity performs quite well, but remains inferior to the methods based on detailed scoring such as the Top 2 scoring method.

Other possible strategies to reduce protocol time may rely on limiting the administration time of each task. However, excessively timed, test-like (and, by extension, stressful) conditions may be detrimental to both the manifestation of creativity (e.g., Torrance, 1988) and the reliability of creativity scores, as suggested by the study of Benedek et al. (2013) mentioned above. Time limitation may also eliminate potential differences in intrinsic motivation for the task – some participants may be uninterested and thus want to spend just a couple of minutes on a task while others may want to invest more time and effort. Given that intrinsic motivation is so important for creativity (e.g., Amabile, 1996), it is probably best to avoid compromising its potential beneficial effect through the enforcement of a strict time limitation. Actually, reducing time on task has even been used to induce need for cognitive closure, which is arguably detrimental to creativity (e.g., Kruglanski & Webster, 1996; Leung & Chiu, 2010, study 4). Overall, as regard lowering implementation costs of creativity tasks without compromising quality of measurement, it seems that creativity research has reached a plateau. Indeed, both time on task limitations and/or very simplified evaluation procedures (i.e., snapshot) have important drawbacks. In these circumstances, the technique of planned missing data could be very relevant (e.g., Graham, Taylor, Olchowski, & Cumsille, 2006). Basically, this method enables some tasks or items to be omitted for some participants, typically about 2/3, while preserving the total number of items in the study. This, of course, should be done completely at random (MCAR; see e.g., Graham, 2009). Although the analyses of datasets presenting such a structure of missing data require the use of relatively modern methods (e.g., full information maximum likelihood), these are now fairly accessible. The potential benefits are great, especially if it is used both on the “participant side” and on the “evaluation side” – i.e., participants complete only 2/3 of the tasks and raters assess only 2/3 of these ideas. This approach also avoids problems related to an excessive cognitive workload on evaluators (Forthmann et al., 2017).

Synthesis and objective of the present study

This paper proposes a procedure based on the tasks and scoring methods described above, combined with the planned missing data technique. The exact procedure is described below in the Method section, but an outline is given here.

Basically, three tasks are used: (1) a classic divergent thinking task; (2) a writing task and (3) a drawing task. These tasks were chosen to target everyday creativity (in the sense proposed by Richards, 2007) and “little-c” artistic creativity (as defined in Kaufman & Beghetto, 2009). The divergent thinking task (the cardboard boxes by Torrance, 1966) represents creativity in handiwork and DIY (reflecting an ability to solve everyday problems); the writing task represents little-c in a verbal domain; the drawing task represents little-c in a figural domain. The latter two tasks require participants to come up with small but actual creative products (and not just a list of ideas); they are thus midway between DT tasks and tasks typically used in CAT procedures.

In line with the planned missing data procedure described above, participants do not complete all the tasks, but only 2 out of 3 (3-form design; Graham et al., 2006). Likewise, the evaluators (peers and experts) assess only 2/3 of the total available ideas (more details are given in the Procedure section). Overall, the procedure provides, for each task, a fluency score and an overall subjective creativity score for each of the Top 2 ideas or products.

As regards the convergent validity, the Boxes and Story tasks (both with a strong verbal component) are expected to correlate quite strongly (about .50). Because of domain specificity (Baer, 1993; Barbot & Tinio, 2015; Kaufman & Baer, 2005), the figural task may somewhat stands alone; however, some degree of overlap with the two other tasks is nonetheless anticipated. Overall, as a further indication of validity, given the results of recent research on indicators of creativity and intelligence similar to those included in the present study (e.g., Batey, Chamorro-Premuzic, & Furnham, 2009; Jauk, Benedek, & Neubauer, 2014; Nusbaum & Silvia, 2010), it is also expected that scores derived from creativity tasks will be positively correlated with general intelligence (encompassing fluid intelligence, crystalized intelligence, and broad retrieval abilities) and with self-reported creativity. These correlations are expected to be in the .20-.40 range.

Method

Participants

The total sample consisted of 149 first-year undergraduate psychology students of the University of Geneva (122 women) between 18 and 50 years of age (M = 21.6, SD = 7.1). They participated voluntarily and without payment. The experiment was conducted in small groups of 2 to 10 people in a computer laboratory.

Overall procedure

The data collection began with the creativity tasks, administered using a paper-and-pencil method. Following a 3-form planned missing data design, each participant was given two randomly chosen creativity tasks out of three (described below). Each task was split into two phases: first, participants were invited to find as many creative ideas as possible; then they were asked to select their Top 2 ideas, which would be rated for creativity afterwards. Time was not limited but it was suggested to spend about five to fifteen minutes on each task. The mean time spent on both tasks was 26.9 minutes (SD = 8.73). After completing the creativity tasks, participants answered creativity questionnaires and completed three intelligence tasks (also described below). Creativity questionnaires were administered to all participants, but intelligence tasks were randomly assigned with the same logic as for creativity tasks (each participant completed only 2 out of 3 tasks). On average, the total data collection procedure lasted about 45 minutes.

Material

Creativity tasks

Cardboard boxes. The divergent thinking task involved cardboard boxes (Torrance, 1966). In this task, participants were asked to find many unusual, creative uses for cardboard boxes. The instructions explicitly encouraged participants to think of different types of boxes, and to find several new uses, different from those already known.

Writing task. The writing task was the “end of story task” by Lubart, Besançon, & Barbot (2011). In this task, the short beginning of a story was presented to participants (a mysterious character is walking in a park and suddenly hears a strange noise); then they were asked to propose several endings to this story. A few questions helped the participants get started (e.g., Who could this mysterious character be? What could he do after having heard the noise?). Participants were explicitly asked to find several creative short endings (just a few sentences long).

Drawing task. The drawing task was the “drawing completion” task by Lubart, Besançon, & Barbot (2011). In this task, participants were given an initial simple, “J-shaped” curve and asked to invent a drawing incorporating this basic shape. The instructions explicitly suggested that the original shape could be made smaller or bigger, rotated, or combined with any number of other elements. Again, participants were encouraged to make several creative drawings using the basic target shape.

Intelligence tasks

Fluid intelligence. Set I of Raven’s Advanced Progressive Matrices (Raven, Raven, & Court, 2003) was used to assess fluid intelligence. This consists of 12 items of mid-level difficulty, which is appropriate for adults with slightly above-average intelligence. This version of the test was chosen because it is short and arguably of an adequate difficulty level in a sample of university students. Descriptive statistics showed, however, that scores suffered from a slight ceiling effect (quartile 1 = 10; median = 11; quartile 3 = 12).

An additional indicator of intelligence was constructed on the basis of the correct answers and the time spent on these tasks (automatically measured by the computer interface). This variable was dubbed *Efficiency* and consisted of the number of correct answers divided by the time spent on the task.

Crystallized intelligence. A French version of the Mill Hill vocabulary test (Deltour, 1993) was used to assess crystallized intelligence. Given the nature of the sample, only the 12 most difficult items of this test were administered. Unlike for the Raven progressive matrices, the final scores did not suffer from any ceiling effect (quartile 1 = 4; median = 6; quartile 3 = 7).

Verbal fluency. Verbal fluency was assessed with two tasks: a *conceptual fluency* task (“name all the fruits you can think of”) and a *phonemic fluency* task (“name all the words beginning with the sound ‘fa’ you can think of”). For both tasks, time was limited to 2 minutes, which was automatically enforced by the computer interface. These two tasks (r = .33; p = .002) are combined in one single indicator in the following analyses (arithmetic mean of the standardized scores).

Creativity questionnaires

Idea generation. General idea generation ability was assessed with the scale developed by Fürst, Ghisletta, and Lubart (2014). This short 6-item scale assesses the self-reported abilities and pleasure in generating many original ideas and combinations thereof (e.g., “I easily come up with a lot of ideas”; “I like to play with ideas just for the fun of it”; “I seek to make original idea associations"). The reliability (Cronbach's alpha) of this questionnaire in the present sample was .86.

Creative activities. Interests, activities and achievements in four artistic domains were assessed using a short 4x10-item questionnaire based on similar available instruments (e.g., Jauk et al., 2014; Silvia et al., 2012; Verhaeghen, Joorman, & Khan, 2005) but specifically shortened for the purpose of this study. (An extensive inventory of several creative activities did not seem necessary in the context of the present study, focused on creativity tasks.) Specifically, this questionnaire included artistic domains only and grouped them into four rough clusters: *visual arts* (e.g., drawing, painting, photography); *writing* (e.g., writing novels, poetry, scenarios); *music* (e.g., playing an instrument, composing, singing); *performance arts* (e.g., dancing, acting).

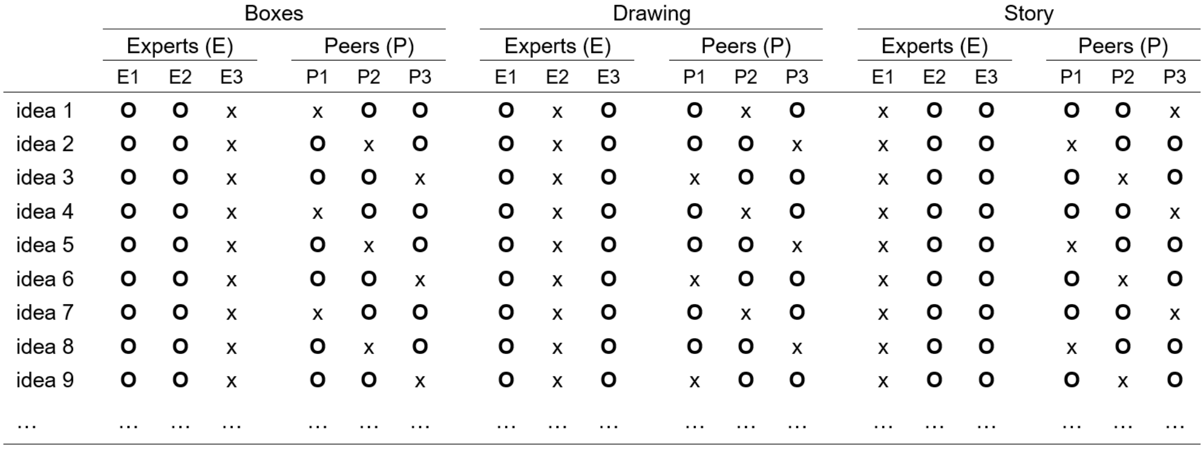
For each domain, a yes/no filter question asked participants whether they had an interest in the target domain (the mean proportion of negative answers was about 30%, and only 7.9% of participants reported no interest at all in all four domains). When participants declared an interest in a given domain, 10 additional questions were asked: 2 related to the intensity of their interest (e.g., "I have a strong interest in this domain"), 4 related to practice (e.g., "I have an active practice, daily or almost, in this domain."), and 4 related to achievements (e.g., "I have received awards in this domain."). Cronbach's alphas were .86 (visual arts), .87 (writing), .90 (music) and .91 (performance arts).

Ideas rating procedure

Following data collection, the creativity of the ideas was assessed by six raters: three of them were creativity researchers (henceforth referred to as experts); the other three raters were research assistants (henceforth referred to as peers). Once again, following a planned missing data scheme, each peer evaluated only two thirds of the ideas in each task, while the experts evaluated all the ideas in two tasks. In the end, every rater thus evaluated 66% of all the ideas, although the missing data structure was different for peers and experts. The data structure can be visualized in Table 1. This procedure was designed to ensure that every idea of each task was rated by two experts, and that all the ideas were evaluated four times. The reliability of the scores for each idea was, in most cases, satisfactory (mean Cronbach's α of .72; see Table 3 for details). The whole evaluation procedure was randomized and Little's test (Little, 1988) showed that the data were missing completely at random (χ^2 = 1972.803, DF = 2023, p = .784).

Table 1. Missing data pattern across raters.

"**O**" = available data; "**x**" = planned missing data



	Boxes						Drawing						Story					
	Experts (E)			Peers (P)			Experts (E)			Peers (P)			Experts (E)			Peers (P)		
	E1	E2	E3	P1	P2	P3	E1	E2	E3	P1	P2	P3	E1	E2	E3	P1	P2	P3
idea 1	O	O	x	x	O	O	O	x	O	O	x	O	x	O	O	O	O	x
idea 2	O	O	x	O	x	O	O	x	O	O	O	x	x	O	O	x	O	O
idea 3	O	O	x	O	O	x	O	x	O	x	O	O	x	O	O	O	x	O
idea 4	O	O	x	x	O	O	O	x	O	O	x	O	x	O	O	O	O	x
idea 5	O	O	x	O	x	O	O	x	O	O	O	x	x	O	O	x	O	O
idea 6	O	O	x	O	O	x	O	x	O	x	O	O	x	O	O	O	x	O
idea 7	O	O	x	x	O	O	O	x	O	O	x	O	x	O	O	O	O	x
idea 8	O	O	x	O	x	O	O	x	O	O	O	x	x	O	O	x	O	O
idea 9	O	O	x	O	O	x	O	x	O	x	O	O	x	O	O	O	x	O
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

The instructions given to all the raters stressed that to be rated as highly creative, an idea must be *original* (that is, infrequent, uncommon, astonishing, surprising, not obvious, offbeat, remote) and somehow *relevant* or *appropriate* (that is, perspicacious, insightful, smart, or witty, whether it respected the instructions or diverted them in a clever, ironic or humorous way). This procedure was based on analogous procedures used by Finke, Ward, & Smith (1992), Fürst, Ghisletta, & Lubart (2017), Runco & Charles (1993) and Silvia, Winterstein, Willse et al. (2008). With these definitions in mind, raters were asked to provide an initial rating using the following 5-point Likert scale: 1 = "poor/very uncreative; an idea that is unintelligible or very commonplace"; 2 = "weak/slightly creative; not as poor as category 1 but quite uncreative nonetheless"; 3 = "average/mildly creative; the idea seems somewhat creative, but something is lacking or uncertain (e.g., it is original in a 'weird' way)"; 4 = "good/quite creative; the idea is quite clearly creative, but it feels like yet more creative ideas exist"; 5 = "very good/very creative; this is a strikingly creative idea".

Raters were encouraged to consider all the ideas once before starting to sort and compare them, and also to gradually adjust the rating accordingly, ensuring that the scores roughly followed a normal distribution (about 15% of '1', 20% of '2', 30% of '3', 20% of '4', and 15% of '5'). At the end of this rating process, raters were then asked to give a very small number of ideas (about 5%) the extreme ratings of "0" or "6", for those that seemed exceptionally uncreative or creative, respectively.

Data analysis and modeling

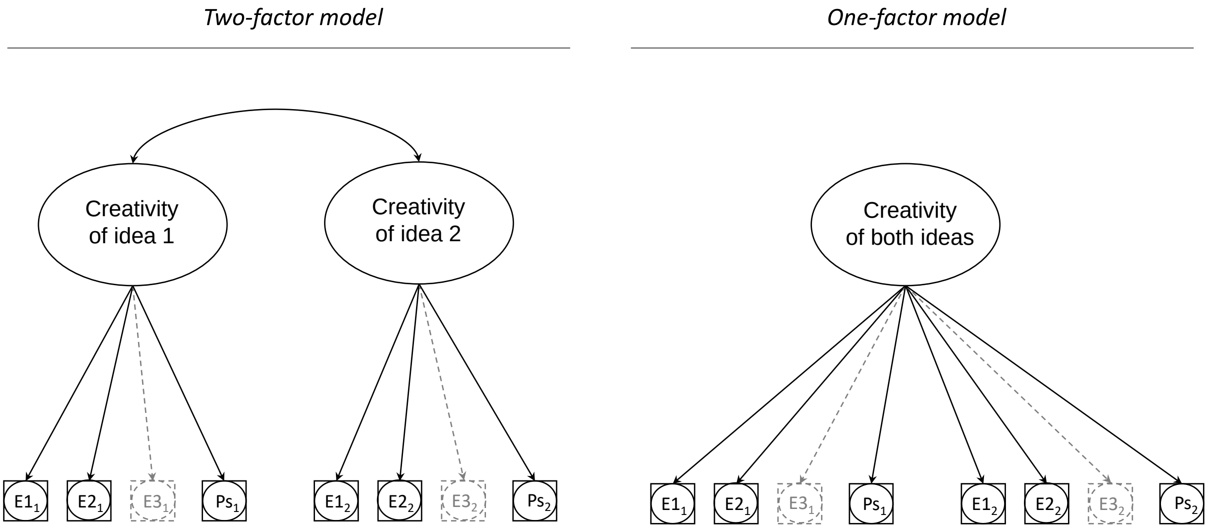
The analyses were divided into two main steps. First, separate structural equation models were estimated for each creativity task (i.e., Boxes, Story, and Drawing). These measurement models were used to investigate whether creativity in each task was adequately represented by the expert and peer ratings, and to clarify in what extent the two ideas selected as "top-2" were correlated. As depicted in Figure 1 (left panel), a factor representing the creativity of an idea was estimated on the basis of three manifest variables: two expert rating variables (without missing data) and a variable grouping two peer evaluations (arithmetic mean).[1] These two peer evaluations were randomly distributed between three raters (i.e., three raters with 33% of missing data, leading to two ratings for each idea, as represented in Table 1). For each task, the correlation between the creativity of idea 1 and the creativity of idea 2

[1] In a preliminary version of this paper, models were estimated with raw scores (that is, with 3 different manifest variables for peer ratings instead of one combining them all) but this led to a relatively poor model fit, arguably because such models were too complex given the relatively small sample size and/or the high proportion of missing data in the peer ratings variables.

was expected to be high, but not so high that a single-factor model (right panel of Figure 1) would fit the data as well as a two-factor model.

To assess model fit, the following indices and cut-off values were used: the ratio between the χ^2 and the degrees of freedom (χ^2/DF); the Root Mean Square Residual (SRMR), the Root Mean Squared Error of Approximation (RMSEA), and the Comparative Fit Index (CFI). Based on the recommendations of Hu and Bentler (1999), a model was considered to fit the data well if the fit indices reached the following cut-off values: SRMR ≤ .08; RMSEA ≤ .06; CFI ≥ .95. In line with less strict recommendations, especially relevant for small samples such as the one in this study, a model was considered acceptable if the fit indices reached the following values: χ^2/DF ≤ 2; SRMR ≤ .10; RMSEA ≤ .08; CFI ≥ .90 (Beauducel & Wittmann, 2005; Iacobucci, 2010).

Figure 1. Basic structure of measurement models.



Note. E = expert rating; Ps = peer rating. This example specifically represents models for the Boxes task; ratings of Expert 3 are not available because, by design, Expert 3 did note rate the ideas in this task (Expert 2 did not rate the ideas of the Story task; Expert 1 did not rate the ideas of the Drawing task). See text for further details about missing data. For both types of model, the loading of Expert 1 was fixed to 1 for identification.

At the end of the first step, each participant's factor scores were estimated[2] and saved to be used in the second step of the analysis. This two-step procedure was selected because the simultaneous estimation of many parameters would not have been reasonable given the small sample size. The second step basically consisted of validity analyses, that is, investigating correlations between all the creativity ratings, along with their correlations with intelligence and questionnaire variables. These analyses were based (1) on a detailed correlation matrix of all variables and (2), more critically, on the estimation of a final regression model with latent variables, enabling a synthetic view of the results and an estimation of the predictive power of intelligence tasks and creativity questionnaires, specifically for each creativity task. All the structural models were estimated with Mplus 6 (Muthén & Muthén, 2007).

Results

Measurement models

The results of model fit are summarized in Table 2. For all three creativity tasks, the two-factor model (distinguishing the two ideas) yielded a better fit. All the two-factor models had a good to excellent fit (e.g., $p > .05$; RMSEA ≤ .08; SRMR < .08; CFI ≥ .94)[3]. Key parameter estimates of these models are presented in Table 3. Factor means were all between 2.9 and 3.1. All factor loadings were significant, ranging from .41 to .77 (median = .60). Factor correlations were .45 ($p < .01$) in the Boxes task, .62 ($p < .001$) in the Story task, and .43 ($p < .01$) in the Drawing task. Following this first round of model estimation, factor scores were saved to be used in further analysis.

[2] Using the regression method (i.e., function *SAVEDATA: SAVE = FSCORES;* in Mplus)

[3] Additional models – not detailed here – were tested in order to test if residual correlation parameters (e.g., between the two ratings provided by Expert 1 [idea 1 and idea 2]) could further improve the model fit. Such additional parameters were not significant and never yielded substantial improvements in the model fit.

Table 2. Fit indices of the two-factor measurement models of the creativity tasks.

	DF	χ2	χ^2/DF	*p*-val	RMSEA (90% IC)	p RMSEA <= .05	SRMR	CFI
Boxes (1F)	9	12.54	1.39	0.18	.06 (0; .14)	0.34	0.064	0.92
Boxes (2F)	**8**	**3.00**	**0.38**	**0.93**	**0 (0; .031)**	**0.97**	**0.031**	**1.00**
Story (1F)	9	23.31	2.59	0.01	.13 (.06; .19)	0.03	0.071	0.84
Story (2F)	**8**	**9.14**	**1.14**	**0.33**	**.04 (0; .13)**	**0.51**	**0.043**	**0.99**
Draw. (1F)	9	34.42	3.82	<.001	.17 (.11; .23)	0.001	0.089	0.75
Draw. (2F)	**8**	**13.79**	**1.72**	**0.09**	**.08 (0; .16)**	**0.20**	**0.073**	**0.94**

Note. See text for details about model specification and fit indices. Best models in bold.

Table 3. Factor loadings from measurement models.

	Boxes		Story		Drawing	
	idea 1	idea 2	idea 1	idea 2	idea 1	idea 2
Descriptive statistics						
Mean	2.88	2.93	2.89	2.94	3.01	3.08
Standard error	.11	.10	.10	.09	.12	.14
ICC [a]	.63	.47	.60	.63	.71	.61
Factor loadings						
Expert 1	.53***	.62***	.55***	.38***	NA	NA
Expert 2	.69***	.41**	NA	NA	.73***	.45***
Expert 3	NA	NA	.48***	.67***	.53***	.69***
Peers [a]	.59***	.48*	.77***	.74***	.72***	.74***
Factor correlation	.45**		.62***		.43**	

Note. *: *p*<.05; **: *p*<.01; ***: *p*<.001. Unilateral *p*-values.

[a] For the computation of ICC (one-way random intra-class correlation) and the estimation of measurement models, the three peer evaluations were grouped together (arithmetic mean) in one single variable (see text and Figure 1 for further details).

Correlations

Table 4 shows that, as expected, there was some degree of correlation between all the creativity indicators, but not all the indicators were significantly correlated. First, the correlations between fluency indicators were relatively low. The correlation between fluency in the Boxes task and the Drawing task was significant (*r* = .38, *p* < .01); however, the fluency in the Story task was not significantly correlated with fluency in the other two tasks. As regards creativity ratings, only the Boxes and Story tasks were significantly correlated ($r = .46$, $p < .001$). Overall, ratings were only slightly related to fluency scores; intra-task correlations between these indicators were .18 for the Boxes task, .33 for the Drawing task and .14 (non-significant) for the Story task.

The correlations between intelligence and creativity tasks also revealed several significant coefficients, although no clear pattern emerged. Fluency in the Boxes task was related to fluid intelligence (Raven) and to verbal (phonemic and conceptual) fluency ($r = .29$ and .26, respectively). Fluency in the Story task was also correlated with fluid intelligence ($r = .28$, $p < .05$). As regards creativity ratings, scores in the Boxes task were correlated with vocabulary knowledge (Mill Hill) and verbal fluency ($r = .24$ and .38, $p < .05$ and $p < .01$ respectively). Creativity ratings in the Story task were also correlated with verbal fluency ($r = .23$, $p < .05$) while creativity ratings in the Drawing task were correlated with efficiency ($r = .26$, $p < .05$). These non-negligible exceptions aside, scores in the creativity tasks were not significantly related to intelligence, or perhaps too weakly to reach significance in this sample. There were indeed several correlations in the .10-.20 range, sometimes just at the limit of the significance level. The final model (below), investigating correlations between the creativity tasks and intelligence at the level of a latent variable, provides valuable additional insights.

As regards the correlations between the creativity tasks and the creativity questionnaires, a clearer pattern emerged. First, fluency in creativity tasks appeared virtually unrelated to the self-report measures of creativity, with the exception of the correlation between fluency in the Boxes task and creative activities in music ($r = .23$, $p < .05$). Likewise, creativity assessed with self-report was also virtually unrelated to the creativity ratings in the Boxes task. On the other hand, the creativity ratings in both the Story task and the Drawing

Table 4. Correlations between creativity tasks, intelligence tasks and creativity questionnaires.

	Creativity, fluency			*Creativity, ratings*			*Intelligence tasks*				*Creativity, self-report*				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Creativity, fluency															
1. Boxes	1														
2. Story	.22	1													
3. Draw	.38**	.16	1												
Creativity, ratings															
4. Boxes	.18*	.00	.21	1											
5. Story	.10	.14	-.11	.46***	1										
6. Draw	.25*	-.10	.33***	.11	-.09	1									
Intelligence tasks															
7. Raven	.29*	.28*	.08	.16	.13	.11	1								
8. Efficiency	.14	.18	.18	.12	.04	.26*	.34***	1							
9. Mill Hill	.12	.05	-.18	.24*	.15	-.03	.26*	.05	1						
10. Verbal fluency	.26*	.01	.12	.38**	.23*	.16	.38**	.28*	.33*						
Creativity - self-report															
11. Generation	-.03	.04	.03	.09	.27**	.18*	.03	.14	.15	.08	1				
12. Visual arts	.04	-.14	.05	-.01	.26**	.29**	-.02	.07	.10	.00	.32***	1			
13. Writing	-.08	.01	.05	.08	.25**	.18*	-.03	.10	.26**	.14	.40***	.31***	1		
14. Music	.23*	-.05	-.03	.10	.15	.10	-.05	.04	.04	.22*	.33***	.32***	.25***	1	
15. Performing arts	-.05	-.11	-.16	.07	.24**	-.16	-.33**	-.27**	-.01	.04	.14*	.27***	.21**	.36***	1

Note. *: *p*<.05; **: *p*<.01; ***: *p*<.001. Unilateral *p*-values. Pairwise missing data deletion.

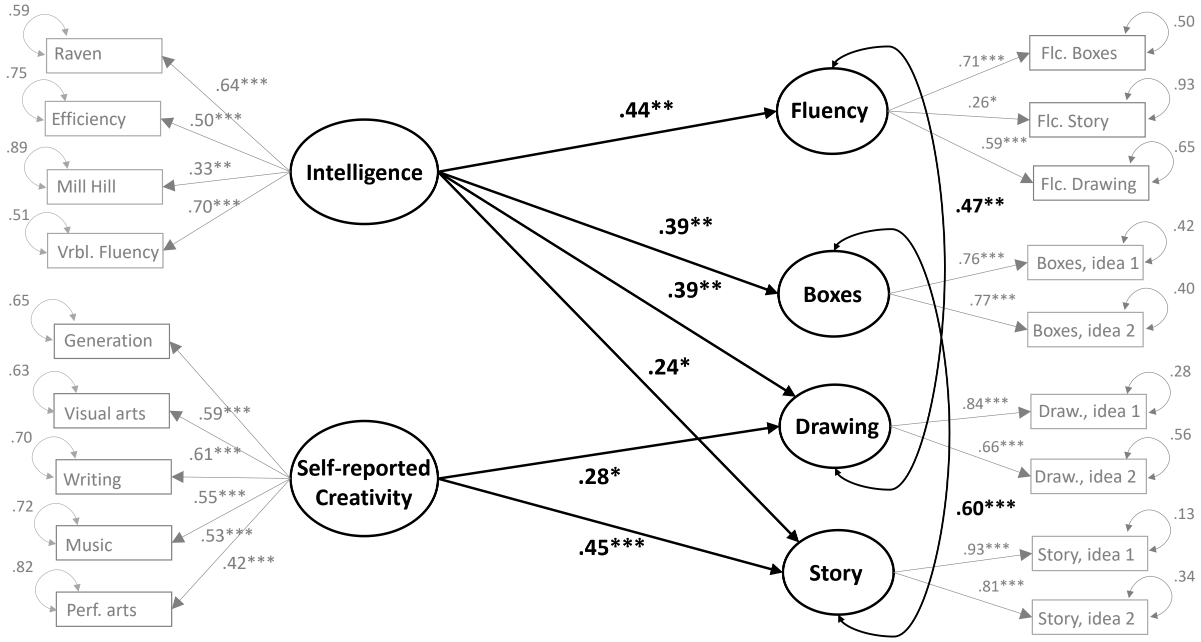
task were related to several questionnaire variables. Creativity ratings in both tasks were related to idea generation abilities (r = .27 and .18, p < .01 and p < .05, respectively). The Story task was also significantly related to creative activities and achievement in writing, in the visual arts, and in the performing arts (rs about .25, p < .01). As regards the Drawing task, it was especially correlated with self-reported creativity in the visual arts (r = .29, p < .05) and, to a lesser extent, in writing (r = .18, p < .01).

Integrative model

The final, integrative model represented in Figure 2 offers a synthetic and statistically more powerful view of these results. This model includes six latent variables: *Intelligence*, combining four indicators; *Self-reported Creativity*, combining idea generation and creative activity in four domains; *Fluency*, combining the fluency of each creativity task; and the three key creativity variables of this study, namely *Boxes*, *Story*, and *Drawing*, combining the ratings of the two best ideas in each task (i.e., factor scores saved in the first step of analyses). This model has a good to excellent fit (see footnote of Figure 2) and provides a clearer view of the relationship between the key variables of this study.

Overall, this model shows that intelligence is a significant predictor of all the creativity factors based on tasks. By contrast, the Self-reported Creativity factor is a significant predictor of only two factors based on creativity tasks. More specifically, the Fluency and Boxes factors are predicted only by Intelligence (R^2 = .19 and .15, respectively). The Drawing factor is also quite strongly predicted by Intelligence, but also by Self-reported Creativity (R^2 = .25) The Story factor has the reverse pattern; it is more strongly predicted by Self-reported Creativity than by Intelligence (R^2 = .29). Two residual correlations are significant: between Story and Boxes (r = .60, p < .001) and between Drawing and Fluency (r = .47, p < .01).

Figure 2. Integrative model of intelligence, creativity tasks, and the creativity questionnaire.



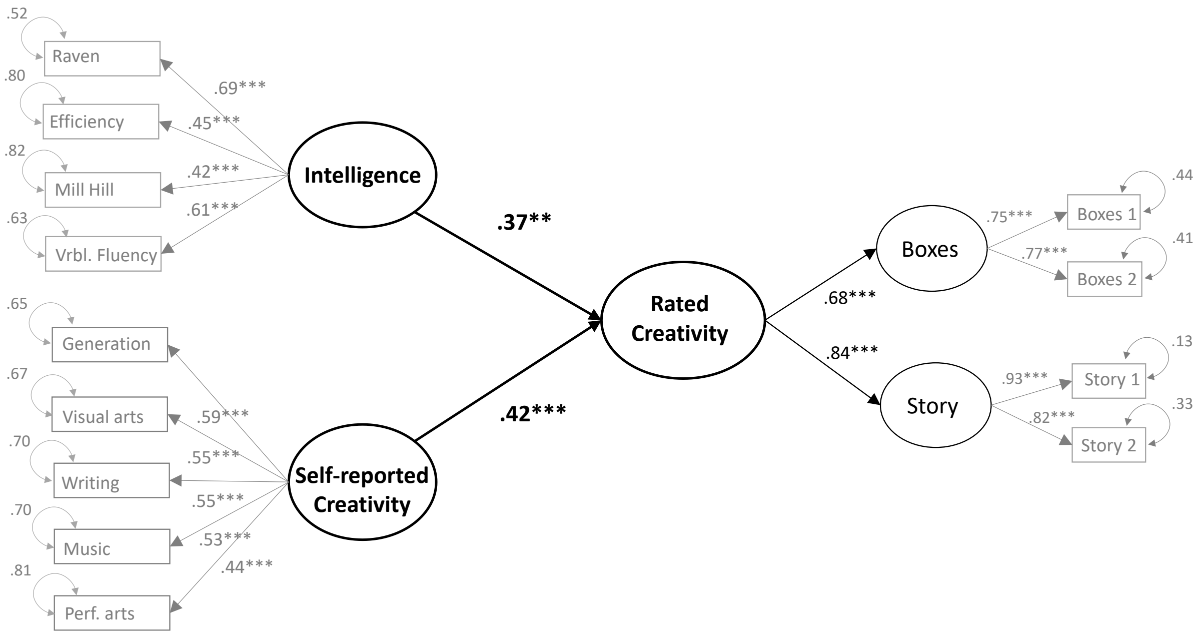
Note. *: p<.05; **: p<.01; ***: p<.001. Unilateral p-values. Details of model fit: χ^2 (107) = 117.5; p = .23; RMSEA = .026 [0; .051]; SRMR = .091; CFI = .97. For all factors, one loading was fixed to 1 for identification; for factor with two indicators only, residual variance of manifest variables were constraint equal to avoid local underidentification. All results reported on this figure are fully standardized. Additional parameters estimated but not represented on the figure (all non-significant): Self-reported creativity on Fluency; Self-reported creativity on Boxes; Drawing with Boxes; Drawing with Story; Fluency with Boxes; Fluency with Story; Fluency with Drawing; Self-reported creativity with Intelligence).

Finally, a complementary integrative model was estimated (Figure 3). This model demonstrates that it was possible to extract a second-order creativity factor, based on the Boxes and Story task – the two tasks that were the most strongly correlated[4]. This factor thus

[4] Various alternative models including the Drawing and/or the Fluency factor were also explored. However, the loadings of these constructs on the second-order creativity factor were too weak, and the overall model fit was not satisfactory.

represents some kind of broad creative abilities, mostly verbal, encompassing two tasks. This model offers a powerful synthesis and confirms that task-based creativity ratings, as assessed in this study, are quite strongly correlated with both intelligence ($\beta = .37$, $p < .01$) and self-report creativity ($\beta = .42$, $p < .01$). Together, intelligence and self-report creativity account for 34% of the variance of this broad creativity factor.

Figure 3. Alternative integrative model with second-order creativity factor.



Note. *: p<.05; **: p<.01; ***: p<.001. Unilateral p-values. Details of model fit: χ^2 (63) = 80.6; p = .07; RMSEA = .043 [0; .069]; SRMR = .087; CFI = .94. For all factors, one loading was fixed to 1 for identification; for factor with two indicators only, residual variance of manifest variables were constraint equal to avoid local underidentification. All results reported on this figure are fully standardized. Additional parameter estimated but not represented on the figure (non-significant): Self-reported creativity with Intelligence.

Discussion

The aim of this paper was to develop and test a creativity assessment procedure based on creativity tasks and subjective ratings, implemented in the context of a planned missing data design. Overall, the results are support the viability of this new approach.

First, despite a relatively limited number of expert raters and the use of peer-rating, the scores' reliability and the fit of measurement models were satisfactory. With most ICC values ranging from .61 to .71 (the only exception being idea 2 in the Boxes task, with ICC of .47), inter-rater agreement was, overall, fair to good (Cicchetti, 1994). For all tasks, the best factorial structure was a two-factor model, distinguishing between each Top 2 idea. The fit indices of these models were all good to excellent. The peer-rating variable converged quite well with the expert-rating one and did not appear to be especially weaker – quite the contrary, actually (probably because this variable uses two informants rather than one, thereby being more reliable).

Moreover, the two ideas in each task were systematically positively correlated, enabling the estimation of a second-order factor representing overall creativity in each task. Taken together, these results provide evidence in favor of the reliability of this approach and are in line with past research suggesting that a relatively few number of raters – and not necessarily experts – can be sufficient to assess creativity reliably, as long as these raters are provided with specific instructions (e.g., Baer et al., 2004; Silvia et al., 2008).

The validity of the scores was first investigated through correlational analysis. First, correlations between creativity ratings and fluency indices were relatively low; this is in line with Silvia, Winterstein, Willse, et al. (2008) and probably reflects a strength of the peer rating procedure, not a weakness. It is also possible, however, that the fluency scores in the drawing and writing tasks had slightly different properties than those of the classic divergent thinking task. Indeed, participants had to develop their idea significantly in those tasks, which may have impacted the very nature of the fluency score. Notwithstanding this possibility, other results (those reported on Figure 2 in particular) support the validity of these scores. Second, regarding the correlations between ratings of each task, the results show that only the Boxes and Story tasks were quite strongly and significantly correlated. This constitutes the first piece of evidence in favor of the convergent validity of these tasks. However, the fact that the Drawing task was not correlated with these two tasks does not symmetrically preclude its validity – this lack of correlation is probably due to domain specificity (e.g., Baer, 1993).

Correlations between the creativity tasks and intelligence were somewhat inconsistent. Fluency indices were correlated with reasoning abilities and efficiency, as well as with verbal fluency, but not systematically. Creativity ratings were correlated with reasoning efficiency, vocabulary knowledge, and verbal fluency, but not systematically either. This lack of a systematic pattern may be due to the statistical power issues related to the use of a planned missing data design in a relatively small sample (most pairwise correlations involving a creativity task and an intelligence task were estimated on a basis of N = 50). This limitation was not as severe when considering latent variable models (more on this further below).

Issues related to a low statistical power were not as serious when considering the correlations between the creativity tasks and the creativity questionnaire, since all participants completed all creativity questionnaires. Moreover, a quite clear pattern emerged: all creative fluency indices and creativity ratings in the Boxes task were unrelated to self-reported creativity. By contrast, creativity ratings in both the Story and the Drawing tasks were significantly related to most indicators, sometimes with very specific evidence of validity (e.g., the Drawing task was especially correlated with creative activities in the visual arts).

Beyond the detailed analysis of individual zero-order correlations, it is the first integrative model (Figure 2) that provides the most conclusive validity evidence. First, the fact that fluency scores were best represented in a specific Fluency factor,[5] rather than combined with rating scores, indicates that creativity scores based on ratings are definitely not confounded with fluency (again in line with Silvia et al., 2008). Moreover, the positive relationship between Fluency and Intelligence is typical of the results of past research (e.g., Plucker, 1999; Silvia, 2008) thus suggesting that the underlying fluency scores are quite like the fluency scores in classic divergent thinking tasks. The fact that this Fluency factor was not predicted by Self-reported creativity is in line with a recent study (Jauk et al., 2014) that found that fluency only weakly predicted creative activities and achievement. However, the relevance of the present study to this specific matter may be limited because the questionnaire of creative activities and achievements was a somewhat simplified one.

More importantly, the model depicted on Figure 2 also provides strong evidence in favor of the validity – both discriminant and convergent – of the creativity ratings. All three factors (Boxes, Drawing and Story) were significantly predicted by intelligence; likewise, creativity in the last integrative model (Figure 3) was also predicted by intelligence[6]. These

[5] It is worth noting that despite the low pairwise correlations between fluency in the three creativity tasks (reported in Table 4), the loadings of the fluency factor estimated on the basis of the whole sample were all significant. This is a good illustration of the power of planned missing data and FIML algorithm.

[6] It may be argued that in both integrative models (Figure 2 and Figure 3), the predictive power of intelligence is boosted by inclusion of verbal fluency as an indicator of the intelligence latent variable. In the first integrative model (Figure 2), without verbal fluency as an indicator of intelligence, the predictions of intelligence are the following: on fluency in creativity tasks: $\beta = .33$, p<.05 (instead of .44, $p < .01$); on Boxes: $\beta = .12$, *ns* (instead of .39, $p < .01$); on Drawing: $\beta = .22$, $p < .05$ (instead of .39, $p < .01$); on Story: $\beta = .24$, $p < .05$ (unchanged). In the last model (Figure 3), without verbal fluency as an indicator of intelligence, the prediction of intelligence is $\beta = .30$, $p < .05$ (instead of .37, $p < .01$). Hence, it is true that, to a certain extent, verbal fluency boosts the predictive power of intelligence. However, even without verbal fluency, intelligence remains a significant predictor of creativity in most cases. Moreover, theoretically speaking, it can be argued that the notion of “general intelligence” is better represented together with verbal fluency (i.e., broad retrieval abilities).

results are in line with many recent studies (e.g., Batey et al., 2009; Jauk et al., 2014; Nusbaum & Silvia, 2010; Silvia, 2008). In addition, both the Drawing and the Story tasks were significantly predicted by the Self-reported Creativity factor. These specific results suggest that the requirement for idea development in these tasks carries some important implications for validity; both tasks seem closer to "real-life" creativity than the classic divergent thinking task (Boxes). Finally, the drawing task was also correlated with the Fluency factor, thus providing supplementary evidence of validity for both constructs[7].

Taken together, these results show that no single creativity task is redundant with another or interchangeable. This strongly supports the inclusion of several creativity tasks in research protocols, and hence – given the inevitable subsequent increase in cost – demonstrates the relevance of the method discussed here. Nevertheless, even though every task seems to have a large degree of specificity, the present results also suggest that the domain (or task) specificity of creativity measures should not be exaggerated. As mentioned in the introduction, it seems possible to conceive of some "broad creative abilities", at least at the everyday creativity level (Kaufman & Baer, 2005; Kaufman & Beghetto, 2009).

Indeed, the final model (Figure 3) has shown that it was possible to estimate a second-order factor representing such "broad creative abilities". This broad creativity factor combines four creative ideas (or products) in two different tasks, all rated by three experts and three peers, hence ensuring a relatively comprehensive assessment of creativity. The R^2 of this latent variable was relatively high (about one third of explained variance); both intelligence

[7] Generally speaking, it must be acknowledged that r or β values in the .30-.50 range are relatively low for convergent validity. However, creativity research faces a difficult trade-off: either we use very similar tasks and have great convergent validity (and excessively high domain or task specificity), or we use more diverse tasks and necessarily end up with lower figures. The right balance is hard to find, especially when, given the high logistic cost of creativity tasks, it is almost impossible to design studies including more than three or four tasks.

and self-reported creativity independently and substantially contributed to it. This provides one more piece of evidence that the method proposed in this paper is viable and makes sense. The relatively “low cost” protocol of this study was able to replicate results that are very similar to the best recent studies on creativity.

Yet, although the method discussed in this paper is very promising, a few words of caution are in order. First, the planned missing data approach may entail some challenges for the computation of the reliability of ratings, because every single participant has a least one missing data and most software uses listwise deletion when calculating reliability indices. Reliability, then, can either be calculated “by hand” [8] or, in order to avoid the listwise deletion problem, using variables that average two or more raters. Both methods were used in this paper (although only the latter is reported in Table 3) and both lead to similar results[9].

Second, it is critical for the design to ensure that the missing data are missing completely at random (MCAR). One specific reason for this is that a substantial proportion of the variance of creativity ratings is due to participant × rater interaction (Silvia, Winterstein, Willse, et al., 2008). Hence, if the same participants are systematically evaluated by the same raters, severe systematic error (bias) may occur in the estimation of creativity scores.

Third, the number of raters should not be too low. In this regard, the present study probably edges towards a lower limit; fewer raters would presumably raise reliability issues. Without missing data, three expert raters may be enough, but with peer rating only and a planned missing data design, it is probably safer to aim for five or six raters.

It is also essential for the planned missing data design to be coupled with the estimation of latent variables (using the FIML algorithm). Indeed, unless such latent variables

[8] Compute each correlation between every pair of raters, then compute the mean of all these correlations, then compute the standardized Cronbach's alpha. This method yields slightly higher reliability values.

[9] Yet another method, based on multilevel modeling, has been proposed by Putka, McCloy, & Diaz (2008).

are estimated, the planned missing data approach is of little interest. This approach makes sense for latent variables based on several indicators, and random missing data across all these indicators. The advantages of the planned missing data approach are best illustrated by the final model in this paper (Figure 3).

To a certain extent, the latent variable approach also allows for the correction of unreliability, since in such factorial models idiosyncrasies of a specific expert's rating is represented in the error variance parameter of each manifest variable, not in the factor score. Yet, the latent variable approach used here is not necessarily the ultimate solution to all potential issues. For one thing, reliability may be further improved using multidimensional item response models (Wang, Chen, & Cheng, 2004). Also, generalizability theory and multilevel models may constitute valuable complementary frameworks to assess specific issues such as potential rater drift (e.g., the impact of extreme ratings in the context of research designs in which mean comparisons are of main interest).

Finally, other possible caveats may imply sample size and choice of creativity tasks. With respect to sample size, this study again represents a lower limit – although sample size requirements fundamentally depend on effect sizes and model complexity. Nonetheless, the relatively small sample size of the present study has led to important limitations. Structural models had to be estimated through a two-step procedure, and the peer ratings with missing data had to be averaged into a single variable in order to achieve acceptable model fit. As regards the sampling of creativity tasks, it is arguably preferable to choose tasks that can be grouped together in second-order factors, aiming at "broad creativity abilities", which would also be potentially compatible with the approach proposed by Wang et al. (2004).

Conclusion

Despite potential limitations, the method introduced in this paper appears viable and promising. Since it has performed to satisfaction in this "proof of concept" study, it can be expected to work well in larger samples, with more creativity tasks, and more advanced data analysis. Moreover, there is no reason to restrict the use of this method to the assessment of everyday creativity. Such a multi-task and multi-rater approach implemented in a planned missing data design can be adapted to any domain and any level of creativity, thus opening the way for further insights into the demanding world of creativity assessment.

References

Amabile, T. M. (1982). Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology, 43*(5), 997-1013.

Amabile, T. M. (1996). *Creativity in context: Update to The Social Psychology of Creativity*. Boulder, CO, US: Westview Press.

Barbot, B., & Tinio, P. P. (2015). Where is the “g” in creativity? A specialization–differentiation hypothesis. *Frontiers in human neuroscience*, *8*, article 1041.

Baer, J. (1993). *Creativity and divergent thinking. A task-specific approach.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the Consensual Assessment Technique to Nonparallel Creative Products. *Creativity Research Journal*, *16*(1), 113-117.

Baer, J., & McKool, S. S. (2009). Assessing Creativity Using the Consensual Assessment Technique. In C. S. Schreiner (Ed.), *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education* (p. 65-77). London, UK: IGI Global.

Batey, M., Chamorro-Premuzic, T., & Furnham, A. (2009). Intelligence and personality as predictors of divergent thinking: The role of general, fluid and crystallised intelligence. *Thinking Skills and Creativity*, *4*(1), 60-69.

Beaty, R. E., Benedek, M., Wilkins, R. W., Jauk, E., Fink, A., Silvia, P. J., … Neubauer, A. C. (2014). Creativity and the default network: A functional connectivity analysis of the creative brain at rest. *Neuropsychologia*, *64*, 92-98.

Beauducel, A., & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data with Slightly Distorted Simple Structure. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 41-75.

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of Divergent Thinking by means of the Subjective Top-Scoring Method: Effects of the Number of Top-

Ideas and Time-on-Task on Reliability and Validity. *Psychology of aesthetics, creativity, and the arts*, *7*(4), 341-349.

Besemer, S. P., & O'Quin, K. (1999). Confirming the Three-Factor Creative Product Analysis Matrix Model in an American Sample. *Creativity Research Journal, 12*(4), 287-296.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, Validity, and Factor Structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, *17*(1), 37‑50.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290.

Deltour, J. J. (1993). *Echelle de vocabulaire Mill Hill de J. C. Raven: Adaptation française et normes comparées du Mill Hill et du Standard Progressive Matrices (PM38).* Braine le Château, Belgique: Application des Techniques Modernes.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*(Supplement C), 129‑139.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. Cambridge, MA, US: MIT Press.

Fürst, G., Ghisletta, P., & Lubart, T. (2014). Toward an Integrative Model of Creativity and Personality: Theoretical Suggestions and Preliminary Empirical Testing. *The Journal of Creative Behavior*, *50*(2), 87‑108.

Fürst, G., Ghisletta, P., & Lubart, T. (2017). An experimental study of the creative process in writing. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 202‑215.

Gough, H. G. (1979). A creative personality scale for the Adjective Check List. *Journal of Personality and Social Psychology*, *37*(8), 1398-1405.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323-343.

Guilford, J. P., Christensen, P. R., Merrifield, P. R., & Wilson, R. C. (1978). *Alternate Uses: Manual of instructions and interpretations*. Orange, CA: Sheridan Psychological Services.

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, *6*(1), 1-55.

Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90-98.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, *41*(4), 212-221.

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The Road to Creative Achievement: A Latent Variable Model of Ability and Personality Predictors. *European Journal of Personality, 28*(1), 95-105.

Kaufman, J. C., & Baer, J. (2005). The Amusement Park Theory of Creativity. In J. C. Kaufman & J. Baer (Ed.), *Creativity across domains: Faces of the muse.* (p. 321-328). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

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.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology*, *13*(1), 1‑12.

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.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind:" Seizing" and" freezing.". *Psychological review*, 103(2), 263-283.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404), 1198‑1202.

Lubart, T., Besançon, M., & Barbot, B. (2011). *Evaluation du Potentiel Créatif [Evaluation of Creative Potential]*. Paris, France: Hogrefe.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus 6*. Los Angeles, CA: Muthén & Muthén.

Nusbaum, E. C., & Silvia, P. J. (2010). Are intelligence and creativity really so different?: Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence*, *39*(1), 36‑45.

Plucker, J. A. (1999). Is the Proof in the Pudding? Reanalyses of Torrance's (1958 to Present) Longitudinal Data. *Creativity Research Journal*, *12*(2), 103‑114.

Plucker, J. A., & Renzulli, J. S. (1999). Psychometric Approaches to the Study of Human Creativity. In *Handbook of creativity* (p. 35‑61).

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. Jo*urnal of Applied Psychology*, 93, 959‑981.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* San Antonio, TX: Harcourt Assessment.

Richards, R. (Ed.). (2007). *Everyday Creativity and New Views of Human Nature: Psychological, Social, and Spiritual Perspectives*. Washington, DC: American Psychological Association.

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, *15*(5), 537‑546.

Runco, M. A., & Mraz, W. (1992). Scoring Divergent Thinking Tests Using Total Ideational Output and a Creativity Index. *Educational and Psychological Measurement*, *52*(1), 213‑221.

Silvia, P. J. (2008). Creativity and Intelligence Revisited: A Latent Variable Analysis of Wallach and Kogan (1965). *Creativity Research Journal*, *20*(1), 34‑39.

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*(2), 79‑85.

Silvia, P. J., Wigert, B., Reiter-Palmon, R., & Kaufman, J. C. (2012). Assessing creativity with self-report scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 19‑34.

Silvia, P. J., Winterstein, B. P., & Willse, J. T. (2008). Rejoinder: The madness to our method: Some thoughts on divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 109‑114.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (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.

Torrance, E. P. (1966). *The Torrance Tests of Creative Thinking-Norms-Technical Manual Research Edition-Verbal Tests, Forms A and B-Figural Tests, Forms A and B*. Princeton, NJ: Personnel Press.

Verhaeghen, P., Joorman, J., & Khan, R. (2005). Why we sing the blues: the relation between self-reflective rumination, mood, and creativity. *Emotion*, *5*(2), 226-232.

Wallach, M., & Kogan, N. (1965). *Modes of thinking in young children*. New York, NY, US: Holt, Rinehart and Winston.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. Psychological Methods, 9(1), 116.

Tables

Table 1. Missing data pattern across raters

	Boxes						Drawing						Story					
	Experts (E)			Peers (P)			Experts (E)			Peers (P)			Experts (E)			Peers (P)		
	E1	E2	E3	P1	P2	P3	E1	E2	E3	P1	P2	P3	E1	E2	E3	P1	P2	P3
idea 1	**O**	**O**	x	x	**O**	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	**O**	**O**	**O**	x
idea 2	**O**	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	x	**O**	**O**	x	**O**	**O**
idea 3	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	**O**
idea 4	**O**	**O**	x	x	**O**	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	**O**	**O**	**O**	x
idea 5	**O**	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	x	**O**	**O**	x	**O**	**O**
idea 6	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	**O**
idea 7	**O**	**O**	x	x	**O**	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	**O**	**O**	**O**	x
idea 8	**O**	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	x	**O**	**O**	x	**O**	**O**
idea 9	**O**	**O**	x	**O**	**O**	x	**O**	x	**O**	x	**O**	**O**	x	**O**	**O**	**O**	x	**O**
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Note. “O” = available data; “X” = planned missing data.

Table 2. Fit indices of the two-factor measurement models of the creativity tasks.

	DF	χ2	χ^2/DF	*p*-val	RMSEA (90% IC)	p RMSEA <= .05	SRMR	CFI
Boxes (1F)	9	12.54	1.39	0.18	.06 (0; .14)	0.34	0.064	0.92
Boxes (2F)	**8**	**3.00**	**0.38**	**0.93**	**0 (0; .031)**	**0.97**	**0.031**	**1.00**
Story (1F)	9	23.31	2.59	0.01	.13 (.06; .19)	0.03	0.071	0.84
Story (2F)	**8**	**9.14**	**1.14**	**0.33**	**.04 (0; .13)**	**0.51**	**0.043**	**0.99**
Draw. (1F)	9	34.42	3.82	<.001	.17 (.11; .23)	0.001	0.089	0.75
Draw. (2F)	**8**	**13.79**	**1.72**	**0.09**	**.08 (0; .16)**	**0.20**	**0.073**	**0.94**

Note. See text for details about model specification and fit indices. Best models in bold.

Table 3. Factor loadings from measurement models.

	Boxes		Story		Drawing	
	idea 1	idea 2	idea 1	idea 2	idea 1	idea 2
Descriptive statistics						
Mean	2.88	2.93	2.89	2.94	3.01	3.08
Standard error	.11	.10	.10	.09	.12	.14
ICC [a]	.63	.47	.60	.63	.71	.61
Factor loadings						
Expert 1	.53***	.62***	.55***	.38***	NA	NA
Expert 2	.69***	.41**	NA	NA	.73***	.45***
Expert 3	NA	NA	.48***	.67***	.53***	.69***
Peers [a]	.59***	.48*	.77***	.74***	.72***	.74***
Factor correlation	.45**		.62***		.43**	

Note. *: p<.05; **: p<.01; ***: p<.001. Unilateral p-values.

[a] For the computation of ICC (one-way random intra-class correlation) and the estimation of measurement models, the three peer evaluations were grouped together (arithmetic mean) in one single variable (see text and Figure 1 for further details).

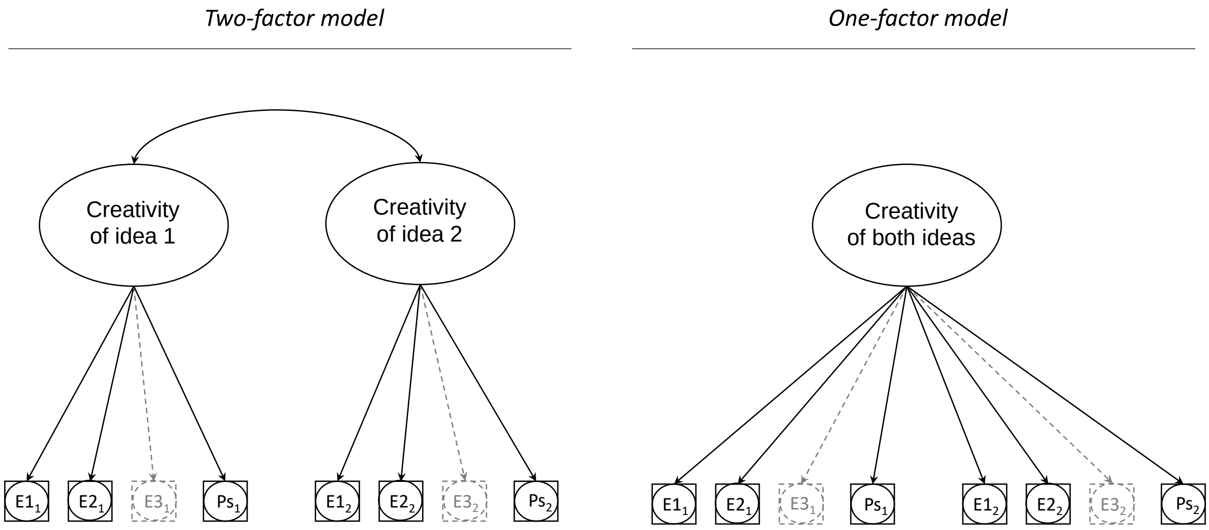
Table 4. Correlations between creativity tasks, intelligence tasks and creativity questionnaires.

	Creativity, fluency			*Creativity, ratings*			*Intelligence tasks*				*Creativity, self-report*				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Creativity, fluency															
1. Boxes	1														
2. Story	.22	1													
3. Draw	.38**	.16	1												
Creativity, ratings															
4. Boxes	.18*	.00	.21	1											
5. Story	.10	.14	-.11	.46***	1										
6. Draw	.25*	-.10	.33***	.11	-.09	1									
Intelligence tasks															
7. Raven	.29*	.28*	.08	.16	.13	.11	1								
8. Efficiency	.14	.18	.18	.12	.04	.26*	.34***	1							
9. Mill Hill	.12	.05	-.18	.24*	.15	-.03	.26*	.05	1						
10. Verbal fluency	.26*	.01	.12	.38**	.23*	.16	.38**	.28*	.33*						
Creativity - self-report															
11. Generation	-.03	.04	.03	.09	.27**	.18*	.03	.14	.15	.08	1				
12. Visual arts	.04	-.14	.05	-.01	.26**	.29**	-.02	.07	.10	.00	.32***	1			
13. Writing	-.08	.01	.05	.08	.25**	.18*	-.03	.10	.26**	.14	.40***	.31***	1		
14. Music	.23*	-.05	-.03	.10	.15	.10	-.05	.04	.04	.22*	.33***	.32***	.25***	1	
15. Performing arts	-.05	-.11	-.16	.07	.24**	-.16	-.33**	-.27**	-.01	.04	.14*	.27***	.21**	.36***	1

Note. *: p<.05; **: p<.01; ***: p<.001. Unilateral p-values. Pairwise missing data deletion.

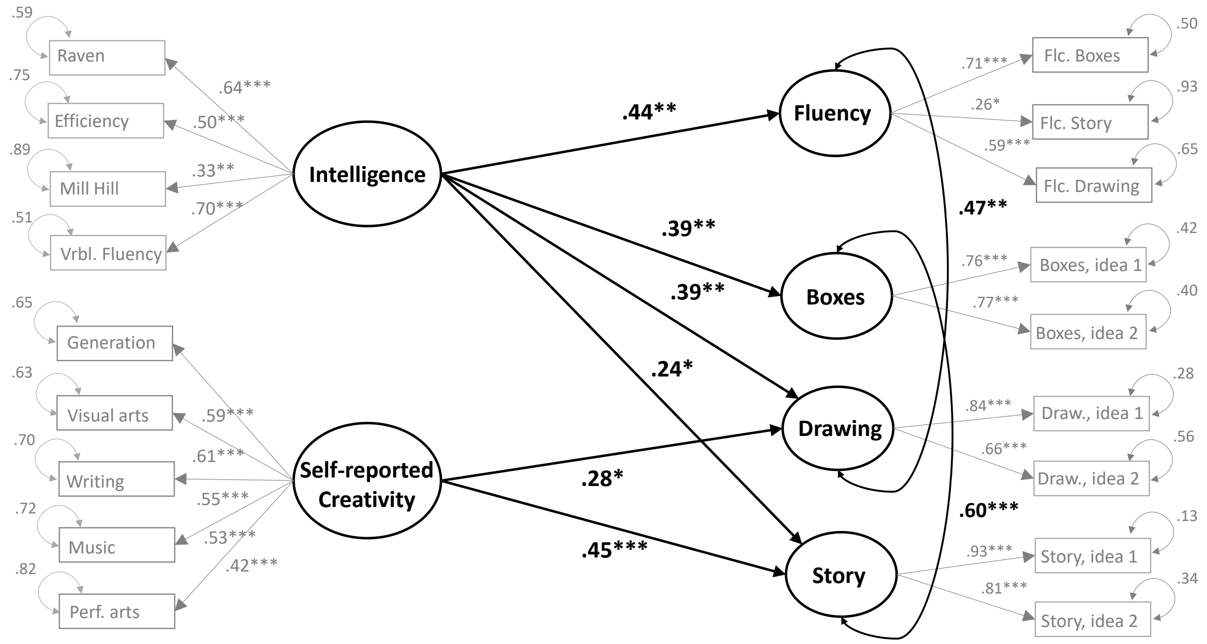
Figures

Figure 1. Basic structure of measurement models.



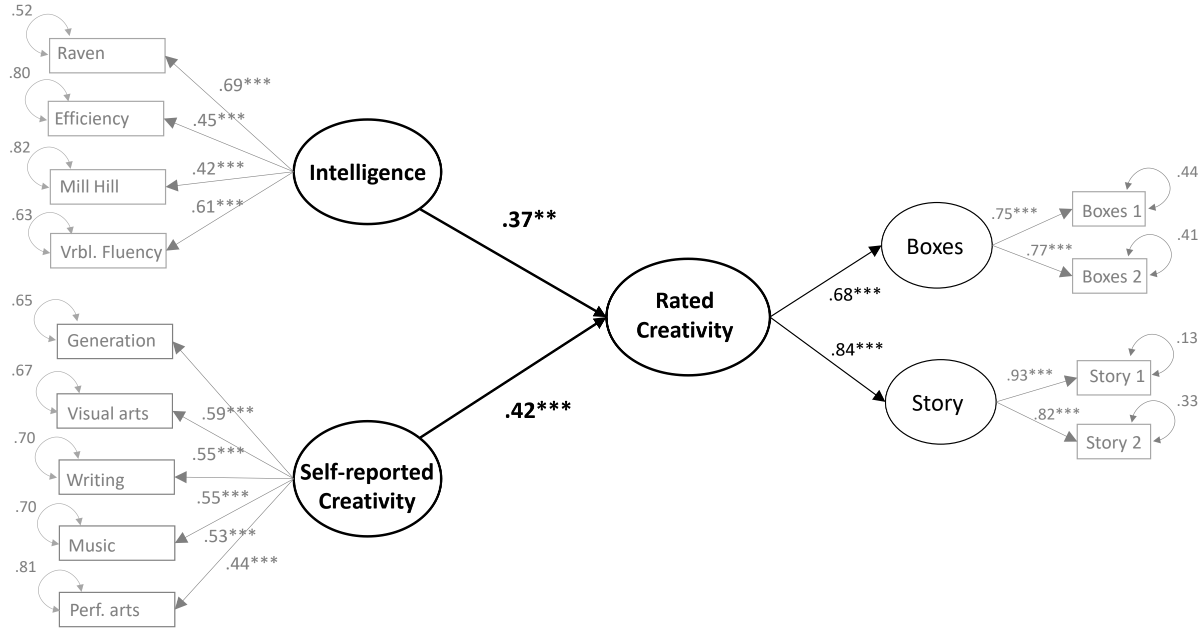
Note. E = expert rating; Ps = peer rating. This example specifically represents models for the Boxes task; ratings of Expert 3 are not available because, by design, Expert 3 did note rate the ideas in this task (Expert 2 did not rate the ideas of the Story task; Expert 1 did not rate the ideas of the Drawing task). See text for further details about missing data. For both types of model, the loading of Expert 1 was fixed to 1 for identification.

Figure 2. Integrative model of intelligence, creativity tasks, and the creativity questionnaire.



Note. *: p<.05; **: p<.01; ***: p<.001. Unilateral p-values. Details of model fit: χ^2 (107) = 117.5; p = .23; RMSEA = .026 [0; .051]; SRMR = .091; CFI = .97. For all factors, one loading was fixed to 1 for identification; for factor with two indicators only, residual variance of manifest variables were constraint equal to avoid local underidentification. All results reported on this figure are fully standardized. Additional parameters estimated but not represented on the figure (all non-significant): Self-reported creativity on Fluency; Self-reported creativity on Boxes; Drawing with Boxes; Drawing with Story; Fluency with Boxes; Fluency with Story; Fluency with Drawing; Self-reported creativity with Intelligence).

Figure 3. Alternative integrative model with second-order creativity factor.



Note. *: *p*<.05; **: *p*<.01; ***: *p*<.001. Unilateral *p*-values. Details of model fit: χ^2 (63) = 80.6; *p* = .07; RMSEA = .043 [0; .069]; SRMR = .087; CFI = .94. For all factors, one loading was fixed to 1 for identification; for factor with two indicators only, residual variance of manifest variables were constraint equal to avoid local underidentification. All results reported on this figure are fully standardized. Additional parameter estimated but not represented on the figure (non-significant): Self-reported creativity with Intelligence.