





DENIS DUMAS   
 SELCUK ACAR  
 KELLY BERTHIAUME  
 PETER ORGANISCIAK  
 DAVID EBY  
 KATALIN GRAJZEL   
 THEADORA VLAAMSTER  
 MICHELE NEWMAN  
 MELANIE CARRERA

## What Makes Children's Responses to Creativity Assessments Difficult to Judge Reliably?

### ABSTRACT

Open-ended verbal creativity assessments are commonly administered in psychological research and in educational practice to elementary-aged children. Children's responses are then typically rated by teams of judges who are trained to identify original ideas, hopefully with a degree of inter-rater agreement. Even in cases where the judges are reliable, some residual disagreement on the originality of the responses is inevitable. Here, we modeled the predictors of inter-rater disagreement in a large (i.e., 387 elementary school students and 10,449 individual item responses) dataset of children's creativity assessment responses. Our five trained judges rated the responses with a high degree of consistency reliability ( $\alpha = 0.844$ ), but we undertook this study to predict the residual disagreement. We used an adaptive LASSO model to predict 72% of the variance in our judges' residual disagreement and found that there were certain types of responses on which our judges tended to disagree more. The main effects in our model showed that responses that were less original, more elaborate, prompted by a Uses task, from younger children, or from male students, were all more difficult for the judges to rate reliably. Among the interaction effects, we found that our judges were also more likely to disagree on highly original responses from Gifted/Talented students, responses from Latinx students who were identified as English Language Learners, or responses from Asian students who took a lot of time on the task. Given that human judgments such as these are currently being used to train artificial intelligence systems to rate responses to creativity assessments, we believe understanding their nuances is important.

**Keywords:** creativity, creativity assessment, divergent thinking, elementary school, creativity judgments.

In order for an idea to be creative, it must exhibit originality: or a sense of novelty, unusualness, surprisingness, or cleverness within the context it was produced (Acar, Burnett, & Cabra, 2017; Runco & Jaeger, 2012; Simonton, 2012; Wilson, Guilford, & Christensen, 1953). Therefore, as creativity researchers, we commonly administer tasks and assessments to our participants that are designed to elicit the production of original ideas (e.g., divergent thinking tasks; Dumas & Dunbar, 2014; Forthmann, Paek, Dumas, Barbot, & Holling, 2020; Forthmann, Szardenings, & Holling, 2020; Said-Metwaly, Taylor, Camarda, & Barbot, 2022). Within the creativity research literature, a relatively wide variety of tasks have been devised for this purpose, which differ depending on the population in which they are intended to be used (e.g., children or adults; Orwig, Diez, Vannini, Beaty, & Sepulcre, 2021; Richard, Aubertin, Yang, & Kriellaars, 2020), the domain in which they are nested (e.g., arts or sciences; Dumas, Schmidt, & Alexander, 2016; Pürgstaller, 2021), and the format in which they are composed (e.g., verbal or drawing; Fink, Reim, Benedek, & Grabner, 2020), among many other differences (e.g., instructions; Acar, Runco, & Park, 2020). Despite all these sources of variation, one aspect that tends to unite creativity assessments is their open-ended nature (Dumas, Doherty, & Organisciak, 2020; Dumas, Organisciak, Maio, & Doherty, 2020).

In other areas of empirical work within the psychological and education sciences, it is common to employ close-ended and selected-response type measures such as multiple-choice knowledge tests (e.g., McMullen, Hannula-Sormunen, Lehtinen, & Siegler, 2022) or Likert-style self-report measures (e.g., Coleman, Dong, Dumas, Owen, & Kopta, 2022). Although creativity researchers have fruitfully utilized such selected-response assessments (e.g., the Remote Associates Test; Mednick, 1962), creativity research in general is characterized by the collection and interpretation of open-ended and ill-structured data wherein tasks and assessments are administered to participants, and participants respond to those stimuli by positing ideas in either a spoken, written, or drawing format (Barbot, Hass, & Reiter-Palmon, 2019). Human raters are often utilized to carefully read participants' creative responses and judge the degree to which they exhibit originality by designating a numeric code to each response that represents where it falls on an underlying scale of originality (Long & Wang, 2022). Indeed, even in recent research that has utilized automated scoring methods for originality based on text-mining models (e.g., Beaty & Johnson, 2021; Dumas, Doherty, & Organisciak, 2020; Dumas, Organisciak, et al., 2020; Sung, Cheng, Tseng, Chang, & Lin, 2022), human judges are often used as an important criterion against to ascertain the validity of the text-mining scores. In addition, as supervised machine learning methods designed to automatically recognize originality in participant responses begin to proliferate in the creativity literature (see Cropley & Marrone, 2022; Organisciak, Newman, Eby, Dumas, & Acar, 2022; Organisciak, Acar, Dumas, & Berthiaume, 2022 or Patterson, Barbot, Lloyd-Cox, & Beaty, 2022 for recent additions to the literature and also see Forthmann & Doebler, 2022 or Paulus & Renzulli, 1968 for a classic and often-overlooked approach), those models will require reliable and valid human judgments of originality in order to be trained: a situation that underscores the importance of human judges rather than undermines it.

In extant measurement theory surrounding psychological judgments and ratings (see Shavelson, Webb, & Rowley, 1989 for a classic presentation), each human rater is considered to provide important information related to the target construct. However, each individual rater is also expected to bring with them their own unique and idiosyncratic sources of measurement error as well. This error is understood to exist regardless of whether the judge is an expert in the domain from which the responses were collected (e.g., expert design engineers judging students' design projects; Lu & Luh, 2012), whether they are a systematically trained coder (e.g., researchers and graduate students in psychology labs; Kaufman, Lee, Baer, & Lee, 2007), or whether they are laypeople with no particular interest or knowledge in the area (e.g., workers on MTurk; Hass, Rivera, & Silvia, 2018). Given how individual judges are known to bring measurement error into a study, multiple raters are generally utilized in modern inquiry, and the degree to which they agree with each other on their ratings is used as evidence that the rating scheme has been applied reliably.

It is important to note that inter-rater reliability can be conceptualized in two ways, as either absolute agreement or consistency (Hallgren, 2012). In the case of absolute agreement, human judges are expected to precisely correspond on the level of the attribute that is expressed by a response, and any disagreement among the raters is taken as evidence that reliability is not perfect. On the other hand, consistency reliability does not require that raters exactly agree, but instead that their ratings covary perfectly; meaning that they agree on their ratings in a relative sense, but perhaps not in an absolute sense. These two conceptualizations of reliability are most appropriate in somewhat different measurement contexts. Whereas consistency reliability is sufficient for correlating test scores with other measures, absolute agreement is necessary when tests feature higher-stakes cut-scores, above or below which participants qualify for programs or services. In our literature review here, we use the terms 'agreement' and 'disagreement' more generally to refer to the reliability (or lack thereof) of any coding scheme. In the empirical part of the current study, we report both absolute agreement and consistency reliability indices, and we use response-level inter-rater variance as our operationalization of rater disagreement. Of course, whether reliability is conceptualized as absolute agreement or as consistency, in practice, the reliability of any coding scheme cannot be perfect because human judges naturally hold individualized perspectives on the target construct that introduce measurement error. But with careful training, raters can be taught to code (even for a complex construct such as originality) with satisfactory levels of reliability (Stemler & Kaufman, 2020).

In the field of creativity research, the process of selecting and training judges, enacting the coding scheme, and calculating the reliability of the ratings is most often seen as a means to an end: the main intention in the field is to study originality or creativity itself, and therefore, we aim for reliable codes in order to methodologically support our substantive work. However, a small but notable body of research has focused on the process of human rating *eo ipso*, empirically investigating and making recommendations on the types of tasks, participants, coders, and trainings that best support rater agreement and reliability (see

Wang & Long, 2022 for a recent example). In this paper, we build on that existing literature to develop a model to inform the field's understanding of rater variance when coding the originality of children's responses to creativity assessments. In the next section, we summarize extant work on human judges and the quality of their originality ratings within the creativity literature, before positing several additional potential sources of rater variance that have not yet been empirically examined in the literature.

#### SUMMARY OF PAST WORK ON JUDGMENT QUALITY IN CREATIVITY ASSESSMENT

Although rarely its own subject of inquiry, human judgments and their inter-rater agreement have been a critical aspect of creativity assessment since its inception (e.g., Torrance, 1969). Colloquially, the first and second author of this current paper have often taken part in conversations with other creativity researchers at conferences about topics such as what types of stimuli yield meaningful data for different populations of participants, how to train raters – most of whom are graduate students – to reliably identify originality, and whether responses from certain groups of students (e.g., age groups, English language learners, etc.) are more or less difficult for raters to judge reliably. So, from our perspective, the type of scholarly conversations that the current paper is meant to support are already occurring, but mostly outside of the peer-reviewed literature.

That said, a small but informative literature has emerged about the quality of human judgments in creativity research, most of which focuses either on aspects of the raters themselves (Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Gentile, & Baer, 2005) or on aspects of the assessments that support the collection of responses that are more straightforward to judge (Forthmann et al., 2017). In the former vein, the most commonly studied aspect of raters that has been systematically examined is rater expertise. For example, Long & Pang (2015) used a mixed methods approach to generate the process by which humans rated sixth grade students' responses to a creativity assessment. Using three groups of raters that differed in their vantage point on these students (i.e., researchers working in an educational psychology lab, teachers who worked with 6<sup>th</sup> graders, and undergraduate students with no prior experience in this area) these scholars were able to identify that the differences in the cognitive process of rating were driven by variation in raters' mental schema of a normative response to the task: essentially, the prior knowledge that the raters brought with them into the rating process was the key proximal influence on their ratings. What this meant was that the groups of raters differed in their mean level of rated originality (with teachers being the most generous and researchers being the least generous), but their levels of inter-rater agreement, calculated within groups, were relatively similar. This finding points to a general need to recruit raters that are relatively homogenous in terms of their expertise in order to achieve agreement: if raters differ too much on their prior knowledge in the domain, high degrees of inter-rater variance are likely to result.

Also in regard to studying characteristics of teacher-raters, Benedek et al. (2016) developed a measure of creativity evaluation skills in order to determine what psychological attributes of preservice teachers supported their judgments of student creativity. Using their measure, these scholars found that while on average preservice teachers underestimated the creativity of ideas, those who were more creative themselves, more open to experience, more intelligent, and who had a greater competence with language, were more accurate in their ratings. These findings are interesting because they suggest that potentially, educators who are creative themselves may be better able to recognize creativity in student ideas and, therefore, might be better able to support student creative thinking (see also Gurak-Ozdemir, Acar, Puccio, & Wright, 2019). In addition, while this study did not examine predictors of rater disagreement, it might suggest that those raters with more accurate judgments of creativity would also agree more and hence produce more reliable ratings.

Using a rating scheme that they termed 'dynamic evaluation', in which raters judged a creative product multiple times throughout the process of creation, Kozbelt and Serafin (2009) showed that raters with demonstrated expertise in the domain being judged (in this case visual artists judging drawings) differed in key ways from nonexperts in how they judged creative quality. Specifically, expert artists focused on the originality of a drawing while nonexperts were only able to focus on its realistic-ness. This finding makes sense in light of raters' prior knowledge: experts possessed the requisite prior knowledge of drawing in order to identify a work that seemed original to them, while nonexperts had only prior knowledge of the things the drawings were attempting to represent, and, therefore, had to focus their ratings on the degree of realistic representativeness. In another study focusing on experts rating the creativity of products, Lu and Luh (2012) showed that, when rating high school and undergraduate engineering students' design ideas, expert design engineers exhibited a notably lower inter-rater reliability than did nonexperts. What this finding implies is that, as their expertise developed in the domain of engineering design, engineers accumulated

knowledge that diverged from one another in meaningful ways such that, when they each compared the students' work to their own prior knowledge, their ratings exhibited more variance than nonexperts. Another, perhaps less generous possible implication is that experts may not be sufficiently motivated to rate students' work, and perhaps individuals who work with that population of students (e.g., teachers) or individuals who are motivated to produce high-fidelity research (e.g., graduate student researchers) may actually be more motivated judges in creativity research. Also related to the expertise of human judges, Dumas, Doherty, and Organisciak (2020); Dumas, Organisciak, et al. (2020), contended that professional stage or screen actors might be considered experts at verbal divergent thinking, given that those divergent thinking tasks bear a close resemblance to many exercises on which actors are trained. So, these scholars invited a professional actor to be the first judge of the responses in their dataset and had that actor help develop a training for their graduate students who would continue to rate the responses. In this work, it was found that the graduate students, when trained by a professional actor, were able to rate ideas with a high degree of agreement with that actor. This suggests that the effect of expertise on human judgments in the area of creativity assessment might be teachable if researchers learn what that expert views as original based on their expertise.

On the opposite side of the expertise spectrum, Hass et al. (2018) found that they were able to train very novice raters – who were crowdsourced via the website MTurk and paid a nominal fee – to reach reliability when they rated participants' individual responses to a divergent thinking task. In contrast, when these novices were asked to bundle multiple responses into a single overall rating (referred to as snapshot scoring; Silvia, Martin, & Nusbaum, 2009; or total ideational output; Runco & Mraz, 1992), they were not able to reach reliability. In similar work on human judges, but outside of the creativity literature, Organisciak (2015) found that raters hired on MTurk were better able to reach reliability when their training involved specific examples of responses that would validly receive specific codes, and when they rated individual responses one at a time. This finding was corroborated and explained in further detail by Forthmann et al.' (2017), who calculated a cognitive complexity coefficient for various kinds of rating tasks and showed that instances where raters were asked to bundle multiple ideas into a single originality rating required a much greater load of cognitive complexity. Likely because of this cognitive complexity, the raters exhibited significantly more disagreement in cases when they utilized snapshot scoring for larger and more diverse sets of ideas. Also interestingly, Forthmann et al. (2017) found that when divergent thinking tasks featured the explicit instruction to participants to 'be creative', the judges showed significantly more disagreement in rating the responses. This finding is highly psychometrically relevant because it suggests that possibly the originality of the responses themselves might make those responses easier or harder to rate with a high degree of agreement among the judges: an idea that we explore further in the current study.

When attempting to produce human-rated scores for originality that exhibit a high degree of inter-rater agreement, psychological attributes of the raters other than expertise are also relevant. For instance, Zhou, Wang, Song, and Wu (2017) conducted an in-depth sequence of studies on the role of individual judges' motivational orientations (i.e., prevention or promotion foci; cf. Rosenzweig & Miele, 2016) as well as attributes of the context they were situated in (i.e., whether the culture valued creativity or not), as well as the interactions among them, in influencing ratings of creativity. What they found was that individuals holding a promotion-oriented motivation focus perceived more novelty in the ideas they were judging. In addition, those individuals who were nested in organizational cultures that explicitly valued creativity also perceived more novelty. Although this work did not model inter-rater variance as an outcome, it did show the important role that context and individual differences may play in affecting the way raters judge the creativity of responses. In related and recent work, Ceh, Edelmann, Hofer, and Benedek (2021) examined how the personality attributes of human judges can influence their ratings using 166 novice judges who were current undergraduates. These researchers found that judges who were higher in openness to experience, divergent thinking (see also Guo et al., 2022), self-reported creative achievements, and who spent more time on the ratings were better able to discern high versus low creative responses. By using discernment as the outcome, these results are simultaneously relevant to both the judges' sensitivity to creativity, and the degree to which raters are expected to agree with one another. This means that if all the raters operated with high levels of discernment, presumably, inter-rater variance would be low and agreement would be very high. Ceh et al. (2021) also administered a self-reported Dark Triad scale to their novice judge participants but found no significant affects for those attributes on the ratings.

To summarize the extant literature, rater variance appears to be generally lowest, hence reliability could be expected to be highest, in cases where a reasonably high number of relatively homogenous raters are

utilized. In cases where lower numbers of raters are used, or when those raters differ from one another substantially in terms of their prior knowledge in the domain or their perceptions of creativity, large amounts of inter-rater variance would be hypothesized to be present. In addition, because the cognitive process of rating involves a relatively complex mapping between the response being rated and the raters' schema for a normative response, taking meaningful steps to limit the cognitive load of the rating task (i.e., rating only one response at a time, not rating them in bundles) is a good strategy for helping raters do their best work. However, it is also clear that much remains to be learned about what types of tasks, responses, and participants are best served by human raters, and how variance in ratings may or may not be a validity threat within educational contexts. For instance, it is not known whether some salient demographic attributes of participants (e.g., gender, race) could influence the reliability of human judgments, even in cases where the judges are totally blind to the participants being rated. In the current work, we hypothesize several additional sources of inter-rater variance that have not been fully examined in past research. These sources of inter-rater variance are summarized in the next section.

#### SOME FACTORS HYPOTHESIZED TO AFFECT JUDGMENT RELIABILITY

Here, we take special interest in unexplored sources of inter-rater variance for children's responses to creativity assessments, and we organize these potential sources of variance into three areas: (a) attributes of the assessment items, (b) attributes of children's responses, and (c) attributes of the children themselves.

##### Attributes of the items

Because of the varying contexts in which we collect data, creativity researchers must continually apply their own creativity in devising and developing appropriate assessments of creative thinking. For this reason, many different types of assessments exist in the field today, and it is largely unknown how differing stimuli may affect the inter-rater variance. Hass et al. (2018) were one team that did systematically examine differences in rater reliability across two common divergent thinking tasks, the Alternate Uses Task (AUT) and the Consequences Task finding that judgments were slightly more reliable for the AUT (see also Silvia, 2011; Silvia et al., 2008 for differences in inter-judge reliability between AUT and Consequences). Of course, given the wide variation in stimuli within the field of creativity, it would be untenable for any one study of rater agreement to include all possible assessments. In general, following Forthmann et al.' (2017) findings, that tasks which require a larger cognitive load for raters might produce lower inter-rater agreement, although it is not clear to us which divergent thinking tasks those may be, especially in cases where all responses regardless of task are not being bundled into a single originality rating but being rated one by one. In the current study, we include three verbal divergent thinking tasks that vary from very common in the literature to relatively uncommon: an AUT, an Instances tasks, and a Complete the sentence task. More details about these tasks are presented in the Method section.

##### Attributes of the responses

When participants respond to creativity assessments, their responses can differ in several ways. Likely the first that comes to mind is the actual originality of the responses, and it could be that lower or higher originality responses are harder for human judges to rate reliably. But typically the variance in the ratings and the level of the ratings are confounded together because the variance of the ratings is on the scale of the ratings themselves. So, the originality of a response could not typically be validly used to predict the variance in ratings of the originality of that response (Feng & Hancock, 2022). One modern method that could help with this issue is to utilize text-mining based originality scores as well as human judgments of originality. In this method the originality of the response, operationalized as the semantic distance between the prompt and the response (Dumas, Doherty, & Organisciak, 2020; Dumas, Organisciak, & Doherty, 2021), could be used as a predictor of the variance in human-judged originality. As far as we are aware, it is not known whether verbal divergent thinking responses that exhibit a greater semantic distance from the prompt are easier or harder for humans to rate reliably. On the one hand, responses that are more original in terms of semantic distance might be more open to human interpretation and, therefore, could display greater variance in their human-judged originality. On the other hand, responses with recognizably high levels of originality might be particularly straightforward for humans to identify, because they are trained to be sensitive to creative quality, and therefore, human raters might be likely to agree readily on such responses. Still, the originality of the response might predict rater variance in a way that interacts with other aspects of the

response, or attributes of the items or the participants themselves, such that highly original responses are straightforward for humans to identify for certain tasks or children, but not others.

Analogously, the elaboration of the responses could also be hypothesized to influence rater variance (Dumas, Organisciak, et al., 2020). It is known in the literature that, as responses get wordier, text-mining based methods often exhibit greater error in producing originality scores (Acar et al., 2023; Forthmann, Oyeade, Ojo, Günther, & Holling, 2019), but it is not known as far as we are aware whether human judges might display a similar issue. It could be that more elaborate responses create more variance in human ratings because more words increase the subjective nature of originality ratings by introducing variance not just about the originality of the idea itself but also in how that idea is verbally expressed (this would be essentially the same effect with humans as is known to occur with text-mining models). Raters would by necessity focus on the same word in single-word response whereas various raters may focus on different words of a multiword response, which may create disagreement among the raters. This hypothesized pattern would logically follow from Forthmann et al. (2017) findings concerning increases in rater disagreement when the cognitive complexity of responses is higher: note that their cognitive complexity measure was similar to the elaboration measure proposed by Dumas, Doherty, and Organisciak (2020) and Dumas, Organisciak, et al. (2020). In contrast, it could also be hypothesized that, as responses get wordier, the participants' intended meaning becomes less ambiguous to human judges, and therefore, human judges have an easier time identifying and rating originality in a reliable way, because the meaning of the response is clearer. Of course, different individuals have a greater or lesser ability to express their ideas clearly in words, and especially when the participants are children, the influence of elaboration on rater variance could interact with other attributes of the children themselves (e.g., age).

Besides originality and elaboration, a third attribute of creativity assessment responses that might affect how reliably judges are able to rate originality is the time that participants spent to formulate their response. Commonly for creativity assessment, meaningful individual variation exists in the amount of time that each participant spends to generate each response (Acar & Runco, 2019; Hass, 2015; Paek, Abdulla Alabbasi, Acar, & Runco, 2021). So, in our view, this makes the time each participant spends to formulate each of their responses a relevant variable in creativity assessment. Moreover, in cases where children are being assessed, the time spent on each response likely captures variance associated with many outside attributes of the children that would otherwise be considered random measurement error such as motivation, attention, reading speed, typing or writing speed, among others. Because response-level variance in time per response is likely influenced by this wide-reaching suite of individual and task-level attributes, it is difficult to hypothesize precisely how it might predict variance in human judgments of response originality and it could influence ratings either as a main effect or as an interaction with other attributes (e.g., elaboration, task type, special education status, etc.). For these reasons, we investigate the influence of time spent per response on rater variance in the current study.

#### Attributes of the children

In any situation where a measure is designed to yield psychological inferences across diverse subgroups of a population (e.g., elementary students from various gender or race groups), the way that the demographic attributes of the participants may affect the scoring of that measure is of key importance (Dumas & Grajzel, 2022). In psychometric areas where closed-ended items are common, differential-item-functioning or measurement invariance methods are utilized to determine whether demographic membership significantly moderates the way the item-response is weighted into the latent score (see Meredith, 1993 for a classic methodological guide). Within the creativity literature, some investigations of measurement invariance exist, most of which is focused on the functioning of Torrance Test of Creative Thinking across demographic groups (e.g., age and gender groups; Kim, Cramond, & Bandalos, 2006; Said-Metwaly, Van den Noortgate, & Barbot, 2021). However, these investigations have been situated to detect demographic group moderation at the link between item level quantities (i.e., the originality of a students' response to given item) and the overall latent score for the test. In contrast, what we are interested in here is whether participants' demographic characteristics may influence the degree of rater variance that is displayed for a given response, even when the raters are totally blind to the students being rated. For example, it does not seem unreasonable to suggest that grade level, gender, race, English language learning status, special education status, and gifted status, among others, might influence the ratings. These demographic characteristics might affect the way students describe their ideas, which could potentially make those responses more or less easily identifiable to raters as original. For example, for tasks administered in English, students who are fluent in



another language but are still learning English (referred to as English Language Learners; ELLs) might use language in an ambiguous or hard-to-rate way for English speaking raters, leading to more inter-rater disagreement. The same could potentially be said of younger students, or students with special education or gifted identification, which could lead the judges to disagree with one another resulting in larger rater variance in scoring. In the United States, race groups are a particularly sensitive grouping variable across which educational and psychological researchers work hard to make valid and fair inferences (see Dumas, Dong, & McNeish, 2022 for one psychometric perspective). Because race groupings could indicate cultural differences that may affect how students express their ideas and formulate their responses, therefore, potentially creating larger rater variance for certain race/ethnicity groups than others. Of course, recruiting reasonably diverse judges, at least in terms of gender and race, is also a clearly prudent method for limiting issues associated with these grouping variables, and it is a strategy we take in the current research.

#### GOAL OF CURRENT STUDY

In this work, we draw on a large dataset of elementary student responses to a verbal creativity assessment, collected in a school setting, to systematically examine whether and how attributes of the test items, attributes of the children's responses, and attributes of the children themselves may predict variance in judges' rating of originality. In addition, we examine interactions among these attributes to determine how those interactions predict disagreement. After identifying the key predictors of rater variance in our pool of responses, we also offer example responses from children to illustrate our findings. Our intention is to add to the ongoing scholarly conversation surrounding the refinement of creativity assessment, help in the training of human raters, and design tasks that yield responses that are more straightforward to judge when possible.

#### METHODS PARTICIPANTS

The responses analyzed in the current study were drawn from 387 elementary students in grades 3, 4, and 5. The students were evenly drawn from each of the grades with 137 (35.4%) in 3<sup>rd</sup> grade, 132 (34.1%) in 4<sup>th</sup> grade, and 118 (30.5%) in 5<sup>th</sup> grade. The average age of the participants was 9.36 years ( $SD = 0.97$ ). These students were recruited from five publicly funded elementary schools within a major metropolitan area in the central United States: their parents or guardians provided written consent for them to participate in this study, and they each received a toy as compensation for participation. These data were collected as part of a large and ongoing research project funded by the Institute of Education Sciences within the U.S. Department of Education with the goal of developing, validating, and norming modern assessments of creativity for elementary school students.

Demographic information for the participating children was provided to the research team by schools and was based on existing administrative data. Participants were recruited evenly based on sex (male  $N = 182$ ; 47.1%; female  $N = 205$ ; 52.9%). Also, as is typical within large metropolitan areas of the U.S., this sample of students was highly diverse in terms of race/ethnicity: 125 (32.5%) reported a Hispanic or Latinx identification; 114 (29.6%) reported Black/African American; 74 (19.2%) reported Asian/Pacific Islander; 60 (15.6%) were White/European American; 3 (1%) reported an American Indian/Indigenous identity; and 8 reported having multiple ethnic identities (2.1%). Three students did not report their ethnicity.

Additional key school-related grouping variables associated with these students were also reported to our research team by school administration: English Language Learning (ELL) status; Gifted/Talented (G/T) identification; and special education (SPED) status. Approximately one third of the sample ( $N = 255$ ; 34.1%) were classified as ELLs, meaning they spoke a first language other than English (most commonly Spanish). In addition, approximately one out of every seven students was identified G/T ( $N = 60$ ; 15.5%) and another one out every eight students ( $N = 1323$ ; 12.6%) was identified SPED. It should be noted that our sample did not include students with severe intellectual disability, for whom verbal creativity assessments are likely not informative. Instead, the SPED identifications in this sample encompassed students who teachers believed could reasonably respond to a written test in a group setting on their own.

#### ASSESSMENT

A 27-item verbal divergent thinking assessment, with three scales of nine items each, was administered to participants. For the purposes of higher-stakes scoring and norming of the measure, the first item of each scale was intended to be a practice item for the children (leaving eight items on each scale to be weighted

into the final score), but because the current study was focused on rater variance, we had our judges rate the practice items and included them here to maximize our sample of responses. Because each of the 387 students responded to 27 items, it gave us a total of 10,449 responses that were rated by our team of judges as a part of this study. Students were taught to refer to the assessment as the “Game of Surprises” and throughout the administration of the measure, they were directed to think of ideas that were ‘surprising’. Our team determined that ‘surprising’ was the most helpful way to describe the type of responses we wanted from students through piloting. We learned from students and teachers that terms like ‘creative’ or ‘original’ can be ambiguous to children, but ‘surprising’ carries a more specific connotation. In addition, surprise is a key attribute of originality and creativity (Simonton, 2012). For this reason, the explicit instructions that students were given for each item always utilized the idea of surprisingness to prompt children to provide creative ideas.

The three scales of the measure were composed of original items written by the research team, but they followed formats that generally exist within the creativity literature. The first scale was an AUT presented to the children as “How would you use it?” (an example item from that scale was “What is a surprising way to use a backpack?”). The second scale was an Instances task that was presented to the children as “What’s an example?” (an example item from that scale was “What is a surprising example of something that is loud?”). The third scale was more atypical in the current creativity literature called the “Complete the sentence” task. This task was written to appear similar to common complete the sentence-type tasks on reading comprehension tests, which children tend to be familiar with. But instead of completing the sentence with a correct response, in this case children were asked to complete the sentence with whatever they could think of that would be surprising. Although this item format is less common in the literature, it is conceptually akin to the Complete the Story task described by Barbot et al. (2019). An example item from this scale was “Complete this sentence in a surprising way: On the playground, the kids. . .” The three scales of the assessment were always administered in the following order: What’s an example?, How do you use it?, and then Complete the Sentence. However, to limit order effects, the items were randomized within each scale. Individual items were also not time-limited (although the time that each child took to respond to each item was logged by the computerized assessment system), but each of the three scales were time-limited. To avoid the fluency confound and use a larger number of prompts (Acar, 2023; Forthmann, Paek, et al., 2020; Forthmann, Szardenings, & Holling, 2020), we asked participants to give one response per prompt. Children had 12 minutes and 20 seconds to complete the “How would you use it?” task, 7:40 minutes to complete the “What’s an example?” task, and 5:40 minutes to complete the “Complete the sentence” task. These time-limits were based on pilot samples for this assessment and time-limits were set at 1.5 standard deviations above the average time it took pilot participants to complete a nontime-limited version of the assessment. Although the children were aware that they did not have the entire school day to respond to the study measures (there is an implicit understanding of schedule and timing in the school environment) they were not made aware of the individual task time limits in order to avoid time-limit related anxiety. Because of these time limits, there was some missing data in terms of children not responding to particular items (but these were certainly missing-at-random given the randomization of the items): 788 instances of missing responses existed in the current dataset; accounting for 7.5% of the data.

#### DATA COLLECTION PROCEDURES

The description of the project and recruitment materials were both electronically and on paper to parents and guardians via school coordinators. Parents and guardians provided written consent for their child’s participation in the study. Children’s assent to participate in our project was verbally given on the day of the activities. A post-doctoral researcher and a research assistant welcomed students and explained to them the general procedures and the types of activities that they would be asked to do. To limit students’ test anxiety and promote a creative environment, we strongly emphasized to students that they would be playing ‘games’ rather than taking a test. Our research team worked continually to limit distraction and maintain children’s focus, while also promoting a welcoming environment in which the children felt free to think creatively.

This study took place in a highly ecologically valid setting, with all the attributes of a school-based environment. All data were collected in spaces provided by schools such as multipurpose rooms, classrooms, and computer labs. The average group size of participants was 10.75 children ( $SD = 5.35$ ), and the largest group of children participating at a single time was 25. Each student used either a laptop or a desktop computer and a pair of headphones to complete the activities. Headphones were provided because all the



instructions were electronically read aloud to students in order to limit variance in the responses based on reading ability. In larger groups, study carrels were used to limit distractions for individual students. Research team members monitored students' progress and assured that students were using their devices only for the study activities. If students completed the activities before their group testing session was complete, they were instructed to sit quietly, use their device for a school-approved program, read, or draw. After completing all activities, students received a toy as compensation for participation. With the goal of limiting distraction, participants did not receive their toy until after all students completed the tasks in the session.

## RESPONSE CODING

### Precoding processing of spelling errors and blinding

Perhaps not surprisingly, the elementary students who participated in this study were quite physically capable of typing their responses to the assessment using their provided laptop; however, their spelling was sometimes incorrect. Our team held the viewpoint that children's scores on this measure should be affected as little as possible by variance in outside attributes such as spelling. For this reason, the third author of this paper read over all 10,449 responses in this dataset to streamline the spelling before the responses were scored. In this process, spelling was only changed if the intended word could relatively easily be deciphered, and grammatical or spelling flourishes that the students added intentionally (e.g., choosing to write a particular word all in capital letters for emphasis) were retained. After editing spelling, the third author also organized the responses based on the prompt and blinded the dataset so the judges would not be able to see any information about the child who supplied the responses.

### Who were the judges?

Five researcher assistants served as judges for the responses in the current study, with all five judges rating all of the 10,449 responses in the dataset. All raters were funded by the research grant that supported this project, and all five had been involved in the development of the assessment and the collection of data. Although some of the raters were more involved with the school-based data collection, and others were more involved with measure design or building the computerized assessment system, all had a well-developed sense of the purpose of this research endeavor, and all had been regularly attending lab meetings with the investigators. At the time of the coding, two of the five raters were doctoral students, two were masters students, and one was a closely involved and funded undergraduate research assistant. The coders ranged in age from 21 to 37 with the average age being 29.4. Four of the five raters identified as female, and one as trans nonbinary. One of the five raters reported a White ethnicity, one reported a Black/African American identity, one reported a Hispanic/Latinx identity, and two were multiethnic (one was African American and European American, and the other was European American and American Indian-Choctaw). Two of the five raters spoke first languages other than English (one being Spanish, and one being Hungarian). We report these data about our raters to demonstrate the diversity of perspectives that they brought to the study and give a sense of who formulated the ratings that are the focal point in the current study.

### How were the judges trained and what was their process?

The raters were trained by the first and second author. Our two doctoral student raters had considerable prior experience and knowledge for rating the originality of children's responses. Therefore, they received the training again, but also exercised a degree of leadership during the training session given that they had completed a separate rating project recently. In the training, the raters were taught to score responses on an originality scale ranging from 1 (totally unoriginal) to 5 (maximally original). They were specifically directed to conceptualize the ratings as arising from an underlying normal distribution such that the large majority of responses would fall in the middle of the scale at a 2, 3, or 4, and the extreme points of the scale (i.e., 5 and 1) would be relatively rarely used.

All five of the raters judged all 10,449 responses in the dataset. The raters were directed to judge the responses in the order of the prompts, so they judged all the responses for a given prompt, and then moved on to the responses given to the next prompt. They were also asked to peruse all the responses to each prompt before rating them. This was done so that the raters could initially familiarize themselves with the range of responses that could be expected from elementary school students and reduce rater drift over the course of the rating process. The raters coded each of the responses blind to the judgments of the other

raters. When they were finished, they sent their datasheet to the third author of this paper, who organized their ratings into single dataset and ascertained rater agreement.

#### Rater agreement and distribution of ratings

Although the current paper is focused on modeling rater variance, our intention when we trained the judges was to reduce this inter-rater variance across the responses as much as possible and have ratings that were generally reliable. For this reason, the current paper is focused on modeling and understanding the residual inter-rater variance that we were unable to eliminate via our training. Our five trained raters provided relatively highly reliable ratings ( $\alpha = 0.844$ ), which reasonably minimized the residual inter-rater variance in these data. Across all 10,449 responses, the average originality rating was 2.99 ( $SD = 0.59$ ). In order to ascertain whether our training was successful in conceptualizing an underlying normal distribution for the originality ratings, we fit a continuous normal distribution and performed an Anderson-Darling test, which is a highly sensitive test for normality (Anderson & Darling, 1954). Results of the Anderson-Darling test ( $A^2 = 45.31$ ;  $p < .001$ ) indicated that a continuous normal distribution with parameters  $\mu = 2.99$  and  $\sigma = 0.59$  fit the observed ratings across the 10,449 responses very well. See Figure 1 for a histogram of the response ratings, with superimposed normal distribution. These results strongly corroborate our decision to treat these codes as continuous and conceptualize reliability as consistency in the covariances among the raters. If the codes were treated as categorical, and therefore, exact agreement was required among them to indicate reliability, the reliability of the coding used here would have been much lower (exact agreement ICC = 0.224).

#### CALCULATING TEXT-MINING BASED ORIGINALITY AND ELABORATION

This study also included the calculation of text-mining based originality and elaboration scores for all of the responses. In order to accomplish this, we utilized the online freeware Open Creativity Scoring (OCS; Organisciak & Dumas, 2020; <https://openscoring.du.edu/>). OCS was validated in most previous work using samples of divergent thinking responses from adult participants, but they recently added functionality to specifically handle responses from elementary-aged children. In a recent preprint from the OCS research team (Organisciak, Acar, et al., 2022; Organisciak, Newman, et al., 2022), they detail a new corpus of child-facing text including 582 million words from children's books, 12.4 million words from U.S. Children's TV show subtitles, 9.2 million words from child-facing YouTube videos, and 27 million words from simple English Wikipedia. In this study, we utilized the OCS system and a Global Vectors for Word Representation (GloVe; Pennington, Socher, & Manning, 2014) text-mining model to estimate the semantic distance

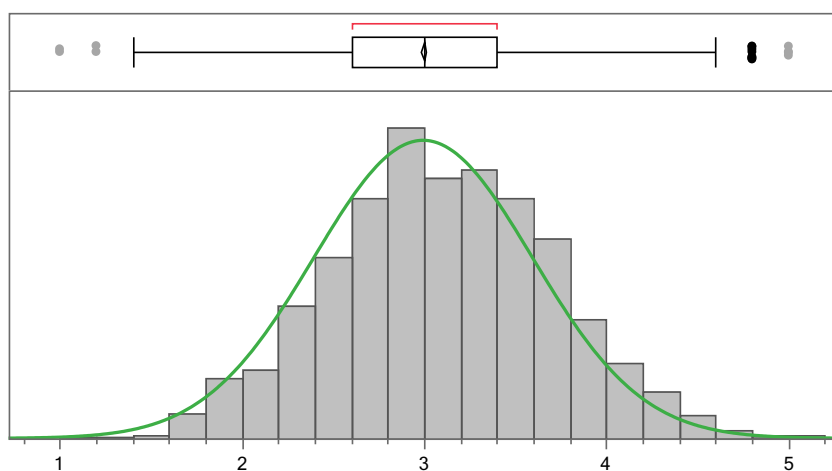


FIGURE 1. Histogram of the Originality Ratings. *Note.* Data displayed for 10,449 responses, averaged across the five judges. The originality ratings were highly normally distributed, as indicated by the superimposed continuous normal distribution. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

between each child's response (entered into the OCS after spelling corrections were applied) and the prompt following the procedures that are currently most recommended in the literature, including the deletion of very common words from the responses (i.e., stopword removal; Forthmann et al., 2019) and the weighting-up of very uncommon words in the response (i.e., term-weighting; Maio, Dumas, Organisciak, & Runco, 2020). Besides automatically scoring the originality of the responses based on that system, the OCS also provides automatically calculated term-weighted elaboration scores, which are word-counts for the responses that have additional weighting of each word depending on its rarity such that uncommon words count more toward the elaboration score and very common words count for less or not at all (see Forthmann, Szardenings, & Holling, 2020 for a detailed presentation of this method). In this study, we generated originality and elaboration scores via the OCS for every response in the dataset. We did not aggregate these response-level scores into an overall measure for each participant, but used the originality and elaboration at the response-level for the analysis here.

#### ANALYSIS OVERVIEW

After ascertaining the level of reliable agreement across our five raters and demonstrating the normality of the distribution of the ratings, we undertook the current analysis as a way to understand when and why our raters did not exactly agree. In other words: could we predict the variance across our five raters based on attributes of the assessment items, the responses that were coded, or the children who supplied the responses? In order to accomplish this goal, we organized our data into a long format (with 10,449 rows) such that the following analysis unfolded entirely at the level of the responses. We took the variance of the five judges' ratings and inserted them into the dataset as a new variable, such that we had one variance term for every response in the dataset. In cases where a variable is normally distributed, as was our average originality ratings here, the variance of that variable is mathematically known to follow a chi-squared distribution (Satterthwaite, 1941), which is a special case of the gamma distribution (Willink, 2003). For this reason, we modeled the variance of the ratings as a gamma distributed outcome. Moreover, the variance was modeled with a zero-inflated gamma (ZIG) distribution, because there were 588 cases where all five reviewers agreed exactly on their ratings, resulting in a variance of zero. Because the gamma distribution typically only holds positive values, the ZIG distribution allowed us to retain those 588 responses in the current analysis.

In modeling the variance among the ratings, our overall intention was to create a model that would be as informative as possible to the field in planning their own assessments and training their own raters. For this reason, we modeled all our key predictors as well as all two-way interactions as a way to best capture how the predictors operated in reality. In thinking about replicability and not overfitting our results to the current dataset – especially because it is very large and we are using a large number of predictors – we used an adaptive Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996) to select the predictors and interaction terms to retain in our model in a way that was most likely to replicate (i.e., be least effected by shrinkage) in future datasets (Zou, 2006). We fit these models in SAS JMP PRO version 16.0 using the adaptive double LASSO estimation procedure. The adaptive double LASSO penalizes the predictors in the model in order to avoid overfitting and support replicability and is designed to take an initial relatively large set of predictors and winnow them down to predictors that are nonspurious and expected to replicate (Urminsky, Hansen, & Chernozhukov, 2016). Double LASSO is conceptually akin to a 'leave-one-out' methodology for the predictors because it algorithmically enters the predictors and interactions into the model, removes them, and re-enters them based on the variance that the predictor explains in the outcome, as well as whether that predictor covaries substantially with the other predictors and may be an important covariate (Belloni, Chernozhukov, & Hansen, 2014).

## RESULTS

### LASSO PREDICTORS AND RESULTS

In Table 1, we list all the predictors that we initially placed in the model, as well as their levels if they were categorical. Asterisks indicate the reference level for categorical predictors (i.e., the category that would be coded as '0' using a numeric nominal coding). We fit the double adaptive ZIG LASSO model to all 10,449 rater variance terms in the dataset, attempting to predict why our raters did not exactly agree for each of the responses (even though they generally agreed well). In choosing what predictors to retain, we used the AIC as the validation method for the model, with the final model showing an AIC of  $-1,525.23$ . The ZIG LASSO model used here exhibited an  $R$ -square of 0.721, indicating that 72.1% of the variance in

TABLE 1. Predictors Initially Entered into the Adaptive LASSO Model

Conceptual Category	Predictor Name	Levels
Attributes of the Assessment	Assessment Scale	1. "How do you use it"* 2. "What's an example" 3. "Complete the Sentence"
Attributes of the Responses	OCS Originality	Continuous
	OCS Elaboration	Continuous
	Time on response	Continuous
Attributes of the Children	Gender	1. Male* 2. Female
	Race/Ethnicity	1. White/European American* 2. Black/African American 3. Hispanic/Latinx 4. Asian/Pacific Islander 5. American Indian/Indigenous 6. Multiethnic
	English Language Learner (ELL)	1. ELL 2. Not ELL*
	Gifted/Talented Identification (G/T)	1. G/T 2. Not G/T*
	Special Education Identification (SPED)	1. SPED 2. Not SPED*
	Age	Continuous

*Note.* All two-way interactions among these predictors were also entered into the model, \* indicates the reference level for categorical predictors.

rater disagreement was able to be predicted. As can be seen in Table 2, the ZIG LASSO model retained 17 significant predictors. Six of these 17 significant predictors were main effects indicating that responses showed significantly greater inter-rater disagreement if they were less original, from the Uses task, were generated by younger students, were more elaborate, or were from male students. The remaining 11 significant predictors were 2-way interactions among the terms in the model. For instance, despite the main effect of originality being in the negative direction, originality interacted with the task scale, Gender, Age, and G/T identification all in the positive direction, suggesting that when those attributes of the responses were combined with the response being highly original, our judges showed significantly more inter-rater disagreement.

In addition, although the main effect of Age was in the negative direction as well, it interacted with the task scale and with Gender in the positive direction, such that greater participant age predicted more inter-rater disagreement if the responses were from the Complete the Sentence task or from a Female student. Although there was no main effect of race/ethnicity, responses from Latinx students showed greater inter-rater disagreement if those students were also identified as ELL or SPED, and responses from Asian students showed greater inter-rater variance if those students also took greater than average time to formulate the response.

Because these 17 significant predictors of rater variance can be complex to interpret, especially because many of the categorical predictors have multiple levels, and the predictors interact together, we translate the model's key implications in the next section.

TABLE 2. Adaptive LASSO Results for Retained Predictors and Interactions

Predictor	LASSO Results			
	Coefficient Estimate	SE	Wald $\chi^2$	p-Value
Originality	-0.80	0.09	80.49	<.0001
Scale [Sentence]	-0.12	0.01	77.05	<.0001
Scale [Example]	-0.14	0.02	68.39	<.0001
Age	-0.05	0.01	46.90	<.0001
Elaboration	0.02	0.00	37.21	<.0001
Scale [Sentence]*Originality	0.66	0.12	32.24	<.0001
Gender*Originality	0.36	0.09	16.66	<.0001
Gender	-0.05	0.01	12.86	.0003
Scale [Example]*Originality	0.33	0.11	9.45	.0021
Scale [Sentence]*Age	0.03	0.01	9.01	.0027
Gender*Game [Example]	0.06	0.02	7.33	.0068
G/T*Originality	0.32	0.12	7.02	.0081
Ethnicity [Latinx]*ELL	0.04	0.02	6.42	.0113
Ethnicity [Latinx]*SPED	0.06	0.02	5.96	.0146
Gender*Age	0.02	0.01	5.33	.021
Ethnicity [Asian]*Time	0.01	<0.01	5.06	.0245
Age*Game [Example]	0.02	0.01	4.05	.0442

*Note.* Predictors are presented in the order of the strength of their effect. All other predictors and interactions were dropped by the adaptive LASSO model. Brackets are used to indicate which levels of categorical predictors were significant in cases where the predictor had more than two levels.

## DISCUSSION: WHICH RESPONSES WERE MORE DIFFICULT TO RATE RELIABLY?

In this section, we translate the findings from the LASSO model and then provide touchstone examples from our dataset to illustrate the findings, which allow us to offer some conjecture and discussion about why these types of responses might have produced larger inter-rater disagreement.

### LESS ORIGINAL RESPONSES

The strongest predictor in our model of inter-rater variance was the originality of the response, operationalized as semantic distance and quantified by the OCS. Interestingly, this predictor was in the negative direction, indicating that less original responses exhibited greater inter-rater variance than more original responses, after controlling for all the other predictors in the model. To us, this makes theoretical sense because our raters were researchers working in a creativity lab, and as such were likely interested particularly in the most highly original responses. If the raters as a group were most sensitive to the most original responses, their discernment of those very original responses and agreement on them would naturally be higher, while the lower originality responses would have less agreement. This finding is potentially related to Forthmann et al.' (2017) finding that 'be creative' instructions in divergent thinking tasks was associated with greater rater disagreement: possibly the raters themselves were also influenced by the task context and instructions and, therefore, were more capable of reliably identifying the responses that were more original.

As a case in point, one child responded to the Examples task prompt 'loud' with the response 'screaming'. The OCS rated the originality of this response quite low (0.281) because the word 'screaming' is semantically close to the word 'loud', but the human judges were split across multiple codes at the low end of the scale ranging from 1 to 3. It may also have been that our judges, all of whom had an interest in education and children's cognition, felt that screaming might be a more out-of-the-box response than the OCS thought it was, because screaming would be socially inappropriate in the school setting (where the child was when they responded to the assessment), and as such the child was thinking outside of their immediate context. The take-away from this predictive term in the model is that it appears that human judges are more readily in agreement on the highest originality responses, and in less agreement on lower originality responses.

### USES TASK RESPONSES

Going into the current study, we were unsure how aspects of the assessment items would influence the disagreement among the raters. In our model, we found that items written in the ‘Complete the sentence’ and the ‘What’s an example’ format both exhibited significantly less inter-rater variance than did ‘Uses’ task items after controlling for all other predictors in the model. We might hypothesize that the ‘Uses’ task responses intrinsically invite the raters to think of the feasibility or utility of the responses (even though they were specifically directed not to), and therefore, the feasibility issue could perhaps be causing the raters to disagree somewhat more on the ‘Uses’ task than the other tasks. One of the responses in our dataset that exhibited the highest degree of inter-rater variance (with judgments ranging across the entire scale from a 1 to a 5) was in response to the ‘Uses’ task prompt ‘toothbrush’: ‘When it is alive it can brush your own teeth’. In our view, we see why this response was particularly difficult for our judges to agree on, because there really are two parts of it. First the toothbrush comes alive, and then it is brushing teeth. So, if our raters focused on the core use of the object itself, this idea would be quintessentially unoriginal (i.e., the toothbrush is being used to brush teeth), but if they focus on the fact that the toothbrush is alive and doing the brushing on its own, then perhaps the originality is high. According to our LASSO model, the ‘Uses’ task responses appear to have challenged our judges with significantly more opportunities for disagreement such as this one.

### RESPONSES FROM YOUNGER STUDENTS

Another significant main effect from our LASSO model was for age, which predicted inter-rater disagreement in the negative direction, indicating that responses generated by younger children exhibited significantly more inter-rater variance, while older children’s responses were more straightforward to rate reliably. In our sample, the youngest children were 8 years old, and the oldest were 12. One response from an 8-year-old participant to the Uses task prompt ‘Hat’ that our raters disagreed on to a large extent was: ‘you cut off the shade part and it will look silly’. Our raters were in disagreement about the originality of this response, with judgments that ranged from a 1 to a 4. In the perspective of the rater who rated it highly, this child was altering the hat in a meaningful way, but on the other hand, in the perspective of the rater who rated it lower, even after altering the hat, this child is still envisioning wearing it simply as a hat. In our view, this is an example of how young children’s responses can be difficult to score for adult judges: it is possible that this child had never seen a hat with its brim cut off and genuinely thought that a hat without a brim would be very funny and surprising. Some of our raters agreed with this child that a hat without a brim would be surprising and original, but some of our raters did not.

### MORE ELABORATE RESPONSES

The OCS weighted elaboration term predicted rater variance in the positive direction indicating that more elaborate responses exhibited higher inter-rater variance. This makes sense, given that highly wordy responses require human judges to attend to multiple aspects of the response, therefore, creating disagreement among the raters as to how original the response was. A more elaborate response also likely produces a higher cognitive load for the judges, making it difficult to hold all aspects of the response simultaneously in memory in order to give it a rating. This finding dovetails entirely with that of Forthmann et al. (2017), who similarly found that the complexity of responses predicted rater disagreement. One example of a highly elaborate response in our dataset that also exhibited a high degree of inter-rater variance was the following response, which was given to the Uses task prompt ‘lightbulb’:

Use the inner part of it and attach it to something like a lunch pail or something and find a way to power it. Then the light will produce light energy and heat to keep your food warm and make it easier to find things. then you can use the glass to cover the lower part to cover the food so you don’t burn your hand getting it.

As can be seen, this response is quite wordy, and even at times difficult to understand what the child visualized doing with the lightbulb. Our raters ranged from a 3 to a 5 on this task, exhibiting a higher level of disagreement. This disagreement, we suspect, was likely caused at least in part by how wordy the response was: it was difficult for our judges to attend specifically to the core of the idea, and perhaps they even disagreed about what the core idea was. For this reason, highly elaborate responses led to greater inter-rater variance in our dataset.



## GIFTED/TALENTED CHILDREN'S HIGHLY ORIGINAL RESPONSES

Above, we noted that, controlling for all the other predictors in the model, *less* original responses were harder for our judges to agree on. However, our model also showed that highly original responses exhibited more inter-rater variance if those responses were also written by children with a G/T identification. One example of a response that fit this description, and challenged our raters to agree (with our judges giving it ratings ranging from 2 to 4), was in response to the Complete the sentence prompt 'At a sleepover, we...' One G/T student replied: 'Studied'. The OCS system thought that this response was highly original, because the word 'studied' is quite semantically distant from 'sleepover' in our specialized corpus of child-facing text. This is because, sleepovers are most often associated with fun and games, rather than studying, in the children's books and TV shows used to develop the corpus. In line with the OCS and two of the five raters, this child may have been cleverly supplying a surprising response that, in a child's world, runs deeply counter to expectation by studying when it is time to have fun. However, on the other hand (as was thought by three of our five raters) there really is no reason why children could not study at a sleepover, and it may even be somewhat common. For this reason, perhaps this response is not highly original. As can be seen, this G/T student found a way to challenge our human judges and produced a higher level of inter-rater variance among them.

## RESPONSES FROM LATINX CHILDREN WHO WERE EITHER ELL OR SPED

Another type of response that challenged our team of judges to agree on originality ratings were responses written by Latinx children who were either identified as ELLs or SPED. We hypothesize that this was likely because these children were responding to the assessment with somewhat less English vocabulary available to them, and as such tended to express themselves in ways that were slightly ambiguous to our judges. For instance, one Latinx ELL child responded to the Example task prompt 'Smelly' with the response 'A gorilla barfing'. This response exhibited a high degree of inter-rater variance for its originality, with our judges giving it ratings that ranged from 3 to 5. From the perspective of the judge who gave it a 5, this response is quite surprising and original, while other judges saw it as more obvious: especially because the responses 'barf' or 'puke' were common responses to the prompt 'Smelly'. Does the fact that it is a gorilla barfing and not a human make it a fundamentally different and more original response? Our raters disagreed on this point.

## RESPONSES FROM MALE STUDENTS, OR FROM FEMALE STUDENTS THAT WERE HIGHLY ORIGINAL

In our model, the Gender predictor was coded with Males as the reference group, so the categorical predictor can be read as an indicator of being Female. The Gender predictor had a negative main effect, indicating that on average Male children's responses exhibited a greater degree of inter-rater variance. However, Gender also displayed a (stronger) significant interaction with OCS Originality such that Female students' responses showed high degrees of rater disagreement in cases where those responses were also highly original. This particular finding, although robust in the LASSO model presented above, is somewhat difficult to interpret psychologically. One suggestion we would offer for the generally higher level of rater variance for responses given by male participants is that we know from prior research (e.g., Dumas & Strickland, 2018) that male students are more likely to posit violent ideas on the Uses task. Although we did not code specifically for violent or malevolent ideas in this study, if those ideas are harder to code reliably, that could potentially help explain the current finding related to Maleness and inter-rater variance. For example, one Male student responded to the Complete the sentence prompt "At the sleepover, we..." with the response, *eat someone outside in the night*. Our judges rated this response with a high degree of variance from one another, ranging from a 2 to a 4. On the one hand, eating someone is a highly unusual thing for a child to imagine themselves doing at a sleepover, but on the other – given the high degree of violence available in TV shows and movies today – it is also easy to picture them seeing this happen on screen perhaps in the context of a zombie story. For this reason, our judges were somewhat divided on its originality.

Our judges were only likely to display a high degree of inter-rater variance for female students' responses if those responses were also rated as very highly original on the OCS. In discussing the data, it appears that female students who were also highly original may have been more likely to imagine fantastical or magical things in their responses, which may have sometimes led to disagreement among the raters. For instance, in response to the Complete the sentence prompt, "On the school bus, I saw..." one female student wrote *everyone was fairies*. This prompt was rated highly by the OCS (0.916) because fairies and school buses are

relatively highly semantically far apart in the OCS children's text corpus. However, our human judges did not all agree on this, giving this response scores that ranged from a 3 to a 5. In our view, it appears that our raters, at least in this case, were in a disagreement about how surprising or unusual it would be to see fairies on a school bus. Is it maximally original because it is magical? On this issue, our raters were somewhat divided.

#### RESPONSES FROM ASIAN STUDENTS WHO TOOK MORE TIME

The final set of responses that exhibited significantly greater amounts of inter-rater disagreement in their originality ratings were responses from Asian students that took more than average time on that particular assessment item, but not Asian students in general. It is important to note that we imposed time limits on each scale of the assessment here, but within that scale-level time limit, children were free to vary on the amount of time they dedicated to each item. It is again somewhat difficult to conjecture as to why this effect occurred, but to illustrate it, one Asian child responded to the Example task prompt 'Red' with the response *a firetruck made from snow*. It took this child 2 minutes and 48 seconds to generate this response, which was long in comparison to the overall sample. From our perspective, it looks as if this child may have initially thought of a very typical thing that was red (a firetruck), but then perhaps remembered the specific instructions to generate a surprising response and thought for some time about how to augment their answer to make it surprising, eventually deciding that a firetruck made of snow was sufficiently original. Our judges were somewhat divided on whether or not a firetruck made of snow was a surprising example of something that is red, giving ratings for this response that ranged from 3 to 5. Perhaps in our dataset this strategy of augmenting or changing a common idea to make it more surprising was disproportionately used by Asian students who took a long time to respond, and for that reason their responses exhibited greater inter-rater variance for our judges. Of course, only future systematic research could determine if this strategic issue was the cause of the inter-rater disagreement we saw here.

#### CAVEAT: WHAT OUR RATERS DID AND DID NOT KNOW WHEN RATING

In interpreting each of the key findings of the current study, it is important to note the distinction between what our raters could directly know or perceive from the responses, and what they could not. More specifically, the raters only had access to the responses themselves, so proximal attributes of the responses such as to what task (i.e., Uses, Instance, Sentences) the students were responding, and the originality or elaboration of the responses, were known to them. In contrast, our raters did not know the attributes of the students (e.g., gender, race/ethnicity, age) or other attributes of their responding pattern (e.g., time spent on response) while they rated. For this reason, the findings here that are related to child attributes should not be interpreted as direct and proximal causes of rater disagreement. Rather, our key findings suggest that, in some yet-to-be-understood systematic way, students with differing attributes (e.g., G/T, SPED, or ELL status; race/ethnicity; gender) respond in different ways to DT prompts. The content of those prompts would, therefore, theoretically be the more proximal cause of rater disagreement. For instance, preliminary evidence from the responses we quoted above might suggest that our raters disagreed more on malevolent, magical, or oxymoronic responses. However, only a more detailed coding scheme would allow hypotheses such as this to be tested directly.

#### FUTURE DIRECTIONS & CONCLUSION

This study has been a relatively large scale and in-depth investigation of the predictors of inter-rater variance in originality ratings of children's responses to verbal divergent thinking tasks. As the Discussion section above has shown, we uncovered a variety of key findings that are expected to be relevant to the field as researchers write measures and train judges to rate participants' responses. Still, there is much to understand about how raters perceive the originality of responses, and in what cases the meaning that the participant intended is obscured or difficult to judge. Many future directions could be posited given the findings of this work, with one of particular interest to us being the potential use of children themselves as the judges of originality. The OCS's children's corpus is designed to score children's responses to creativity assessments the way a child would: it is built to have an understanding of language based on children's books and other sources of child-facing language. So, it may be that, following the same logic, children would be even better raters of their peers' responses to creativity assessments, although of course that research process would come with many logistical difficulties.

Another key future direction we can see in this line of inquiry is to adopt an open-science paradigm wherein we could gather additional datasets from creativity laboratories around the world, combine them, and then conduct a similar analysis with an even greater number and wider range of predictors. For instance, by combining datasets we might capture a wider range of participant ages (potentially through adulthood) and more types of items – beyond the three used here – such as items in a drawing or figural mode. We might be able to determine how the language used in measure administration and participant response may affect rater disagreement or other aspects of the measure administration that might vary widely across available datasets such as a wider variety of time limits, high- or low-stakes assessment settings, as well as different coding and scoring methods. Perhaps most salient in regard to the existing literature on creativity judgment is the important issue of rater characteristics (Benedek et al., 2016; Long & Pang, 2015; Zhou et al., 2017). In the future, we could use a similar methodology, like the one used in the present work, to test the effects of rater characteristics in a large sample of datasets from around the world. Within the creativity literature, some scholars are finding success in sharing datasets across laboratories and analyzing those datasets to make meaningful findings (e.g., Organisciak, Acar, et al., 2022; Organisciak, Newman, et al., 2022) – a process that may be fruitful in this area as well. Assembling a very large dataset using an open-science paradigm would better represent the scope and complexity of the field and the coefficients in that study would be substantially more likely to be replicated in future work.

For our part, we understand that individuals' subjective perceptions of originality and creativity will differ somewhat, even if they are well-trained researchers who agree with each other to a large extent. However, we do not necessarily see this subjectivity as a problem: instead, it may be an opportunity to study the ways that ideas are translated (or lost in translation) between the mind of the creator and the mind of the perceiver. The way that others perceive an idea is relevant to its creative quality (Simonton, 2000), and in a laboratory setting, our trained judges form a proxy for those general perceptions of others. For this reason, we look forward to more research on judgment and rating issues in creativity in the future.

### CONFLICT OF INTEREST

The authors declare they have no conflicts of interest.

### ETHICAL APPROVAL

This work follows all ethical guidelines of the American Psychological Association and was approved by the Institutional Review Board at the University of North Texas.

### DATA AVAILABILITY STATEMENT

The data used in this article are available from the authors upon request.

### REFERENCES

- Acar, S. (2023). Does the task structure impact the fluency confound in divergent thinking? An investigation with TTCT-figural. *Creativity Research Journal*, 35, 1–14; doi: [10.1080/10400419.2022.2044656](https://doi.org/10.1080/10400419.2022.2044656).
- Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C.T., & Organisciak, P. (2023). Applying automated originality scoring to the verbal form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly*, 67, 3–17; doi: [10.1177/00169862211061874](https://doi.org/10.1177/00169862211061874).
- Acar, S., Burnett, C., & Cabra, J.F. (2017). Ingredients of creativity: Originality and more. *Creativity Research Journal*, 29, 133–144; doi: [10.1080/10400419.2017.1302776](https://doi.org/10.1080/10400419.2017.1302776).
- Acar, S., & Runco, M.A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts*, 13, 153–158; doi: [10.1037/aca0000231](https://doi.org/10.1037/aca0000231).
- Acar, S., Runco, M.A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, 14, 39–49; doi: [10.1037/aca0000256](https://doi.org/10.1037/aca0000256).
- Anderson, T.W., & Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, 765–769; doi: [10.1080/01621459.1954.10501232](https://doi.org/10.1080/01621459.1954.10501232).
- Barbot, B., Hass, R.W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (Re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, 13, 233–240; doi: [10.1037/aca0000233](https://doi.org/10.1037/aca0000233).
- Beaty, R.E., & Johnson, D.R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, 53, 757–780; doi: [10.3758/s13428-020-01453-w](https://doi.org/10.3758/s13428-020-01453-w).
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81, 608–650; doi: [10.1093/restud/rdt044](https://doi.org/10.1093/restud/rdt044).

## Rater Disagreement in Creativity Assessment

- Benedek, M., Nordtvedt, N., Jauk, E., Koschmieder, C., Pretsch, J., Krammer, G., & Neubauer, A.C. (2016). Assessment of creativity evaluation skills: A psychometric investigation in prospective teachers. *Thinking Skills and Creativity*, 21, 75–84; doi: [10.1016/j.tsc.2016.05.007](https://doi.org/10.1016/j.tsc.2016.05.007).
- Ceh, S.M., Edelman, C., Hofer, G., & Benedek, M. (2021). Assessing raters: What factors predict discernment in novice creativity raters? *The Journal of Creative Behavior*, 56, 41–54. doi: [10.1002/jobc.515](https://doi.org/10.1002/jobc.515).
- Coleman, J.J., Dong, Y., Dumas, D., Owen, J., & Kopta, M. (2022). Longitudinal measurement invariance of the Behavioral Health Measure in a clinical sample. *Journal of Counseling Psychology*, 69, 100–110; doi: [10.1037/cou0000524](https://doi.org/10.1037/cou0000524).
- Cropley, D.H., & Marrone, R.L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*; doi: [10.1037/aca0000510](https://doi.org/10.1037/aca0000510). Online ahead of print.
- Dumas, D., Doherty, M., & Organisciak, P. (2020). The psychology of professional and student actors: Creativity, personality, and motivation. *PLoS One*, 15; doi: [10.1371/journal.pone.0240728](https://doi.org/10.1371/journal.pone.0240728).
- Dumas, D., Dong, Y., & McNeish, D. (2022). How fair is my test? A ratio statistic to help represent consequential validity. *European Journal of Psychological Assessment*. doi:[10.1027/1015-5759/a000724](https://doi.org/10.1027/1015-5759/a000724). Online ahead of print.
- Dumas, D., & Dunbar, K.N. (2014). Understanding Fluency and Originality: A latent variable perspective. *Thinking Skills and Creativity*, 14, 56–67; doi: [10.1016/j.tsc.2014.09.003](https://doi.org/10.1016/j.tsc.2014.09.003).
- Dumas, D., & Grajzel, K. (2022). Measuring up: Aligning creativity assessment with the Standards. In M. Runco & S. Acar (Eds.), *Handbook of creativity assessment*. Cheltenham, UK: Edward Elgar Publishing.
- Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15, 645–663; doi: [10.1037/aca0000319](https://doi.org/10.1037/aca0000319).
- Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2020). Four text-mining methods for measuring elaboration. *The Journal of Creative Behavior*, 55, 517–531. doi: [10.1002/jobc.471](https://doi.org/10.1002/jobc.471).
- Dumas, D., Schmidt, L.C., & Alexander, P.A. (2016). Predicting creative problem solving in engineering design. *Thinking Skills and Creativity*, 21, 50–66; doi: [10.1016/j.tsc.2016.05.002](https://doi.org/10.1016/j.tsc.2016.05.002).
- Dumas, D.G., & Strickland, A.L. (2018). From book to bludgeon: A closer look at unsolicited malevolent responses on the alternate uses task. *Creativity Research Journal*, 30, 439–450.
- Feng, Y., & Hancock, G.R. (2022). A structural equation modeling approach for modeling variability as a latent variable. *Psychological Methods*; doi: [10.1037/met0000477](https://doi.org/10.1037/met0000477). Online ahead of print.
- Fink, A., Reim, T., Benedek, M., & Grabner, R.H. (2020). The effects of a verbal and a figural creativity training on different facets of creative potential. *The Journal of Creative Behavior*, 54, 676–685; doi: [10.1002/jobc.402](https://doi.org/10.1002/jobc.402).
- Forthmann, B., & Doebler, P. (2022). Fifty years later and still working: Rediscovering Paulus et al's (1970) automated scoring of divergent thinking tests. *Psychology of Aesthetics, Creativity, and the Arts*. Online ahead of print.
- 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, 129–139; doi: [10.1016/j.tsc.2016.12.005](https://doi.org/10.1016/j.tsc.2016.12.005).
- Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, 53, 559–575; doi: [10.1002/jobc.240](https://doi.org/10.1002/jobc.240).
- Forthmann, B., Paek, S.H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*, 90, 683–699; doi: [10.1111/bjep.12325](https://doi.org/10.1111/bjep.12325).
- Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts*, 14, 94–112; doi: [10.1037/aca0000196](https://doi.org/10.1037/aca0000196).
- Guo, Y., Lin, S., Acar, S., Jin, S., Xu, X., Feng, Y., & Zeng, Y. (2022). Divergent thinking and evaluative skill: A meta-analysis. *The Journal of Creative Behavior*, 56, 432–448. doi: [10.1002/jobc.539](https://doi.org/10.1002/jobc.539).
- Gurak-Ozdemir, S., Acar, S., Puccio, G., & Wright, C. (2019). Why do teachers connect better with some students than others? Exploring the influence of teachers' creative-thinking preferences. *Gifted and Talented International*, 34, 102–115; doi: [10.1080/15332276.2019.1684221](https://doi.org/10.1080/15332276.2019.1684221).
- Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods For Psychology*, 8, 23.
- Hass, R.W. (2015). Feasibility of online divergent thinking assessment. *Computers in Human Behavior*, 46, 85–93; doi: [10.1016/j.chb.2014.12.056](https://doi.org/10.1016/j.chb.2014.12.056).
- Hass, R.W., Rivera, M., & Silvia, P.J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology*, 9; doi: [10.3389/fpsyg.2018.01343](https://doi.org/10.3389/fpsyg.2018.01343). Online ahead of print.
- 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, 171–178; doi: [10.1080/10400410802059929](https://doi.org/10.1080/10400410802059929).
- Kaufman, J.C., Gentile, C.A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, 49, 260–265; doi: [10.1177/001698620504900307](https://doi.org/10.1177/001698620504900307).
- 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, 96–106; doi: [10.1016/j.tsc.2007.04.002](https://doi.org/10.1016/j.tsc.2007.04.002).
- Kim, K.H., Cramond, B., & Bandalos, D.L. (2006). The latent structure and measurement invariance of scores on the Torrance tests of creative thinking-figural. *Educational and Psychological Measurement*, 66, 459–477; doi: [10.1177/0013164405282456](https://doi.org/10.1177/0013164405282456).

- Kozbelt, A., & Serafin, J. (2009). Dynamic evaluation of high- and low-creativity drawings by artist and nonartist raters. *Creativity Research Journal*, 21, 349–360; doi: [10.1080/10400410903297634](https://doi.org/10.1080/10400410903297634).
- Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, 15, 13–25. doi:[10.1016/j.tsc.2014.10.004](https://doi.org/10.1016/j.tsc.2014.10.004)
- Long, H., & Wang, J. (2022). Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational Psychology Review*, 34, 1399–1443. doi: [10.1007/s10648-022-09679-0](https://doi.org/10.1007/s10648-022-09679-0).
- Lu, C.C., & Luh, D.B. (2012). A comparison of assessment methods and raters in product creativity. *Creativity Research Journal*, 24, 331–337; doi: [10.1080/10400419.2012.730327](https://doi.org/10.1080/10400419.2012.730327).
- Maio, S., Dumas, D., Organisciak, P., & Runco, M. (2020). Is the reliability of objective originality scores confounded by elaboration? *Creativity Research Journal*, 32, 201–205; doi: [10.1080/10400419.2020.1818492](https://doi.org/10.1080/10400419.2020.1818492).
- McMullen, J., Hannula-Sormunen, M.M., Lehtinen, E., & Siegler, R.S. (2022). Predicting adaptive expertise with rational number arithmetic. *British Journal of Educational Psychology*, 92, 688–706. doi: [10.1111/bjep.12471](https://doi.org/10.1111/bjep.12471).
- Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69, 220–232; doi: [10.1037/h0048850](https://doi.org/10.1037/h0048850).
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543; doi: [10.1007/BF02294825](https://doi.org/10.1007/BF02294825).
- Organisciak, P. (2015). Design problems in crowdsourcing: Improving the quality of crowd-based data collection (Order No. 10151863). Available from ProQuest Dissertations & Theses Global. (1816981858). <https://www.proquest.com/dissertations-theses/design-problems-crowdsourcing-improving-quality/docview/1816981858/se-2>.
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2022). Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. doi: [10.13140/RG.2.2.32393.31840](https://doi.org/10.13140/RG.2.2.32393.31840).
- Organisciak, P., & Dumas, D. (2020). *Open creativity scoring [Computer software]*. Denver, CO: University of Denver.
- Organisciak, P., Newman, M., Eby, D., Dumas, D., & Acar, S. (2022). How do the kids speak? Modeling child-directed language for educational use. *Information and Learning Sciences*, 124, 25–47.
- Orwig, W., Diez, I., Vannini, P., Beaty, R., & Sepulcre, J. (2021). Creative connections: Computational semantic distance captures individual creativity and resting-state functional connectivity. *Journal of Cognitive Neuroscience*, 33, 499–509; doi: [10.1162/jocn\\_a\\_01658](https://doi.org/10.1162/jocn_a_01658).
- Paek, S.H., Abdulla Alabbasi, A.M., Acar, S., & Runco, M.A. (2021). Is more time better for divergent thinking? A meta-analysis of the time-on-task effect on divergent thinking. *Thinking Skills and Creativity*, 41, 100894; doi: [10.1016/j.tsc.2021.100894](https://doi.org/10.1016/j.tsc.2021.100894).
- Patterson, J., Barbot, B., Lloyd-Cox, J., & Beaty, R. (2022). AuDrA: An automated drawing assessment platform for evaluating creativity; doi: [10.31234/osf.io/t63dm](https://doi.org/10.31234/osf.io/t63dm).
- Paulus, D.H., & Renzuli, J.S. (1968). Scoring creativity tests by computer. *Gifted Child Quarterly*, 12, 79–83. doi:[10.1177/001698626801200202](https://doi.org/10.1177/001698626801200202)
- Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543); doi: [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- Pürgstaller, E. (2021). Assessment of creativity in dance in children: Development and validation of a test instrument. *Creativity Research Journal*, 33, 33–46; doi: [10.1080/10400419.2020.1817694](https://doi.org/10.1080/10400419.2020.1817694).
- Richard, V., Aubertin, P., Yang, Y.Y., & Kriellaars, D. (2020). Factor structure of Play creativity: A new instrument to assess movement creativity. *Creativity Research Journal*, 32, 383–393; doi: [10.1080/10400419.2020.1821567](https://doi.org/10.1080/10400419.2020.1821567).
- Rosenzweig, E.Q., & Miele, D.B. (2016). Do you have an opportunity or an obligation to score well? The influence of regulatory focus on academic test performance. *Learning and Individual Differences*, 45, 114–127; doi: [10.1016/j.lindif.2015.12.005](https://doi.org/10.1016/j.lindif.2015.12.005).
- Runco, M.A., & Jaeger, G.J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24, 92–96; doi: [10.1080/10400419.2012.650092](https://doi.org/10.1080/10400419.2012.650092).
- Runco, M.A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement*, 52, 213–221; doi: [10.1177/001316449205200126](https://doi.org/10.1177/001316449205200126).
- Said-Metwaly, S., Taylor, C.L., Camarda, A., & Barbot, B. (2022). Divergent thinking and creative achievement—How strong is the link? An updated meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*; doi: [10.1037/aca0000507](https://doi.org/10.1037/aca0000507). Online ahead of print.
- Said-Metwaly, S., Van den Noortgate, W., & Barbot, B. (2021). Torrance Test of Creative Thinking-Verbal, Arabic version: Measurement invariance and latent mean differences across gender, year of study, and academic major. *Thinking Skills and Creativity*, 39, 100768. doi: [10.1016/j.tsc.2020.100768](https://doi.org/10.1016/j.tsc.2020.100768).
- Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309–316; doi: [10.1007/BF02288586](https://doi.org/10.1007/BF02288586).
- Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, 44, 922–932; doi: [10.1037/0003-066X.44.6.922](https://doi.org/10.1037/0003-066X.44.6.922).
- Silvia, P.J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, 6, 24–30; doi: [10.1016/j.tsc.2010.06.001](https://doi.org/10.1016/j.tsc.2010.06.001).
- 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, 79–85; doi: [10.1016/j.tsc.2009.06.005](https://doi.org/10.1016/j.tsc.2009.06.005).
- Silvia, P.J., Winterstein, B.P., Willse, J.T., Barona, C.M., Cram, J.T., Hess, K.L., . . . & 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, 68–85; doi: [10.1037/1931-3896.2.2.68](https://doi.org/10.1037/1931-3896.2.2.68).

## Rater Disagreement in Creativity Assessment

- Simonton, D.K. (2000). Creativity: Cognitive, personal, developmental, and social aspects. *American Psychologist*, 55, 151–158; doi: [10.1037/0003-066X.55.1.151](https://doi.org/10.1037/0003-066X.55.1.151).
- Simonton, D.K. (2012). Taking the US Patent Office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal*, 24, 97–106; doi: [10.1080/10400419.2012.676974](https://doi.org/10.1080/10400419.2012.676974).
- Stemler, S.E., & Kaufman, J.C. (2020). Are creative people better than others at recognizing creative work? *Thinking Skills and Creativity*, 38, 100727; doi: [10.1016/j.tsc.2020.100727](https://doi.org/10.1016/j.tsc.2020.100727).
- Sung, Y.T., Cheng, H.H., Tseng, H.C., Chang, K.E., & Lin, S.Y. (2022). Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques. *Psychology of Aesthetics, Creativity, and the Arts*; doi: [10.1037/aca0000450](https://doi.org/10.1037/aca0000450). Online ahead of print.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288; doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x).
- Torrance, E.P. (1969). Prediction of adult creative achievement among high school seniors. *Gifted Child Quarterly*, 13, 223–229; doi: [10.1177/001698626901300401](https://doi.org/10.1177/001698626901300401).
- Urminsky, O., Hansen, C., & Chernozhukov, V. (2016). Using double-lasso regression for principled variable selection. (SSRN Scholarly Paper No. 2733374) doi: [10.2139/ssrn.2733374](https://doi.org/10.2139/ssrn.2733374).
- Wang, J., & Long, H. (2022). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. *Psychology of Aesthetics, Creativity, and the Arts*; doi: [10.1037/aca0000470](https://doi.org/10.1037/aca0000470). Online ahead of print.
- Willink, R. (2003). Relationships between central moments and cumulants, with formulae for the central moments of gamma distributions. *Communications in Statistics – Theory and Methods*, 32, 701–704; doi: [10.1081/STA-120018823](https://doi.org/10.1081/STA-120018823).
- Wilson, R.C., Guilford, J.P., & Christensen, P.R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, 50, 362.
- Zhou, J., Wang, X.M., Song, L.J., & Wu, J. (2017). Is it new? Personal and contextual influences on perceptions of novelty and creativity. *Journal of Applied Psychology*, 102, 180–202; doi: [10.1037/apl0000166](https://doi.org/10.1037/apl0000166).
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429; doi: [10.1198/016214506000000735](https://doi.org/10.1198/016214506000000735).

---

Denis Dumas, University of Georgia

Selcuk Acar, Kelly Berthiaume, University of North Texas

Peter Organisciak, University of Denver

David Eby, University of Illinois Urbana-Champaign

Katalin Grajzel, University of Denver

Theadora Vlaamster, University of North Texas

Michele Newman, University of Washington

Melanie Carrera, University of North Texas

Correspondence concerning this article should be addressed to Denis Dumas, Department of Educational Psychology, University of Georgia: 624 Aderhold Hall, 110 Carlton St., Athens, GA, 30602. E-mail: [denis.dumas@uga.edu](mailto:denis.dumas@uga.edu)

## ACKNOWLEDGEMENTS

This research was funded by a grant from the US. Department of Education's Institute for Education Sciences Grant (#R305A200199) to Selcuk Acar, Denis Dumas, and Peter Organisciak.