Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates

Boris Forthmann1*, Sue Hyeon Paek2, Denis Dumas3, Baptiste Barbot4,5 and Heinz Holling6

1Institute of Psychology in Education, University of Münster, Germany
2School of Psychological Sciences, University of Northern Colorado, Greeley, Colorado, USA
3Department of Research Methods and Information Science, University of Denver, Colorado, USA
4Psychological Sciences Research Institute, Université catholique de Louvain, Belgium
5Yale University, Child Study Center, New Haven, Connecticut, USA
6Institute of Psychology, University of Münster, Germany

Background. The originality of divergent thinking (DT) production is one of the most critical indicators of creative potential. It is commonly scored using the statistical infrequency of responses relative to all responses provided in a given sample.

Aims. Response frequency estimates vary in terms of measurement precision. This issue has been widely overlooked and is addressed in the current study.

Sample and method. Secondary data analysis of 202 participants was performed. A total of 900 uniquely identified responses were generated on three DT tasks and subjected to a 1-parameter logistic model with a response as the unit of measurement which allowed for the calculation of response-level conditional reliability (and marginal reliability as an overall summary of measurement precision).

Results. Marginal reliability of response propensity estimates ranged from .62 to .67 across the DT tasks. Unique responses in the sample (the basis for the classic uniqueness scoring) displayed the lowest conditional reliability (across tasks: ≈ .50). Reliability increased nonlinearly as a function of both the frequency of occurrence predicted by the model (conditional reliability) and sample size (conditional and marginal reliability).

Conclusions. This study indicates that the common practice of frequency-based originality scoring with typical sample sizes (e.g., N = 100 to N = 200) yields unacceptable levels of measurement precision (i.e., in particular for highly original responses). We further offer recommendations to mitigate the lack of measurement precision of frequency-based originality scores for DT research.
Creativity, or the mental ability to generate original and meaningful ideas (Runco & Jaeger, 2012), is a critically important human capacity that must be supported by educators. However, the psychological processes that are antecedent to creative accomplishment are not currently well understood in the educational psychology literature, limiting extant efforts of researchers to provide practically relevant recommendations. One principal area within creativity research that has perennially hindered the field’s relevance to educational practice has been the high degree of measurement imprecision (i.e., low reliability) in estimates of student creative attributes (Plucker & Makel, 2010). In the current research, we apply a novel psychometric focus to the measurement of one highly studied dimension of creativity: originality (Dumas & Dunbar, 2014). In doing so, we sought to improve educational psychologists’ and creativity researchers’ ability to quantitatively tap the original thinking of students, with a more distal goal of improving the field’s psychological understanding of creativity and therefore the capacity of educators to support the creative potential of all students.

Across definitions, originality is the primary facet of creativity. Accounting for it in creativity assessment is therefore mandatory (Runco, 2011; Zeng, Proctor, & Salvendy, 2011). Although numerous resources come into play in creative work (Sternberg & Lubart, 1995), divergent thinking (DT) – the ability to generate multiple novel solutions for a given problem (Guilford, 1967) – is the most studied component, and historically, the most common way to operationalize creative potential (Kaufman, Plucker, & Baer, 2008). This is well documented by 538 hits in the PsychInfo database for work with DT in the title, and 983 hits for work with listing DT as a keyword in a recent search (retrieved on 12 July 2019). In addition, broader creativity or cognitive ability measures in the United States (e.g., Scales for Rating the Behavioral Characteristics of Superior Students-III. Creativity Characteristics; Renzulli, Smith, White, Callahan, & Hartman, 1976; Torrance Tests of Creative Thinking; Torrance, 2017) and Europe (e.g., Berlin structure of intelligence test for youth: assessment of talent and giftedness; Jäger et al., 2006) assess DT as part of their test conception.

Most prominently, and since the pioneering work of Guilford (1968) and Torrance (1963), DT research and its methodological scrutiny (Cropley & Clapson, 1971; Vernon, 1971) are rooted in educational psychology. In this vein, DT has been used to answer big questions regarding whether children’s creative potential is a valid measure to determine their eligibility for special educational opportunities (Runco & Albert, 1985), whether DT scores assessed in childhood can predict creative performance at a later age (Runco, Millar, Acar, & Cramond, 2010), and whether creative potential is distinct from general intelligence (Kim, 2008). This tradition has sustained as found in several recent research endeavours on these classical issues (e.g., Dumas, 2018; Paek & Runco, 2018), but also in newly emerging strands of creativity research regarding what role creative thinking plays in critical thinking, school achievement, arts and even mathematics learning (e.g., Chang, Li, Chen, & Chiu, 2015; Gajda, Karwowski, & Beghetto, 2017; van de Kamp, Admiraal, Drie, & Rijaardsdam, 2015). The common use of DT tests is also further highlighted in developmental psychology (e.g., Charles & Runco, 2001; Wallace & Russ, 2015), clinical psychology (e.g., Acar, Chen, & Cayirdag, 2018; Ludyga et al., 2018), social psychology (e.g., De Dreu et al., 2014; Ritter et al., 2012), organizational psychology (e.g., Carmeli, Gelbard, & Reiter-Palmon, 2013; Lu et al., 2017), and neurocognitive studies on creative thinking (e.g., Gilhooly, Fioratou, Anthony, & Wynn, 2007; Hass, 2017).

Divergent thinking is classically operationalized by open-ended tasks. For example, in the Alternate Uses Task (AUT), participants are instructed to think of multiple uses for an everyday object (e.g., knife; see Table 1) that diverge from the objects’ intended use...
Although there is no definite consensus as to how to score DT tasks (Reiter-Palmon, Forthmann, & Barbot, 2019), most DT studies account at least for ideational fluency, which refers to the count of all valid responses generated by a person. Hence, this score reflects a person’s ideational productivity. Often, only fluency scores are derived from DT task protocols (Runco & Acar, 2012) because this score highly correlates with other summative DT performance scores (Forthmann, Szardenings, & Holling, 2018; Mouchiroud & Lubart, 2001). However, it is increasingly acknowledged that this operationalization loses the theoretical connection with the concept of creativity (e.g., Barbot, 2018; Zeng et al., 2011). More in line with the pioneer conceptualization of DT, other scores of the DT production tapping into the quality of the responses, with its focus on originality, are typically accounted for in DT assessment (Forthmann, Holling, Çelik, Storme, & Lubart, 2017; Runco, 2011; Zeng et al., 2011).

Frequency-based scoring of DT originality
Classically, indicators for originality are associative remoteness, cleverness, and – the most typically used one – uncommonness of the responses (Wilson, Guilford, & Christensen, 1953). In DT research, the latter is generally based on the statistical rarity of the responses (i.e., relative infrequency in the sample; e.g., Forthmann et al., 2017; Mouchiroud & Lubart, 2001; Wallach & Kogan, 1965). For example, if the response screwdriver appears two times as a response to an alternate use of a knife among a sample of five persons, the relative frequency of this response would be .40 (i.e., 2/5) and, thus, its statistical rarity would be .60 (i.e., 1 – .40; see also Table 1). However, such frequency-based originality scores have been criticized for being confounded by fluency scores, for being blind to fuzzy responses, and issues related to sample size dependence of the derived scores (Silvia et al., 2008). While the former issues have been extensively studied in recent years (e.g., Reiter-Palmon et al., 2019; Forthmann, Szardenings, et al., 2018), the present study focuses on the issue of sample size dependence. Specifically, this work focuses on the reliability of response-level frequencies as a function of sample size.

Frequency-based originality scores of DT tests are attractive because of their objectivity (Runco, 2008) and face validity (original ideas should not appear very often; Silvia et al., 2008). In practice, the scoring process starts with a cross-tabulation of persons and responses to calculate the frequency of occurrence of each response (Cropy, 1967; Reiter-Palmon et al., 2019). In scoring DT originality, these frequencies are usually referred to as relative frequencies. Individual responses are weighted by these frequencies.

**Table 1.** Hypothetical frequency table of occurrence for four persons who generated responses on the Alternate Uses Task with stimulus knife

<table>
<thead>
<tr>
<th>Response</th>
<th>Person 1</th>
<th>Person 2</th>
<th>Person 3</th>
<th>Person 4</th>
<th>Absolute frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weapon</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>Dart</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>Screwdriver</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>Cake server</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>.25</td>
</tr>
<tr>
<td>Stirring coffee</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>.50</td>
</tr>
<tr>
<td>Fluency</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

(Guilford, 1967; Wallach & Kogan, 1965).
in various ways and then aggregated into originality scores for each participant. For example, uniqueness scoring awards a point to responses generated by only one person in the tested sample (Murphy, 1973; Silvia et al., 2008; Wallach & Kogan, 1965). Accordingly, all other responses (i.e., those proposed by two or more respondents) are not credited for originality. Similarly, threshold scoring credits originality points according to defined relative frequency thresholds (Cropley, 1967, 1972; Runco, 2008; Torrance, 1966). For example, using 5% threshold scoring, a response is credited for originality if it was proposed by <5% of the tested sample (this approach has been also named unusualness scoring; see Runco, 2008). Finally, relative response frequencies are also often used directly to derive an originality score (Forthmann et al., 2017; Mouchiroud & Lubart, 2001). Accordingly, relative response frequencies are transformed into infrequency weights by subtracting each relative frequency from 1 prior to aggregation of scores (see example above). Then, infrequency weights can be averaged across responses to a DT prompt, to yield a person’s originality score. For example, Person 4 in Table 1 would receive an average weighted originality score of \((.00 + .75 + .50)/3 = .42\).

**Measurement precision of the estimates of relative response frequency**

First, it is important to focus on the frequency of occurrence tables that are used for scoring. Table 1 illustrates a hypothetical example with four persons who generated responses on an AUT for *knife* as a stimulus. The table is arranged as a matrix with responses in rows and persons in columns. In every cell of the matrix is either a 0, indicating that a given person did not provide a given response, or a 1 when a given response was provided by a given person. The row sums presented in the column labelled *absolute frequencies* are the frequencies of occurrence for each response. Dividing them by the number of persons yields the response’s relative frequency of occurrence which builds the basis for further originality scoring (see above). Despite their common use in the literature (e.g., the Torrance Test of Creative Thinking which is a widely used DT measure has scoring rubrics that are referenced on frequency-based scoring; Torrance, 1966, 2017), it has been widely overlooked that these relative frequencies are sample-specific and are therefore only estimates of the probability of a response to be provided in the target population. When such a population parameter is estimated, there is always a degree of uncertainty in that parameter, which can be quantified in terms of measurement precision. However, despite decades of DT research and hundreds of studies using frequency-based originality scoring at the level of individual participants, measurement precision has never before been estimated at the response level. In order to do so, it is first necessary to define how response frequency estimates can be modelled from a psychometric perspective.

As illustrated in Table 1, the frequency of occurrence estimates for each response, or sums by row, reflects the responses’ main effects, whereas fluency scores (sums by column) reflect the person main effects. Thus, the probability \(P\) of a given response being provided can be modelled as a function of the response and the person main effects in the following logistic model:

\[
P(X = 1|\beta_i, \theta_v) = \frac{\exp(\beta_i + \theta_v)}{1 + \exp(\beta_i + \theta_v)},
\]

with \(\beta_i\) being the propensity of response \(i\) to be provided (analogous to item easiness in traditional item response theory [IRT] modelling), and \(\theta_v\), the ideation parameter of
person \( v \) on the particular prompt being scored. At the logit level, the \( \beta_i \) can be assumed to follow a normal distribution with mean zero and variance \( \sigma^2_{\beta_i} \). This model is a variant of the 1-parameter logistic model (1PL; see De Boeck et al., 2011). Hence, this approach allows to examine the probability of a response generated in a DT task from an IRT perspective.

Importantly, using the 1PL allows for the quantification of the reliability of \( \beta_i \) estimates. First, their conditional reliability (i.e., reliability depending on the level of \( \beta_i \); e.g., Green, Bock, Humphreys, Linn, & Reckase, 1984) can be calculated according to Brown and Croudace (2015):

\[
\text{Rel}(\beta_i) = 1 - \frac{SE_{\beta_i}^2}{s_{\beta_i}^2}.
\]

The squared standard error of an estimate of \( \beta_i \) is denoted by \( SE_{\beta_i}^2 \), and the estimated variance of the \( \beta_i \) distribution is denoted by \( s_{\beta_i}^2 \). In addition, marginal (empirical) reliability (\( SE_{\beta_i}^2 \) in the formula above is replaced by the average squared \( SE \) across all \( i \)s) can be calculated to get an overall reliability estimate (Brown & Croudace, 2015; Green et al., 1984).

**Aim of the current study**

This work sought to examine the extent to which sample size affects marginal reliability as a general estimate of measurement precision, as well as conditional reliability of \( \beta_i \) estimates. In other words, the sample dependence of statistical rarity as an indicator of originality was scrutinized with a focus on reliability. Together, this work (1) outlines the importance of accounting for the measurement precision of frequency-based originality scoring in DT research, (2) provides methodological directions to do so, and (3) may help derive recommendations regarding the minimum sample size needed in DT research relying on frequency-based scoring.

**Method**

All raw data necessary to reproduce the reported results and data analysis scripts are published in the Open Science Framework (https://osf.io/gce5k/).

**Dataset**

This work is based on a secondary data analysis. The dataset taken from Forthmann et al. (2017) contained responses to a classic AUT (Guilford, 1967; Wallach & Kogan, 1965), with three objects presented to participants: paperclip, garbage bag, and rope. Participants had 2.5 min to respond to each object-prompt. Explicit instructions to be creative were given (Harrington, 1975): *Please try to write down as many uncommon and creative uses for a [object-prompt] as you can think of.* This instruction is considered as a hybrid instruction, which sets simultaneously the focus on both the productivity and the quality of responses.

The analysis was based on data provided by 202 participants (58 males and 144 females; age: \( M = 24.51, SD = 6.81; 78.22\% \) were university students; 51.49\% reported high-school graduation, 23.27\% university graduation, and 16.34\% a finished apprenticeship as their highest educational level). However, the main focus here was on responses as a unit of measurement. Overall, a total of \( N = 900 \) uniquely identified responses entered
the analysis (see Table 2). These responses were coded for frequency tabulation by the first author. Responses that differed only by functionally irrelevant features were treated as equal (e.g., *drink milk* and *drink juice* as uses for a cup; see also Reiter-Palmon et al., 2019). Absolute response frequencies ranged from 1 to 76 (relative frequencies ranged from .005 to .455). One participant did not provide responses on two of the tasks, but remained in the analysis (see Table 2 for details).

**Data preparation and analytic strategy**

For the analysis, each AUT response was coded as illustrated in Table 1, with either 0 (response not generated) or 1 (response generated) for every person in the sample, and each AUT task, independently. Response and person parameters were estimated by means of the R package *mirt* (Chalmers, 2012). The $\beta$ parameters were estimated by the expected a posteriori method as implemented in the *mirt* package. The 1PL (see Equation 1) was fitted to each task separately. To further check adequacy of the 1PL, this model was compared with a 2PL according to which levels of discrimination of the persons are allowed to vary:

$$P(X = 1|\beta_i, \theta_v, \alpha_v) = \frac{\exp(\alpha_v(\beta_i + \theta_v))}{1 + \exp(\alpha_v(\beta_i + \theta_v))},$$

with $\alpha_v$ being the person discrimination parameter. That is, responses with a lower response propensity $\beta_i$ have a lower likelihood to appear as compared to responses with higher response propensity, given that discrimination is rather high. In other words, response probabilities are more comparable for persons with low discrimination. Thus, the 1PL might be too restrictive in this regard because discrimination parameters are not allowed to vary (across persons in this application). This assumption of constant discrimination is indeed an empirical question and was, thus, tested here. Models were contrasted by likelihood-ratio tests. In addition, model fit was examined by means of covariate-adjusted frequency plots (CAFP; Holling, Böhning, & Böhning, 2015) based on the frequency counts of absolute response frequencies. The distinction between frequency counts and absolute response frequencies can be exemplified in relation to Table 1: The absolute response frequencies occurring were 1, 2, and 4 with frequency counts of 1, 3, and 1 (i.e., one response occurring once, three responses occurring twice, and one response occurring four times in the sample), respectively. Absolute response frequencies based on the 1PL or 2PL are known to follow a generalized binomial distribution (González, Wiberg, & von Davier, 2016; Lord, 1980) and density values for predicted (model-implied) frequency counts, as required for the construction of CAFPs, were calculated with functions provided by the R package *GenBinomApps* (Lewitschnig

<table>
<thead>
<tr>
<th>Table 2. Solution space characteristics for each of the Alternate Uses Task tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paperclip</strong></td>
</tr>
<tr>
<td>Number of unique responses</td>
</tr>
<tr>
<td>Number of non-unique responses</td>
</tr>
<tr>
<td>Number of non-redundant responses</td>
</tr>
<tr>
<td>Total number of responses</td>
</tr>
<tr>
<td>Number of persons</td>
</tr>
</tbody>
</table>
In a CAFP, the observed frequency counts of response frequencies are compared with model-implied frequency counts (the closer the fit, the better for a given model).

**Results**

**Descriptive statistics**

As shown in Table 2, the number of unique responses ranged from 112 to 154 and the number of non-unique responses ranged from 149 to 186 across the three AUT tasks. In addition, the task with object-prompt garbage bag showed the highest number of unique responses and also the highest number of non-redundant responses (see Table 2). Thus, it showed the largest solution space of the three tasks in the study sample. Relatedly, it was observed that the largest amount of responses was provided for the task with object-prompt rope, meaning that the overall number of responses provided does not necessarily correspond with the number of non-redundant responses.

**Fitting the 1PL**

Model fit of the 1PL as indicated by CAFPs was good (see Figure 1). Only the frequency count of unique responses was clearly underestimated by the 1PL across all three AUT tasks (see Figure 1). In addition, CAFPs revealed that the more complex 2PL did not fit better to the observed counts as compared to the 1PL. This was indicated by the almost perfect correspondence between model-implied frequency counts of response frequencies based on the 1PL and 2PL, respectively, across all AUT tasks in the bottom row in Figure 1. In this regard, it is, however, noteworthy that likelihood-ratio tests were in favour of the 2PL for all three AUT tasks (see Table 3). However, information criteria results, also taking model parsimony into account, were all in favour of the 1PL. In addition, marginal reliability estimates did not differ between the 1PL and 2PL (see Table 3) and, hence, we relied on the less complex 1PL for all further analyses. Marginal reliability based on the 1PL across all of the response propensity estimates was found to be .66 (paperclip), .62 (garbage bag), and .67 (rope), respectively. Conditional reliability of single response propensity estimates ranged from .52 to .98 (paperclip), .47 to .97 (garbage bag), and .52 to .97 (rope). In sum, reliability of parameter estimates varied greatly.

**Sample size and conditional reliability**

This variation was further examined in connection with sample size by means of resampling from the original data. Participants were resampled 100 times to create datasets of three different sample sizes: (1) $N = 50$, (2) $N = 100$, and (c) $N = 150$. These sample sizes were chosen based on recent work by Said-Metwaly, Van den Noortgate, and Kyndt (2017) who pointed out that sample sizes in DT research can be as small as $N = 30$ and, hence, $N = 50$ was chosen to approximate a typical small sample. In addition, classic studies such as Wallach and Kogan’s (1965) often had sample sizes in the range from 100 to 200 (e.g., Wallach and Kogan had $N = 151$) and, thus, the range of sample sizes chosen here reflects common sample sizes in DT research.

Then, separately for each AUT task, response propensity reliabilities were estimated in each resampled dataset and averaged across datasets of the same sample size (note that this implies that each possible response was not necessarily drawn in every dataset).
Response propensity reliabilities are depicted as a function of predicted response probability and sample size in Figure 2. First, conditional response propensity reliability increases nonlinearly as a function of estimated response probabilities: Reliability estimates are expected to increase up to a propensity estimate of .50 and to decrease from that point onward (i.e., the relationship follows an inverted U-shape). The maximum of estimated response probabilities for the resampled datasets was below .50. Thus, unique responses had the lowest conditional reliability, and the most common responses had the highest reliability. In addition, conditional reliability increased as a function of sample size (Figure 2). This is also illustrated in averaged estimates of marginal reliability for paperclip with values of .14, .47, and .59 when sample sizes are \( N = 50 \), \( N = 100 \), and \( N = 150 \), respectively. Similar results were obtained for garbage bag \( (N = 50: .03; N = 100: .39; \) and \( N = 150: .54) \) and rope \( (N = 50: .12; N = 100: .47; \) and \( N = 150: .60) \). Thus, the step from \( N = 50 \) to \( N = 100 \) resulted in a larger increase in reliability as compared to the increase from \( N = 100 \) to \( N = 150 \) (see also Figure 2).

**Discussion**

Scoring DT responses for originality is critical given the importance of this facet for creativity (Runco, 2011; Zeng et al., 2011). Originality in DT tests is often based on
<table>
<thead>
<tr>
<th></th>
<th>Paperclip</th>
<th></th>
<th>Garbage bag</th>
<th></th>
<th>Rope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1PL</td>
<td>2PL</td>
<td>1PL</td>
<td>2PL^a</td>
<td>1PL</td>
</tr>
<tr>
<td>AIC</td>
<td>11,252.63</td>
<td>11,334.68</td>
<td>13,584.06</td>
<td>13,632.84</td>
<td>13,571.88</td>
</tr>
<tr>
<td>BIC</td>
<td>11,977.00</td>
<td>12,776.29</td>
<td>14,357.51</td>
<td>15,172.08</td>
<td>14,318.69</td>
</tr>
<tr>
<td>LR test (df)</td>
<td>319.44 (201)***</td>
<td>351.22 (200)***</td>
<td>399.20 (200)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE(\beta)</td>
<td>0.32</td>
<td>0.26</td>
<td>0.31</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>s^2(\beta)</td>
<td>0.95</td>
<td>0.73</td>
<td>0.82</td>
<td>0.72</td>
<td>0.87</td>
</tr>
<tr>
<td>Marginal reliability</td>
<td>.66</td>
<td>.65</td>
<td>.62</td>
<td>.62</td>
<td>.67</td>
</tr>
</tbody>
</table>

Note. AIC = Akaike’s information criterion; BIC = Bayesian information criterion; lower values on AIC and BIC imply that a model is superior in terms of fit and model parsimony as compared to a model with higher values (only models fit for the same task are comparable here); LR test = likelihood-ratio test for comparing the 1PL with the 2PL; \(\text{SE}\(\beta\)\) = average of the squared standard errors for the \(\beta\) estimates; \(s^2\(\beta\)\) = variance of the \(\beta\) estimates. Marginal (empirical) reliability = \(1 - \frac{\text{SE}\(\beta\)\^2}{s^2\(\beta\)}\) (see Brown & Croudace, 2015).

^aThe 2PL for garbage bag had to be estimated with nlminb as optimizer for the M-step in the expectation maximization algorithm because the default L-BFGS-B optimizer did not yield convergent results at the default level of tolerance (for details, see Chalmers, 2012).

***p < .001.
frequencies of occurrence of responses in a study sample. The current work addressed a widely overlooked problem with such originality scorings, namely the potential lack of measurement precision associated with the relative frequency of a given response (i.e., response probability estimate). We proposed to use the 1PL to study this issue and estimate the reliability of parameters relating to response frequencies which are the basis of frequency-based scoring of originality in DT tasks. Hence, this work highlights the importance of measurement precision at the level of DT responses as an initial step to establish reliability of a DT scale.

Overall, our work has demonstrated that for sample sizes of $N = 50$ and $N = 100$, both conditional reliability of unique responses and marginal reliability are far below acceptable levels. With marginal reliability estimates of around $.60$, a sample size of $N = 150$ was closer to acceptable values, but still deemed unreliable to an unacceptable degree. These observations are particularly important because studies on DT sometimes rely on very small sample sizes of $N = 30$ or smaller (see Said-Metwaly et al., 2017). Moreover, it is not uncommon that frequency norms from samples of $N = 100$ are used (e.g., De Dreu et al., 2014) and exemplary DT research such as Wallach and Kogan’s (1965) classic study used a sample size of $N = 151$ (see also van de Kamp et al., 2015). Even the sample size used in the current study did not yield overall satisfactory results in terms of measurement precision of response frequencies which are the basis for methods.
scoring originality. An extrapolation of the presented findings suggests that, to obtain a good level of marginal reliability for each of these DT task prompts (i.e., > .80), one would need a sample size between $N = 300$ and $N = 400$.

However, it might not be sufficient to obtain acceptable conditional reliability, when particularly rare responses are most salient. Indeed, the results demonstrated that unique responses (those proposed only once) in a sample are associated with the lowest conditional reliability (across tasks: ≈ .50), at least from the IRT perspective used in the current work. This is problematic when using DT tests in research and practice such as for the identification and placement of creative students, because unique responses are strongly weighted in originality scoring (e.g., Wallach & Kogan, 1965). In practice, this means that study participants with the highest level of originality will have the least reliable scores, whereas those participants who propose the most common responses in their DT protocol will have the most reliable scores. By extension, uniqueness scoring will prove particularly unreliable.

Additionally, the work presented here also questions the practice of crediting originality points according to chosen thresholds (e.g., Cropley, 1967; Torrance, 1966) in rather small samples (e.g., Carmeli et al., 2013; De Dreu et al., 2014; von Stumm & Scott, 2019), because measurement precision of frequency estimates can be considerably low in small samples. Indeed, with small sample sizes, it is hard to justify that the frequency of individual responses is indeed below the designated threshold (i.e., crediting a response for originality that is provided by 4% of the cases with a chosen threshold of 5%). Consequently, originality scoring based on frequency thresholds should be used in adequately large samples.

Moreover, researcher-identified thresholds may be particularly problematic because they are always somewhat arbitrarily chosen, and using thresholds for scoring is accompanied by a loss of information about the uncommonness of the responses a participant generated in DT tasks. In addition, a more detailed relative frequency scoring procedure can also be used directly in conjunction with information theory to determine substantively important aspects of participant DT (see Hass, 2016). Thus, one might question the use of threshold scoring at all. However, in applied testing contexts, a simple rule such as threshold scoring is still attractive for some practitioners given its simplicity and ease to be utilized. Moreover, when test norms are created, it is a desirable feature that small differences between scores are not over-interpretable (Kolen, 2006). This feature of test norms can be achieved by avoiding a scoring scale that is overly finely granular (i.e., the scale has more points on it than are useful to test practitioners; Kolen, 2006). In our view, it may be worthwhile to further investigate how threshold scoring could support this purpose. Hence, it seems premature to abandon threshold scoring entirely, but especially in a research setting, one should very carefully weigh the benefits and pitfalls of using such a method.

Based on the current findings, the following tentative solutions for the sample size issue in DT research and measurement practice are anticipated: (1) an adequately large sample size should be used for frequency-based originality scoring, (2) an adequately large and representative norming sample should be consulted when using frequency scoring methods for DT tasks with smaller sample sizes (see Torrance, 2017), (3) rater-based originality scoring may be recommended instead of frequency-based scoring for smaller samples (e.g., Hass, Rivera, & Silvia, 2018; Silvia et al., 2008), or (4) other objective methods such as latent semantic analysis may be recommended for small samples (e.g., Dumas & Dunbar, 2014; see, however, Forthmann, Oyebade, Ojo, Günther, & Holling, 2018, for technical problems that still need to be solved when DT responses are scored by
means of latent semantic analysis). Indeed, more research is needed to accurately determine what sample sizes are required for the reliable identification of the original responses in various types of DT tasks, and using various scoring approaches.

Indeed, our study revealed that marginal reliability does not only depend on the number of persons in the study. Another crucial factor is the solution space. Marginal reliability was found to be lowest for garbage bag which was the AUT task with the least constrained solution space. Thus, the more constrained the solution space (less unique responses and less non-redundant responses), the larger marginal reliability in AUT tasks implying that required sample sizes for such tasks would be lower. In contrast, other families of DT tasks, such as the Consequences Task (e.g., Christensen, Guilford, & Wilson, 1957) with an expectedly less constrained solution space, are likely to have lower marginal reliability as compared to the AUT family.

Another question is whether time-on-task may alter the characteristics of the solution space and, thereby, the measurement precision of response propensity estimates. It is likely that a longer time-on-task yields more unique responses as demonstrated in a serial order effect (Christensen et al., 1957), and in turn, less reliable response propensity estimates. However, recent studies on the serial order effect (Hass, 2017; Hass & Beaty, 2018) found a nonlinear response rate with asymptotic levels of responding after 1–2 min. Hence, for reliable frequency estimates it could be more efficient to test more participants instead of adjusting time-on-task. Future research should test these issues with larger sample sizes and a systematic focus on characteristics of task solution spaces.

Indeed, the measurement of DT and other creative attributes is interesting and necessary across a variety of educational psychology research contexts, because such creative thinking skills are relevant to an array of formal and informal learning. As such, the findings from this investigation can be generalized to any DT tasks as long as the tasks and their solution spaces are comparable to what was tested in the current study (e.g., AUT). Also, the current study provides a useful means of examining measurement precision in any types of DT tasks or other measures with a similar scoring method for originality. For example, such scores have been calculated for negotiation tactics (De Dreu & Nijstad, 2008), brainstorming tasks administered to small groups (Diehl & Stroebe, 1987), mathematical creative problem-solving (Kim, Cho, & Ahn, 2004), word associations (Nemeth & Kwan, 1985), scientific problem finding (Hu, Shi, Han, Wang, & Adey, 2010), melodic originality in music (Hass, 2016), and this list can be easily extended.

In estimating relative frequencies for each response, fixed person and random response propensity effects were fitted in the 1PL. This method allows for the quantification of the marginal and conditional reliability of parameter estimates at the logit level. Therefore, in keeping with efforts to accurately and extensively report on the characteristics of DT tasks used in research (Reiter-Palmon et al., 2019), researchers are encouraged to report the range of conditional reliability of single response frequencies, and marginal reliability estimates for their set of responses when statistical rarity is used to score originality, especially when samples are small.

A limitation of this study is that only one person conducted the cross-tabulation of responses. This is a common scenario in DT research but potentially constitutes another source of measurement error which was not accounted for here. Given the psychometric focus of the current work, this limitation unlikely undermines any of the conclusions drawn from the results of the current study. But it is further recommended that, in future substantively oriented studies of DT, at least two raters cross-tabulate the responses and solve disagreement by discussion, for example. It should further be noted that our choices with respect to inclusiveness of response categories had a direct effect on the response
frequency distribution and, thus, also on reliability. That is, using less inclusive response categories by treating responses with functionally irrelevant features as different would have yielded even more unique responses and overall lower reliability of response propensity estimates. Thus, any choices need to be made carefully in categorizing and scoring responses.

**Conclusion**

Statistical rarity as an indicator of originality is psychometrically unacceptable when frequency estimates are taken in a small sample, and other scoring methods (such as rater-based scoring or semantic network scoring) are warranted in this context. Application of frequency-based originality scoring requires a large sample used for frequency estimates (i.e., >300). Otherwise, a score of a participant’s original thinking that is derived from frequencies would be conflated with considerable measurement error, implying high imprecision. However, given the central role that creative attributes play in the development of talents and, thus, educational and economic success of individuals, it is critical to measure these attributes with the highest degree of precision. To obtain a high degree of precision when planning DT studies or when using DT assessments in educational contexts, researchers and other test users must take sample size and characteristics of the task’s solution space into account.

**Acknowledgements**

This research was supported by grant HO 1286/11-1 of the German Research Foundation (DFG) to Heinz Holling.

**Conflicts of interest**

All authors declare no conflict of interest.

**References**


Received 10 April 2019; revised version received 29 July 2019