Don't Throw the "Bad" Ideas Away! Multidimensional Top-Scoring Increases Reliability of Divergent Thinking Tasks

Boris Forthmann[1], Maciej Karwowski[2], & Roger E. Beaty[3]

[1]Institute of Psychology in Education, University of Münster

[2]Institute of Psychology, University of Wroclaw

[3]Department of Psychology, Pennsylvania State University

Author Note

Boris Forthmann https://orcid.org/0000-0001-9755-7304

Maciej Karwowski https://orcid.org/0000- 0001-6974-1673

Roger E. Beaty https://orcid.org/0000- 0001-6114-5973

All files for analyses are available at Open Science Framework: https://osf.io/7rgsp/?view_only=8cbc6d85b84f4a56ac6c94ec649a0cea.

We would like to thank Paul J. Silvia for providing his data for reanalysis in our Study 2. We further thank Philipp Doebler for his comments on an early draft of this paper.

R.B. is supported by a grant from the National Science Foundation [DRL-1920653].

Correspondence should be addressed to Boris Forthmann, Institute of Psychology in Education, Germany; boris.forthmann@wwu.de.

Abstract

Scoring divergent thinking tasks opens multiple avenues and possibilities – decisions researchers have to make. While some scholars postulate that scoring should focus on the best ideas provided, the measurement of the best responses (e.g., "top scoring") comes along with challenges. More specifically, compared to the average quality across all responses, top scoring uses less information—the "bad" ideas are thrown away—which decreases reliability. To resolve this issue, this article introduces a multidimensional top-scoring approach analogous to linear growth modeling which retains information provided by all responses (best ideas and "bad" ideas). Across two studies, using both subjective human ratings and semantic distance originality scoring of responses to over a dozen divergent thinking tasks, we demonstrated that Maximum (the best idea) and Top2 Scoring (two best ideas) could surpass typically applied average scoring in measurement precision when the "bad" ideas' originality is used as auxiliary information (i.e., additional information in the analysis). We thus recommend retaining all ideas when scoring divergent thinking tasks, and we discuss the potential this new approach holds for creativity research and practice.

Keywords: Divergent Thinking; Top-Scoring; Maximum Scoring; Semantic Distance; Reliability

Don't Throw the "Bad" Ideas Away! Multidimensional Top-Scoring Increases Reliability of Divergent Thinking Tasks

The open-ended response format is one of the defining characteristics of divergent thinking tasks (i.e., tasks that assess the ability to generate multiple distinct solutions). From a measurement perspective, open-endedness allows test-takers to vary with respect to their number of generated responses, yielding a metric of an ideational fluency. However, if the goal is to score these responses for creative quality, numerous methods exist for aggregating responses. Two popular options are to use a person's best response (i.e., maximum scoring; Dumas et al., 2022; Girotra et al., 2010; Reiter-Palmon et al., 2009, 2019) or the average across a person's TopX responses (i.e., the top-scoring method; e.g., when X = 2, only the Top2 responses are averaged; Benedek et al., 2013; Silvia, 2011; Silvia et al., 2008). Another standard aggregation method is to use the average across all responses of a person provided for a given task (i.e., average scoring). Reliability of average scores (also referred to as ratio scores) is theoretically positively related with fluency (e.g., Cronbach, 1941), and recent empirical findings have also shown this at the task-level (Forthmann, Jankowska, et al., 2021). Put it simply, the more responses are available the more reliable are the scores (i.e., at the level of individuals and in general at the sample level). For maximum scoring and top-scoring, however, less information enters the aggregate scores, which potentially decreases measurement precision compared to average scoring (Benedek et al., 2013; Silvia, 2011).[1] Hence, the current work outlines a multidimensional scoring approach for maximum and—more generally—top-scoring of divergent thinking tasks, in which the remaining responses of the idea pool (i.e., not the best—the "bad"—ideas) are used as collateral information to increase top-scoring reliability.

Reliability from A Three-Level Framework for Divergent Thinking Assessment

[1] Notably, reliability of average scores of divergent thinking tests was found to be problematic in previous work for some combinations of tasks and scorings (e.g., Hocevar & Michael, 1979; Runco et al., 1987), yet potential reasons for these past findings go beyond the focus of the current work. The focus of this work was on the reliability of top scoring vs. average scoring and not on reliability of average (ratio) scores per se.

Divergent thinking refers to the ability to think in multiple directions when multiple solutions are possible. Tests to assess divergent thinking ability are often considered measures of future creative potential (Runco & Acar, 2012) and, indeed, divergent thinking tests were associated with creative achievements (Plucker, 1999). In addition, generating multiple ideas is also relevant from the perspective of creative process models (Lubart, 2001; Mumford & McIntosh, 2017)—such as the classic Geneplore model (Finke et al., 1996), where ideas are first divergently generated and then explored/evaluated—highlighting the importance of divergent thinking for creativity. The importance of divergent thinking is further emphasized by its relationship with leader continuance in organizational settings (Zaccaro et al., 2015), its role in collaborative creative problem-solving in safety- critical environments (Bourgeois-Bougrine, 2020) or psychotherapy (Deacon, 2000), for example. Assessing divergent thinking requires test-takers to provide multiple responses on an open-ended task (Guilford, 1967), such as the Alternate Uses Task (AUT; Guilford, 1967), which requires generating creative uses for objects.

These characteristics of divergent thinking tasks (multiple responses and open-endedness) leave many options for scoring. For example, divergent thinking tests can be scored for creative quality, such as originality (Forthmann et al., 2017; Reiter-Palmon et al., 2019). Scoring then typically follows these steps: (a) all responses are evaluated for their originality, (b) scored responses are aggregated for each task and person, and (c) scored tasks are aggregated across people (Forthmann, Jankowska, et al., 2021; Reiter-Palmon et al., 2019). Importantly, issues related to measurement precision as a fundamental psychometric property of divergent thinking tasks can emerge at each of these levels of assessment. For example, at the level of responses, measurement precision of originality depends on the number of raters (when human judges are used) or the number of participants in the sample (when statistical rarity is used; Forthmann et al., 2020; Forthmann, Jankowska, et al., 2021). Also, when responses are scored using semantic distance (i.e., vector models of word meaning; Recchia & Jones, 2009), the size and number of documents used to create semantic spaces potentially affect validity of

response. For an overview of available methods to score originality in divergent thinking, see, for example, Reiter-Palmon et al. (2019).

Silvia (2011) examined maximum reliability at the level of single tasks as a function of divergent thinking tasks (and task-types such as Alternate Uses, Consequences, and Instances) which were scored by means of subjective creativity ratings. He studied two scoring aggregation methods: (a) *average scoring*, which refers to the average score across all responses of a participant for a given task, and (b) *Top2 Scoring*, which refers to the average score across both responses that participants identified as being their two best responses (but top responses can also be identified by ranking them based on response originality; Reiter-Palmon et al., 2019). Most relevant for the context of the current work is Silvia's (2011) finding that generally, Top2 Scoring displayed lower levels of reliability as compared to average scoring (Average Scoring: reliability ranged from .64 to .90; Top2 Scoring: .59 to .79). Relatedly, Benedek et al. (2013) examined internal consistency reliability estimates for a complete test battery (comprised of 3 Alternate Uses tasks and 3 Instances tasks) scored by human raters and the top-scoring method as a function of the number of top ideas (among other factors such as time-on-task). They found that average scoring, based on all generated responses, yielded the highest reliability estimate (Cronbach's $\alpha = .87$), whereas maximum scoring (i.e., Top1 scoring) resulted in the lowest reliability (Cronbach's $\alpha < .60$).

Such findings are explained by the amount of information on which the scores are based. Average scoring is based on all responses, whereas maximum scoring is only based on a single response (Benedek et al., 2013; Silvia, 2011). Despite these lower levels of reliability of top-scoring, however, researchers have argued to focus only on the best ideas (Girotra et al., 2010) or found differential validity patterns for top-scoring (e.g., Shaw, 2021), who found only Top2 scoring to be related to a measure of intelligence, but not average scoring). Hence, the current work proposes and examines a psychometric approach that can potentially increase the reliability of top-scoring.

Furthermore, in accordance with psychometric theory (e.g., Cronbach, 1941), Forthmann et al. (2021) reported a clear dependence of task reliability on fluency when average scoring across all responses was employed. Because fluency can be indirectly controlled via time-on-task (see Forthmann et al., 2021)—and hence, the reliability of average scoring—one wonders how far this dependence generalizes to top scoring as a method of task-level aggregation. Hence, another aim of the current work was to explore the dependence of conditional reliability on fluency for Maximum Scoring and Top2 Scoring.

Using Auxiliary Information to Increase Measurement Precision

Item response theory models have shown that measurement precision of trait estimates increases when all traits are concurrently modeled as multidimensional compared to separately estimating unidimensional models (Bulut, 2013; de la Torre et al., 2011; Wang et al., 2004). The increase of measurement precision for multidimensional models, compared to separate unidimensional models, can be explained by using the full information provided by the correlational structure among latent variables (Bulut, 2013; Wang et al., 2004). Auxiliary information from any variable that correlates with the target ability might further inform ability estimation to increase reliability (de la Torre et al., 2011; Wang et al., 2004). For example, when the items at hand are indicators of two target latent abilities (e.g., inductive reasoning and mental rotation), measurement precision might increase because information from the correlation between the latent variables can be borrowed (e.g., van der Linden, 2010). In addition, collateral information from additional response behavior can be used. For example, processing speed might be auxiliary information when the target ability is measured based on accuracy (van der Linden et al., 2010). The main contribution of the current work is to adapt the idea of multidimensional scoring from item response theory to the context of the divergent thinking top scoring to potentially increase its comparably lower reliability.

Multidimensional Top Scoring of Ideational Output

The multidimensional top scoring model used in this work is inspired by a simple linear latent growth model commonly used to model growth trajectories across time (Preacher et al.,

2008). A linear latent growth model includes a random intercept that most often reflects the score at the first measurement occasion. The random intercept ensures that every person has their initial score reflected in the model (e.g., the initial level of performance of a target cognitive ability). In addition, the growth over time (e.g., learning progress over time) is modeled as a random slope that is also allowed to vary across persons (i.e., some people may stagnate across time, whereas others increase or decrease). In addition, the model estimates the random intercept and random slope as latent variables, including their variances and correlation (plus a residual variance when the dependent variable is modeled as normally distributed). Thus, the linear latent growth model is multidimensional.

To estimate a linear latent growth model, the coding of the time variable defines the interpretation of the random intercept. For example, when four time points are coded as T1 = 0, T2 = 1, T3 = 2, and T4 = 3, the random intercept can be interpreted as a person's initial level (please note that other choices can be made; see, for example, Biesanz et al., 2004; Foorman et al., 1998). If one assumes that from T1 to T2 no growth occurs, this could be reflected by the following time coding: T1 = 0, T2 = 0, T3 = 1, and T4 = 2. With this time coding, the random intercept can be interpreted as a person's average level across T1 and T2. The idea now for a multidimensional top-scoring model is to use such coding—analogous to time in latent growth models—to rank a person's originality of responses.

For example, a person could have four responses, with the most original receiving the first rank (R1), the second-best receiving the second rank position (R2), and so forth. Here, the ranked originality can be coded as R1 = 0, R2 = 1, R3 = 2, and R4 = 3. This coding of ranked originality of the responses is then carried out for all participants who worked on a given task (e.g., the AUT). Notably, the number of ranks varies in the context of divergent thinking because of variation in fluency scores (e.g., some participants have four responses, whereas others may have ten). When based on the coded rank variable, a random-intercept-random-slope model is estimated with response originality as the dependent variable; this results in a random intercept that reflects a person's maximum score within a multidimensional model. When the

coding of the rank variable is changed to R1 = 0, R2 = 0, R3 = 1, and R4 = 2, the random intercept has a different meaning as it now reflects a person's Top2 Score. This principle of rank coding can be adjusted to reflect any number of top responses. Hence, we refer to this approach as a *multidimensional top-scoring method*.

The multidimensional top-scoring method includes a random intercept that reflects a person's score for their top responses (assessed via some aspect of creative quality; e.g., semantic distance). The random slope in this model can be understood as a person's *unevenness*. The slope will be negative for all people because of the coded ranking variable (a person that creates several responses with exactly the same originality score is very unlikely), and a more shallow slope implies that response originality is more even within a person's pool of responses. The notion of unevenness here is in accordance with conceptualizations of intra-individual variability in creativity research (for an overview see Barbot, 2022). Specifically, Barbot (2022) distinguished processing fluctuations which refer to inter-trial variability within the same task and dispersion as unevenness across tasks. The slope here represents a mix of both because the model is employed to several tasks at the same time.

Multidimensional top-scoring can be further extended for the current context into a model that includes varying item difficulties and item-specific deviations from an average slope parameter across all items. Specifically, using similar notation as De Boeck et al. (2011), the multidimensional top-scoring model used in this work can be written as

$$Y_{pir} = \theta_p + \sum_{k=1}^{K} \beta_i X_{ik} + \sum_{k=1}^{K} \delta_i X_{ik} R_{pr} + \gamma_p R_{pr},$$

with Y_{pir} being the originality score of the response with rank r on item i and person p, θ_p referring to the (latent) top score of person p (i.e., the random intercept), β_i referring to item difficulties, X_{ik} as a binary item indicator ($X_{ik} = 1$ when $i = k$ and $X_{ik} = 0$ otherwise; index k has the same range as i), δ_i as item-specific unevenness parameters, R_r referring to the coded rank variable for the responses of person p (as described above), and γ_p as the person unevenness parameter (i.e., the random part of the slope). For pragmatic reasons, we model the dependent variable Y_{pir} as Gaussian in this work, but the approach can be generalized to other distributional

families in a straightforward manner. The random intercept – i.e., θ_p – in these extended models will reflect a person's top score across all items, and the random slope – i.e., γ_p – will reflect a person's unevenness across all items. These random effects can be assumed to follow a multivariate normal distribution with a zero mean-vector (for model identification purposes) and covariance matrix $\begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\gamma} \\ \sigma_{\theta\gamma} & \sigma_\gamma^2 \end{pmatrix}$. Any linear mixed model software package can estimate the multidimensional top top-scoring model.

The Present Research

This work aims to apply the multidimensional top-scoring method and its effect on the reliability of ability estimates. Hence, we compared the measurement precision of multidimensional Maximum Scoring and multidimensional Top2 Scoring with their counterparts that do not take auxiliary information into account. Reliability of average scoring is also assessed in comparison to multidimensional top scoring. As a more practical examination of ability estimates, we further wanted to compare scoring methods with respect to their factor determinacy index (FDI; Ferrando & Lorenzo-Seva, 2018). Commonly, values of the FDI > .80 are considered for estimates to be used for research purposes, whereas an FDI > .90 implies excellent quality that allows using estimates in more practical individual differences contexts (e.g., high-stakes decisions). The FDI represents the correlation of ability estimates with their true values and can be obtained by taking the square-root from marginal reliability (Brown & Croudace, 2015; Ferrando & Lorenzo-Seva, 2018). Finally, we assessed the influence of fluency on conditional reliability of ability estimates and the influence of the used scoring approach on validity findings. All these analysis steps were applied to two different datasets to test the robustness of the findings. The datasets differ in terms of scoring of originality (Study 1: semantic distance; Study 2: human ratings), time on task (Study 1: 30 seconds; Study 2: 3 minutes), and the way to derive maximum and top scores (Study 1: maximum and top scores based on statistics; Study 2: top scores chosen by participants as well as maximum and top scores based on statistics).

Study 1

Method

Participants

In this work, we re-analyzed openly available data from a study by Beaty et al. (2022). The data are available here: https://osf.io/96zge/. We used $N = 149$ participants from that study, who were recruited from Penn State University (PSU; 67.11% women, mean age = 19.31 years, $SD = 1.79$). All study participants could exchange their received credit for a research option in a psychology course. Informed consent was obtained from all participants, and the study received ethical approval from the PSU IRB.

Procedure

Participants were instructed to participate in the online study (based on the platform Pavlovia) in a quiet room with minimal distractions. Participants completed a series of tests, including cognitive ability and personality measures.

Divergent Thinking Assessment

Divergent thinking was assessed by means of the AUT, including 13 items (Beaty et al., 2022). The purpose of the initial study by Beaty et al. (2022) was to identify maximally reliable and valid AUT items (objects) for semantic distanced-based assessment of originality, hence many different AUT items were included in the study. Participants were instructed to "be creative" while working on the task, with the following instructions: *Come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different*. For each task, they were given 30 seconds to generate creative uses. The order of AUT trials was randomized for each participant.

In this work, responses were automatically scored for originality using the open *SemDis* platform (http://semdis.wlu.psu.edu/). *SemDis* leverages distributional semantic models to compute the cosine similarity between two texts, yielding semantic distance values (1 – cosine similarity) that have been shown to correlate with human judgements of creativity (Beaty & Johnson, 2021; Heinen & Johnson, 2018). The application of semantic distance for creativity

research is rooted in the associative theory of creativity—the notion that creativity involves making conceptually distant associations (Kenett, 2019; Mednick, 1962)—with semantically "close" associations typically yielding low semantic distance values (e.g., *brick-house*) and conceptually "far" associations yielding larger distance values.

Semantic distance was computed between the AUT item/object (e.g., brick) and participant responses (Beaty & Johnson, 2021), using the following text preprocessing and semantic model settings on the *SemDis* platform: cleaning type = remove filler and clean (removes "stop words", such as the, an, and a, from a response; Forthmann et al., 2019), semantic space = all (includes all five semantic models available on the platform; Beaty & Johnson, 2021), and compositional model = multiplicative (combines multiple word vectors via multiplication; combining word vectors is necessary for multi-word AUT responses; Beaty & Johnson, 2021; Dumas et al., 2020). The average semantic distance value for the five models (SemDis_mean) was used for analysis.

External Validity Measures

For a check of validity based on external measures, we measured Openness to Experience by means of the NEO FFI Openness (12 items; McCrae & Costa, Jr., 2007), creative self-beliefs by means of the Short Scale of Creative Self (creative self-efficacy and creative personal identity; Karwowski et al., 2018), and engagement in creative behaviors by means of Biographical Inventory of Creative Behaviors (Batey, 2007; Silvia et al., 2021).

Data Analysis

We used the statistical software R (R Core Team, 2021) and the glmmTMB package (Brooks E. et al., 2017). Prior to model fitting, the semantic distance scores were grand mean *z*-standardized. We used the grand mean for standardization (and not the item-specific means) to keep meaningful item difficulty estimates in the models. Simple scoring models for Maximum, Top2, and Average Scoring were models with item difficulty and a random intercept that mimics simple averaging or summing across all 13 AUT objects (these models are referred to as Only Maximum, Only Top2, or Average hereafter). Then, we fitted the multidimensional

scoring models (see Introduction) for maximum scoring and Top2 scoring, which we denote by All Maximum and All Top2, respectively. R code to fit all these models is openly available in the Open Science Framework (https://osf.io/7rgsp/).

Marginal reliability was estimated based on the estimated trait variance and the average of the variances of the trait scores (i.e., the squared standard errors of the trait scores) by the following formula (Brown & Croudace, 2015):

$$Rel(\theta) = 1 - \overline{SE}_{\theta}^{2}/\hat{\sigma}_{\theta}^{2}.$$

We also examined the square-root of marginal reliability, which reflects the correlation between the trait estimates and their true values (i.e., the FDI). In addition, we estimated reliability conditional on each participant's estimated trait level (i.e, θ_j refers to the trait level of person *j*) based on a similar approach:

$$Rel(\theta_j) = 1 - SE_{\theta_j}^{2}/\hat{\sigma}_{\theta}^{2}.$$

To focus on the potential advantages of the multidimensional scoring approach, an index of relative efficiency (RE) was obtained conditional on the estimated trait level:

$$RE(\theta_{j,\text{Only}}, \theta_{j,\text{All}}) = SE_{\theta_{j,\text{Only}}}^{2}/SE_{\theta_{j,\text{All}}}^{2},$$

with $\theta_{j,\text{Only}}$ and $\theta_{j,\text{All}}$ referring to the trait level based on the model in which collateral information is ignored (Only) or when collateral information is used in multidimensional scoring (All). RE values > 1 indicate that the multidimensional scoring increases measurement precision as compared to the scoring based on the Only models. REs were calculated for both scorings maximum scoring and Top2 scoring. We further calculated average REs (ARE) as a summary statistic.

Study 1 – Results and Discussion

Reliability, FDI, and Relative Efficiency

Reliability results were as expected for the models that did not use auxiliary information (see Table 1 and Figure 1). Average scoring and Top2 scoring had comparable marginal reliability (.82; see Table 1) in this case, whereas marginal reliability was somewhat lower for maximum scoring (.77). This observation was further corroborated by inspecting the

distributions of conditional reliability estimates in Figure 1. All FDI values for these three models indicated that ability estimates were of sufficient quality for research purposes. In addition, average scoring and top-scoring FDI values were scarcely higher as compared to the .90 cut-off for practical applications (e.g., high-stakes decisions). For maximum scoring, the FDI did not pass the .90 cut-off (see Table 1).

However, the highest marginal reliability findings were obtained for the multidimensional top-scoring models (see All Top2 and All Maximum in Table 1 and Figure 1). Reliability clearly increased beyond the marginal reliability of average scoring, when multidimensional top scoring was used. In addition, average RE indicated that multidimensional top-scoring mostly dominated scoring without auxiliary information (Maximum Scoring: ARE = 1.59; Top2 Scoring: ARE = 1.66). Examining all REs, it was evident that measurement precision was better for multidimensional top-scoring for almost all cases (see the dashed vertical line in Figure 2). Finally, efficiency results highlighted that Top2 Scoring benefitted slightly more from the multidimensional scoring approach than Maximum Scoring.

Dependence of Reliability on Fluency

Based on theoretical deliberations and recent empirical findings, it was expected that the reliability of average scoring depends heavily on a person's number of generated responses. This expectation was confirmed: Figure 3 shows that the correlation between conditional reliability and fluency was strongest for Average Scoring (r = .84), but decreases for Top2 Scoring (r = .66) and Maximum Scoring (r = .42) when auxiliary information is ignored. This indicates that reliability for Top2 and Maximum Scoring is still a positive function of fluency. However, when multidimensional scoring is used, this relationship between conditional reliability and fluency decreases to a value of $r \approx .15$ (see Figure 3).

Validity

Finally, we assessed how validity findings might be affected by using naïve observed scores—derived from simple models without auxiliary information—and multidimensional top

scoring. We computed correlations between all these scores and the three personality measures (see the bottom three rows in Table 1).[2]

We found that both variants of average scoring did not correlate with any of the personality measures, even to a small degree. Naïve Maximum and Top2 Scoring yielded highly comparable validity evidence in relation to scores based on simple models without auxiliary information. Validity findings for multidimensional Top2 were also on par with the other findings, yet for Openness the highest correlation with Top2 scoring was found for Only Top2. For Maximum Scoring, however, the highest validity coefficients were obtained for multidimensional scoring. Notably, the effect sizes were generally small, albeit consistent with zero-order correlations between creativity measures reported in prior work (e.g., McCrae, 1987).

Study 2

Method

Participants

In Study 2, we reanalyzed Silvia et al.'s (2008; Silvia, 2008) classical dataset. Their final sample comprised of $N = 226$ participants (178 were female and 48 were male; the average age was $M = 19.20$, $SD = 3.14$). For more details than reported here on the participants, procedure, and measures we refer to the original papers (Silvia, 2008; Silvia et al., 2008).

Procedure

Participants participated in 90-minute sessions. Measures were administered in the following order: a) divergent thinking, b) fluid reasoning, c) verbal fluency, d) strategy generation, and e) openness to experience.

Divergent Thinking Assessment

[2] Notably, we derived ability estimates from latent variable models and used them subsequently in this correlational analysis. Hence, the correlations examined here are not corrected for the imperfect reliability of the measures as it is the case for comprehensive latent variable approaches (Wang et al., 2004). Nonetheless, these analyses serve the purpose of this work, which is to examine if the quality of divergent thinking scores is sufficient for subsequent usage in research and practice.

We reanalyzed both AUTs (*brick* and *knife*) administered by Silvia et al. (2008) with a time limit of three minutes to work on the task. Participants were instructed to "be creative" while working on the task, with the following instructions:

> *For this task, you should write down all of the original and creative uses for a brick that you can think of. Certainly there are common, unoriginal ways to use a brick; for this task, write down all of the unusual, creative, and uncommon uses you can think of. You'll have three minutes. Any questions?*

Three raters provided ratings (a 5-point Likert scale was used) based on a coding scheme based on uncommonness, remoteness, and cleverness as three classical indicators of originality (see Wilson et al., 1953). An absolute agreement intra-class correlation for the average scores indicated fair inter-rater reliability (cf. Cicchetti, 2001), ICC(2, 3) = .43, 95%-CI: [.10, .62].

Fluid Reasoning

Fluid reasoning measures are indicators of fluid intelligence (e.g., Carroll, 1993). The fluid reasoning composite here was based on scores of the Raven's Progressive Matrices (18 items; 12 minutes), the Letter Sets task (16 items, 4 minutes), and the Paper Folding task (10 items, 4 minutes; Ekstrom et al., 1976). We averaged across *z*-standardized scores, and Cronbach's α was .63.

Verbal Fluency

Verbal fluency tasks are indicators of broad retrieval ability (Forthmann et al., 2019; Silvia et al., 2013). Here two letter fluency (list words that begin with *letter f* and *letter m*) and two semantic fluency (*animals* and *occupations*) tasks were assessed (cf. Unsworth et al., 2010). Participants were allowed to work on each task for two minutes. We averaged across *z*-standardized scores, and Cronbach's α was .76.

Strategy Generation

Three tasks measured the ability to generate successful strategies for verbal fluency tasks (Philipps, 1999). In these tasks, participants had to list strategies that could be useful for

the generation of responses for a given verbal fluency task (e.g., list parts of the body). Participants had to generate strategies for three different verbal fluency tasks (*parts of the body*, *examples of food*, and *countries*). We averaged across *z*-standardized scores, and Cronbach's α was .65.

Openness to Experience

We used a composite score of three Openness indicators. The composite was based on *z*-standardized scores of the FFI Openness (12 items; Costa & McCrae, 1992), Scale based on the International Personality Item Pool (10 items; Goldberg et al., 2006), and a brief Big 5 scale (2 items; Gosling et al., 2003). We averaged across scores, and Cronbach's α for the composite was .73.

Data Analysis

All data analysis steps were carried out as in Study 1. However, to adapt the approach to Chosen Top2 two steps had to be different. First, for the top responses chosen by participants the intercept of the linear model was specified to reflect the average ratings for the chosen responses. Second, for semantic distance scores it was unlikely that a person had the exact same score for a response. For average ratings, however, the same scores for different responses appeared quite often. Hence, we used a different method to deal with ties in the ranking of responses. Specifically, we used the random method (i.e., we set the `ties.method` argument in the `rank()` function to `"random"`) for the ranking of participants' responses. Again, all R code is openly available in the Open Science Framework (https://osf.io/7rgsp/).

Study 2 – Results and Discussion

Reliability, FDI, and Relative Efficiency

Reliability results extend and replicate the findings from Study 1 (see Table 2). All multidimensional top scoring approaches displayed much higher marginal reliability (range from .92 to .94) than average scoring (.72). This difference in reliability between multidimensional top scoring and average scoring was clearly stronger as compared to Study 1. Furthermore, differences in marginal reliability between multidimensional top scoring and

scoring without auxiliary information (range from .40 to .47) were much stronger (see Table 2). All these observations were further corroborated by inspecting the distributions of conditional reliability estimates in Figure 1 (see right side). The FDI values for Only Chosen Top2 and Only Maximum indicated that ability estimates were not of sufficient quality for research purposes, whereas all other FDI values were larger than .80. In addition, all multidimensional top-scoring FDI values were clearly higher as compared to the .90 cut-off for practical applications (e.g., high-stakes decisions). Yet, this was not the case for average scoring and Only Statistical Top2 (see Table 2).

In addition, average REs were much higher as compared to Study 1 and indicated that multidimensional top-scoring dominated scoring without auxiliary information (Chosen Top2 Scoring: ARE = 4.09; Statistical Top2 Scoring: ARE = 5.04; Statistical Maximum Scoring: ARE = 7.11). Examining all REs, it was evident that measurement precision was better (or equally precise) for multidimensional top-scoring for all cases (see the dashed vertical line in Figure 2 on the right side). Finally, detailed efficiency results highlighted that Chosen Top2 Scoring benefitted for many more cases from the multidimensional scoring approach than Statistical Top Scoring (i.e., the average REs above were influenced by outliers as indicated by the boxplots on the right side in Figure 2). Figure 2 further revealed that Statistical Maximum Scoring benefitted more from multidimensional scoring than Statistical Top2 (this was the other way around in Study 1).

Dependence of Reliability on Fluency

Again, the correlation between conditional reliability and fluency was strongest for Average Scoring (r = .92). The absolute size of the correlation coefficient was equally high or larger for the respective scorings without auxiliary information as compared to its multidimensional counterpart: Chosen Top2 Scoring (r = .03 vs. r = -.03), Statistical Top2 (r = .04 vs. r = .04), and Statistical Maximum Scoring (r = .06 vs. r = -.09). Overall, this shows that the dependence of reliability on fluency was less an issue for all top scoring variants as compared to Study 1 (cf. Figure 3).

Validity

As in Study 1, we assessed how validity findings might be affected by using observed scores, scores from simple models without auxiliary information, and scores based on multidimensional top scoring. We computed correlations between all these scores and fluid reasoning, verbal fluency, strategy generation, and openness (see the bottom rows in Table 2).

Similar to Study 1, we found that average scoring correlated mostly less strong with the validity measures as compared to the top scoring approaches. This was particularly true for the multidimensional top scoring approaches, which provided mostly the highest correlations with validity measures. The only validity measure with a slight difference as compared to the overall pattern was verbal fluency (see Table 2). Here average scoring correlated higher than any of the top scoring approaches, yet when looking at each of the different top scoring approaches (i.e., Chosen Top2, Statistical Top2, and Statistical Maximum), the multidimensional variant displayed either the highest coefficient or was at least on par with the other coefficients, respectively.

General Discussion

Researchers have argued that for practical purposes, it is of utmost importance to focus on the best ideas when assessing divergent thinking (Girotra et al., 2010). However, the measurement of the best responses comes along with challenges because less information is used for scoring (compared to the average quality of all responses; Benedek et al., 2013; Silvia, 2011). In the current work, we addressed this issue by introducing a multidimensional scoring framework that demonstrated increasing measurement precision in the context of divergent thinking assessment. Across two studies we found that Maximum and Top2 Scoring can surpass Average Scoring in measurement precision when the remaining responses (the “bad” ideas), and their originality scores, are used as auxiliary information by multidimensional top scoring. This advantage was boosted by fluency as indicated by a much stronger reliability boost in Study 2 (average fluency was 7.13 and ranged from 2.00 to 20.00) as compared to Study 1 (average

fluency was 2.45 and ranged from 1.00 to 6.23). Yet, still with not much additional information multidimensional scoring increased measurement precision (Study 1).

Moreover, for Maximum Scoring in Study 1, a clear qualitative leap was observed when auxiliary information was used. Without auxiliary information, Maximum Scoring did not pass the FDI cut-off of .90 (Ferrando & Lorenzo-Seva, 2018), which signals that scores can be used in individual assessment contexts (e.g., high-stakes decisions), but including all responses within the multidimensional top-scoring approach yielded a Maximum Scoring with an FDI > .90. Study 2 further corroborated these findings. For Chosen Top2 and Maximum Scoring not even the .80 cut-off for research purposes was passed and the FDI for Statistical Top2 Scoring was also below the .90 criterion. Yet, for all these scorings multidimensional scoring lifted FDI with all values passing the .90 cut-off.

Furthermore, we found in Study 1 that multidimensional top-scoring substantially weakens the well-known (and problematic) relationship between reliability and fluency. Yet, in Study 2 the difference between multidimensional top scoring and scoring based on the top responses only was only minimally present. This difference in findings across Study 1 and Study 2 could most likely be explained again by average fluency. In Study 1, average fluency was close to two response per person and task and many persons had tasks for which only one response was available. This increased the reliability-fluency correlation for the scorings in Study 1. On the contrary, in Study 2 all participants had at least two responses on each of the items. Consequently, the dependence of reliability on fluency was less of an issue for this dataset. Finally, the strongest validity findings were mostly obtained for ability scores based on the multidimensional top-scoring approach. Only in very few cases a validity correlation was negligible higher (by .01) than the correlation obtained for the multidimensional score.

Importantly, the proof of concept provided by the empirical findings in this work can be considered quite strong. First, divergent thinking assessment in Study 1 relied on very short administration times (30 seconds) and, hence, not much information was available beyond the Maximum or Top2 responses. Hence, the clear increase in marginal reliability and FDI, the

advantage in terms of efficiency, the decrease of dependence on the amount of information, and the slight validity advantage were all observed under conditions that are unlikely to reveal the full potential of multidimensional top scoring. Second, Study 2 replicated and extended the findings obtained from Study 1. In Study 2 time-on-task was much longer (3 minutes) which resulted in higher levels of fluency, and much more auxiliary information available for multidimensional scoring, the advantages were clearly more readily detectable. Only the findings related to the dependence of conditional reliability on fluency were specific to Study 1 (as discussed in detail above).

Notably, the validity gain from multidimensional top-scoring was admittedly modest and would not have passed any examination of statistical significance. In relation to this, it should be considered that using ability estimates from such latent variable models does not correct for any attenuation because of imperfect measures (Wang et al., 2004). Hence, the absolute size of the correlation will be higher when a full latent variable approach is used (see also Beaty et al., 2022). Nonetheless, correlations were mostly highest or on par when ability estimates were derived from multidimensional top-scoring models. This is promising in our view. Relatedly, it is yet unknown how the multidimensional top-scoring approach can be implemented in more comprehensive latent variable approaches. Examining the applicability of the approach with other software, and studying the approach in simulation studies, are promising paths for future research. Currently, we recommend using multidimensional top-scoring for situations in which divergent thinking scores are needed for subsequent analysis (e.g., as a dependent variable) or when such scores are needed for practical purposes.

Importantly, the decision to use either average or top scoring should be primarily be guided by theoretical deliberations (e.g., the dual pathway theory of creativity focuses on average originality; Nijstad et al., 2010). The availability of multidimensional top scoring will make such decisions easier because choosing top scoring is not associated with a loss in measurement precision anymore. Instead, multidimensional top scoring boosts reliability as compared to average scoring. Furthermore, choosing one’s best responses requires evaluative

skill and researchers have cautioned that this can lead to a mix of constructs that is being measured (e.g., Runco, 2008). Meta-analytical evidence on the relationship between divergent thinking and evaluative skill seems to emphasize this view as both constructs were found to have a small correlation (r = .13; Guo et al., 2022). However, it can also be argued that such a mix of constructs generalizes better to models of the creative process (e.g., Mumford & McIntosh, 2017). Creative processes require more skills (e.g., problem definition, conceptual combination, idea evaluation, and so forth) than divergent thinking. Thus, a measure that requires both idea generation and evaluation of ideas maps better onto creative process models. Again, we argue that the question, if participants choose their top responses or if top responses are identified on statistical criteria should also be decided based on theoretical deliberations. Our contribution to this debate is that we demonstrated that multidimensional top scoring leads to a boost of measurement precision regardless if top responses were chosen by participants or based on statistical criteria.

Conclusion

Divergent thinking tasks can be scored in various ways, and psychometric issues at the level of responses, tasks, and full test batteries affect their overall psychometric quality. In this work, we proposed a multidimensional top-scoring approach that relies on the information provided by all responses of participants. That is, we have shown that measurement precision of scoring the complete test can be strongly increased when information from every single response is used as auxiliary information. This way, it was possible to invert patterns of results from previous works that found the reliability of top scoring to be inferior compared to average scoring. With multidimensional top-scoring, the reliability can even surpass the reliability of average scoring. Hence, we recommend using this approach whenever top-scoring represents a reasonable scoring approach in research and/or practice, and we provide open-access code for researchers to implement multidimensional top-scoring in their own research.

References

Barbot, B. (2022). Intra-individual variability in creativity: Nature, measurement, and prospects. European Psychologist. *European Psychologist*. Advance online publication. https://doi.org/10.1027/1016-9040/a000470

Batey, M. (2007). *A psychometric investigation of everyday creativity*. University of Londong.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic Distance And the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Creativity Research Journal*, 1–16. https://doi.org/10.1080/10400419.2022.2025720

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 341–349. https://doi.org/10.1037/a0033644

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The Role of Coding Time in Estimating and Interpreting Growth Curve Models. *Psychological Methods*, *9*(1), 30–52. https://doi.org/10.1037/1082-989X.9.1.30

Bourgeois-Bougrine, S. (2020). What Does Creativity Mean in Safety-Critical Environments?. *Frontiers in Psychology*, 2518. https://doi.org/10.3389/fpsyg.2020.565884

Brooks E., M., Kristensen, K., Benthem J.,van, K., Magnusson, A., Berg W., C., Nielsen, A., Skaug J., H., Mächler, M., & Bolker M., B. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378. https://doi.org/10.32614/RJ-2017-066

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate*

applications series. Handbook of item response theory modeling: Applications to typical performance assessment (pp. 307–333). Routledge/Taylor & Francis Group.

Bulut, O. (2013). *Between-person and Within-person Subscore Reliability: Comparison of Unidimensional and Multidimensional IRT Models*. University of Minnesota.

Cronbach, L. J. (1941). The Reliability of Ratio Scores. *Educational and Psychological Measurement*, *1*(1), 269–277. https://doi.org/10.1177/001316444100100121

Deacon, S. A. (2000). Using divergent thinking exercises within supervision to enhance therapist creativity. *Journal of Family Psychotherapy, 11*(2), 67–73. https://doi.org/10.1300/J085v11n02_06

de la Torre, J., Song, H., & Hong, Y. (2011). A Comparison of Four Methods of IRT Subscoring. *Applied Psychological Measurement*, *35*(4), 296–316. https://doi.org/10.1177/0146621610378653

Dumas, D., Dong, Y., Grajzel, K., Forthmann, B., & Doherty, M. (2022). Understanding ideational fluency as a survival process. *British Journal of Educational Psychology*, *92*(2), e12469. https://doi.org/10.1111/bjep.12469

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000319

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and Psychological Measurement*, *78*(5), 762–780. https://doi.org/10.1177/0013164417719308

Finke, R. A., Smith, S. M., & Ward, T. B. (1996). *Creative Cognition*. The MIT Press. https://doi.org/10.7551/mitpress/7722.001.0001

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*(1), 37–55. https://doi.org/10.1037/0022-0663.90.1.37

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing Speed as a Confounding Variable and the Measurement of Quality in Divergent Thinking. *Creativity Research Journal*, *29*(3). https://doi.org/10.1080/10400419.2017.1360059

Forthmann, B., Jankowska, D. M., & Karwowski, M. (2021). How reliable and valid are frequency-based originality scores? Evidence from a sample of children and adolescents. *Thinking Skills and Creativity*, *41*, 100851. https://doi.org/10.1016/j.tsc.2021.100851

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*(4), 559–575. https://doi.org/10.1002/jocb.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*(3), 683–699. https://doi.org/10.1111/bjep.12325

Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea Generation and the Quality of the Best Idea. *Management Science*, *56*(4), 591–605. https://doi.org/10.1287/mnsc.1090.1144

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

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*. Advance online publication. https://doi.org/10.1002/jocb.539

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 144–156. https://doi.org/10.1037/aca0000125

Hocevar, D., & Michael, W. B. (1979). The effects of scoring formulas on the discriminant validity of tests of divergent thinking. *Educational and Psychological Measurement*, *39*(4), 917-921. https://doi.org/10.1177/001316447903900427

Karwowski, M., Lebuda, I., & Wiśniewska, E. (2018). Measuring creative self-efficacy and creative personal identity. *The International Journal of Creativity & Problem Solving,*

28(1), 45–57.

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Lubart, T. (2001). Models of the Creative Process: Past, Present and Future. *Creativity Research Journal*, *13*(3–4), 295–308. https://doi.org/10.1207/S15326934CRJ1334_07

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*(6), 1258–1265. https://doi.org/10.1037/0022-3514.52.6.1258

McCrae, R. R., & Costa, Jr., P. T. (2007). Brief Versions of the NEO-PI-3. *Journal of Individual Differences*, *28*(3), 116–128. https://doi.org/10.1027/1614-0001.28.3.116

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mumford, M. D., & McIntosh, T. (2017). Creative Thinking Processes: The Past and the Future. *The Journal of Creative Behavior*, *51*(4), 317–322. https://doi.org/10.1002/jocb.197

Nijstad, B. A., De Dreu, C. K., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, *21*(1), 34-77. https://doi.org/10.1080/10463281003765323

Plucker, J. A. (1999). Is the Proof in the Pudding? Reanalyses of Torrance's (1958 to Present) Longitudinal Data. *Creativity Research Journal*, *12*(2), 103–114. https://doi.org/10.1207/s15326934crj1202_3

Preacher, K., Wichman, A., MacCallum, R., & Briggs, N. (2008). *Latent Growth Curve Modeling*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984737

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.1.2). R Foundation for Statistical Computing. https://www.r-project.org/

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*,

41(3), 647–656. https://doi.org/10.3758/BRM.41.3.647

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

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

Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 93–96. https://doi.org/10.1037/1931-3896.2.2.93

Runco, M. A., & Acar, S. (2012). Divergent Thinking as an Indicator of Creative Potential. *Creativity Research Journal*, *24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment*, *5*(2), 149-156. https://doi.org/10.1177/073428298700500206

Shaw, A. (2021). It works…but can we make it easier? A comparison of three subjective scoring indexes in the assessment of divergent thinking. *Thinking Skills and Creativity*, *40*, 100789. https://doi.org/10.1016/j.tsc.2021.100789

Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual differences*, *44*(4), 1012-1021. https://doi.org/10.1016/j.paid.2007.10.027

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, *6*(1), 24–30. https://doi.org/10.1016/j.tsc.2010.06.001

Silvia, P. J., Rodriguez, R. M., Beaty, R. E., Frith, E., Kaufman, J. C., Loprinzi, P., & Reiter-Palmon, R. (2021). Measuring everyday creativity: A Rasch model analysis of the

Biographical Inventory of Creative Behaviors (BICB) scale. *Thinking Skills and Creativity*, *39*, 100797. https://doi.org/10.1016/j.tsc.2021.100797

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. 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*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

van der Linden, W. J. (2010). *Linear models for optimal test design*. Springer.

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327-347. https://doi.org/10.1177/0146621609349800

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*, *9*(1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116

Zaccaro, S. J., Connelly, S., Repchick, K. M., Daza, A. I., Young, M. C., Kilcullen, R. N., ... & Bartholomew, L. N. (2015). The influence of higher order cognitive capacities on leader organizational continuance and retention: The mediating role of developmental experiences. *The Leadership Quarterly*, *26*(3), 342-358. https://doi.org/10.1016/j.leaqua.2015.03.007

Table 1

Study 1 – Reliability, FDI, and Validity Findings

		Marginal Reliability	FDI	1	2	3	4	5	6	7	8	9	10
Average – observed	1	-	-										
Average	2	.82	.91	.98									
Top2 – observed	3	-	-	.88	.85								
All Top2	4	.89	.94	.88	.88	.98							
Only Top2	5	.82	.90	.87	.86	.99	.99						
Maximum – observed	6	-	-	.76	.73	.95	.91	.93					
All Maximum	7	.88	.94	.84	.83	.98	.99	.99	.95				
Only Maximum	8	.77	.88	.76	.73	.95	.91	.94	1.00	.96			
Creative Self-Concept	9	-	-	.07	.05	.15	.16	.16	.15	.17	.15		
BICB	10	-	-	.02	.03	.06	.08	.07	.07	.09	.07	.36	
Openness	11	-	-	.02	.03	.09	.09	.10	.09	.12	.09	.50	.35

Notes. BICB = Biographical Inventory of Creative Behaviors. FDI = Factor determinacy index.

Table 2

Study 2 – Reliability, FDI, and Validity Findings

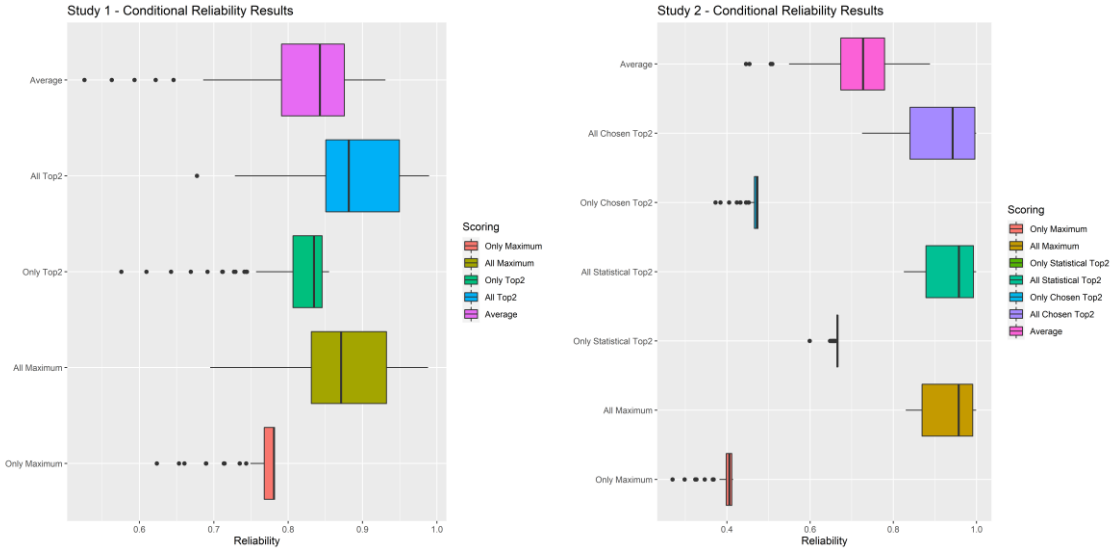
		Marginal Reliability	FDI	1	2	3	4	5	6	7	8	9	10	11	12	14	14
Average – observed	1	-	-														
Average	2	.72	.85	.98													
Chosen Top2 – observed	3	-	-	.82	.82												
All Chosen Top2	4	.92	.96	.88	.90	.89											
Only Chosen Top2	5	.47	.68	.81	.82	1.00	.89										
Statistical Top2 – observed	6			.84	.85	.83	.95	.83									
All Statistical Top2	7	.94	.97	.87	.89	.85	.99	.85	.97								
Only Statistical Top2	8	.66	.82	.84	.85	.84	.96	.84	.99	.98							
Maximum – observed	9	-	-	.79	.81	.79	.89	.79	.92	.91	.96						
All Maximum	10	.93	.97	.89	.91	.86	.97	.86	.97	.99	.98	.92					
Only Maximum	11	.40	.63	.80	.81	.80	.89	.80	.92	.91	.96	1.00	.92				
Fluid Reasoning	12	-	-	.19	.20	.22	.25	.22	.22	.24	.22	.20	.23	.20			
Verbal Fluency	13	-	-	.19	.19	.12	.15	.11	.12	.14	.11	.09	.13	.10	.13		

Strategy Generation	14	-	-	.16	.16	.12	.24	.12	.22	.24	.24	.23	.22	.22	.21	.25	
Openness	15	-	-	.05	.07	.06	.14	.07	.11	.14	.11	.12	.12	.11	.07	.16	.31

Notes. FDI = Factor determinacy index.

Figure 1

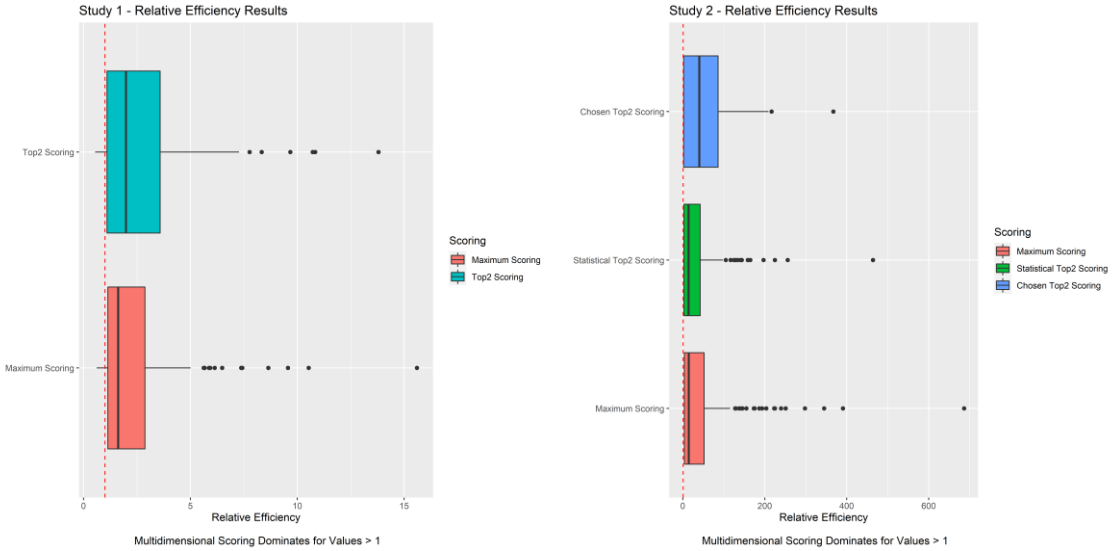
Conditional Reliability as a Function of Scoring



Notes. Left side: Study 1 results. Right side: Study 2 results. Average = Ability Scoring is based on a model including item difficulty and a random intercept across persons. Only (Statistical) Top2 = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only statistically determined Top2 responses for each item enter the model). All (Statistical) Top2 = Ability Scoring is based on multidimensional Top Scoring (i.e., responses not among the statistically determined Top2 serve as auxiliary information). Only Chosen Top2 = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only Top2 responses chosen by participants for each item enter the model). All (Statistical) Top2 = Ability Scoring is based on multidimensional Top Scoring (i.e., responses not among the Top2 chosen by participants serve as auxiliary information). Only Maximum = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only the best responses for each item enter the model). All Maximum = Ability Scoring is based on multidimensional Maximum Scoring (i.e., responses not among the best ones serve as auxiliary information).

Figure 2

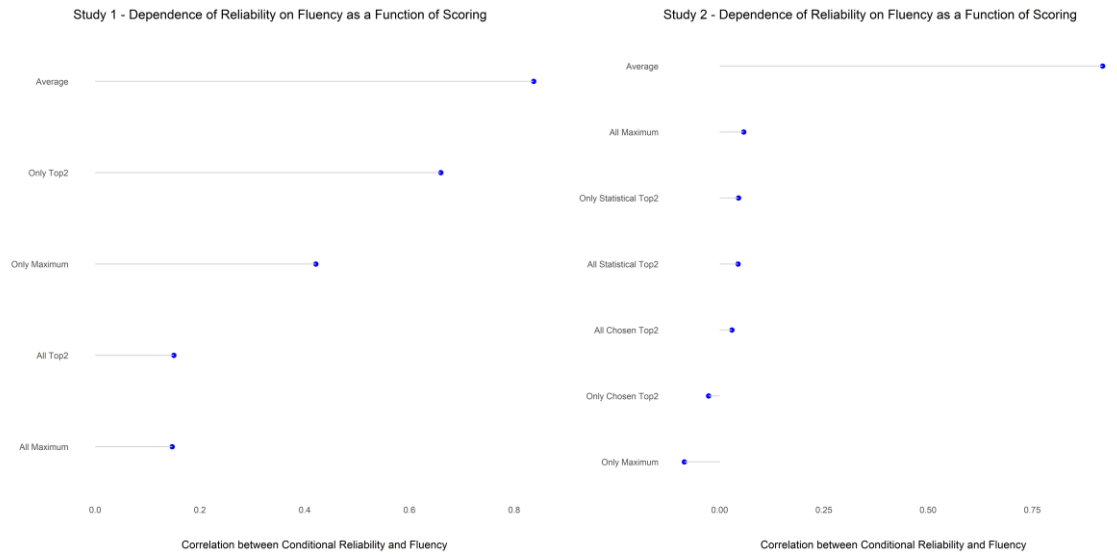
Relative Efficiency for Maximum Scoring and Top2 Scoring



Notes. Left side: Study 1 results. Right side: Study 2 results. Multidimensional Top-Scoring models (i.e., All Maximum, All (Statistical) Top2, and All Chosen Top2 models) are contrasted with their counterparts that do not rely on available auxiliary information (i.e., Only Maximum, Only (Statistical) Top2 models, and Only Chosen Top2 models). The multidimensional Scoring yields higher measurement precision when relative efficiency is > 1.

Figure 3

Dependence of Reliability on Fluency as a Function of Scoring



Left side: Study 1 results. Right side: Study 2 results. Average = Ability Scoring is based on a model including item difficulty and a random intercept across persons. Only (Statistical) Top2 = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only statistically determined Top2 responses for each item enter the model). All (Statistical) Top2 = Ability Scoring is based on multidimensional Top Scoring (i.e., responses not among the statistically determined Top2 serve as auxiliary information). Only Chosen Top2 = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only Top2 responses chosen by participants for each item enter the model). All (Statistical) Top2 = Ability Scoring is based on multidimensional Top Scoring (i.e., responses not among the Top2 chosen by participants serve as auxiliary information). Only Maximum = Ability Scoring is based on a model including item difficulty, a random intercept across persons, and auxiliary information is ignored (i.e., only the best responses for each item enter the model). All Maximum = Ability Scoring is based on multidimensional Maximum Scoring (i.e., responses not among the best ones serve as auxiliary information).