Planning Missing Data Designs for Human Ratings in Creativity Research: A Practical Guide

Boris Forthmann[1], Benjamin Goecke[2], & Roger E. Beaty[3]

[1]Institute of Psychology, University of Münster

[2]Institute for Psychology and Pedagogy, Ulm University

[3]Department of Psychology, Pennsylvania State University

Author Note

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

Benjamin Goecke https://orcid.org/0000-0002-3050-1848

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

All files for analyses are available at Open Science Framework: https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70.

R.E.B. is supported by grants from the National Science Foundation [DRL-1920653; DUE- 2155070].

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

Abstract

Human ratings are ubiquitous in creativity research. Yet the process of rating responses to creativity tasks—typically several hundred or thousands of responses, per rater—is often time consuming and expensive. Planned missing data designs, where raters only rate a subset of the total number of responses, have been recently proposed as one possible solution to decrease overall rating time and monetary costs. However, researchers also need ratings that adhere to psychometric standards, such as a certain degree of reliability, and psychometric work with planned missing designs is currently lacking in the literature. In this work, we introduce how judge response theory and simulations can be used to fine-tune planning of missing data designs. We provide open code for the community and illustrate our proposed approach by a cost-effectiveness calculation based on a realistic example. We clearly show that fine tuning helps to save time (to perform the ratings) and monetary costs, while simultaneously targeting expected levels of reliability.

Keywords: creativity assessment; human ratings; judge response theory; planned missing data

Planning Missing Data Designs for Human Ratings in Creativity Research: A Practical Guide

Human ratings are ubiquitous in creativity research. Beginning with early work by Guilford, which required human judges for the scoring of creative thinking performance (P. R. Christensen et al., 1957; Wilson et al., 1953), to Amabile's Consensual Assessment Technique (CAT) for the assessment of creative products (Amabile, 1982), and more recently, to creative thinking assessment of divergent thinking (Benedek et al., 2013; Forthmann et al., 2017; Silvia et al., 2008), creative metaphor (Beaty & Silvia, 2013; Primi, 2014) or humor production (A. P. Christensen et al., 2018; Nusbaum et al., 2017), human ratings seemingly follow a long tradition of being indispensable for the assessment of creativity. In a sense, they are considered a gold standard in creativity research. This status of human ratings is further emphasized by the fact that most recent attempts of automated scoring of creative thinking aim at the prediction of human ratings (Beaty & Johnson, 2021; Buczak et al., 2022; Dumas et al., 2020; Stevenson et al., 2020). Automated scoring is motivated by the fact that human ratings are associated with two drawbacks: scoring might be affected by various idiosyncrasies of raters (Mouchiroud & Lubart, 2001; Robitzsch & Steinfeld, 2018), and scoring of creativity tasks by human raters does not always result in agreements between raters (Forthmann et al., 2017), but is very laborious and may even take weeks (Benedek et al., 2013; Shaw, 2021; Silvia et al., 2009).

The current work addresses the latter issue by demonstrating how Judge Response Theory (JRT; (Myszkowski, 2021; Myszkowski & Storme, 2019)—i.e., the application of item response modeling to human ratings in creativity research—and simulation techniques can be leveraged for effective distribution of rater work (i.e., by means of a planned missing data design), while at the same time a sufficient degree of measurement precision of the final scores is to be expected. Such a careful planning approach to rating designs can reduce the amount of coding work put on the shoulder of raters (preventing adverse effects of rater fatigue; e.g., (Forthmann et al., 2017), allows the overall coding task being finished in comparably less time (i.e., compared to having all products rated by all available raters), and provides an empirical rationale for saving valuable project money. We argue that such work will be highly useful for the field as long as automated scorings are not yet fully available for all creativity measures that typically require human ratings.

Human Ratings in Creativity Research

Roughly speaking, creativity refers to the novelty and usefulness of a perceptible product (Plucker et al., 2004). This notion of creativity as a property of perceptible products is important for the current work, because it is open to encompass rather small expressions of thought. From this perspective, products include responses generated in a divergent thinking task (Runco et al., 2001), a list of ideas obtained from a small-group brainstorming session (Reinig & Briggs, 2013), a written story (Kornilov et al., 2016; Taylor & Barbot, 2021), responses to a scientific creative thinking task (Long, 2014; Long & Pang, 2015), as well as drawings or a design for furniture. Indeed, all these kinds of products are rated by human judges in creativity research which highlights the broad range of usage contexts of human rater scores.

Beyond the product, it is important to look at the samples of raters used. For example, the famous consensual assessment technique (Amabile, 1982) is considered a valid measure of creativity because experts of the respective product domain assess a product's creativity (e.g., furniture designers rate the creativity of furniture designs). However, quasi-experts (e.g., design students which are not yet fully developed experts; (Kaufman et al., 2013) have also been sampled for providing creativity ratings as well as laypersons (also named novices or naïve raters; (Hass et al., 2018; Kaufman et al., 2013). While researchers have cautioned against using other than expert samples, empirical work suggests that laypersons may provide valid and reliable ratings when adequately prepared for the rating task (Hass et al., 2018; Storme et al., 2014). Either way, rater-characteristics should be taken into account when considering the provided responses in sophisticated statistical models (Myszkowski & Storme, 2019; Primi et al., 2019; Robitzsch & Steinfeld, 2018).

When considering the full product range, it becomes further clear that for some products (e.g., responses generated in divergent thinking tasks) no group of experts is readily available.

Who can be reasonably considered being an expert on how to creatively use a spoon? In such situations, raters are typically equipped with a more extensive coding guide instructing them that more creative responses tend to be uncommon, remote, and clever (Silvia et al., 2008). These three classical indicators of originality were already used by Guilford and colleagues for the scoring of divergent thinking tasks (Wilson et al., 1953). For example, responses for the Plot Titles task were scored for cleverness, whereas responses to the Consequences task were judged by human raters for their remoteness (P. R. Christensen et al., 1957). Similar scores as obtained by human ratings were used for tasks requiring the production of creative metaphors or humor (Beaty & Silvia, 2013; A. P. Christensen et al., 2018; Nusbaum et al., 2017; Primi, 2014).

Beyond these various characteristics of human ratings such as the type of products to rate, the sample of raters, or the scoring dimensions which differ from study to study, there is one common aspect of all such rating tasks that should be emphasized: the high workload put on the raters. For example, in a study focusing on divergent thinking, thousands of ratings might be needed (Kleinkorres et al., 2021) and researchers have already considered approaches to reduce the amount of work this brings along. For example, rating all responses generated by a participant at once (i.e., the full response set and not each response separately) has been proposed to reduce overall rating time (Shaw, 2021; Silvia et al., 2009), but scoring all responses at once comes along with the need to rate a comparably more complex product (i.e., all responses vs. only one response) and it has been shown that increasing complexity of sets of responses is associated with larger rater disagreement (Forthmann et al., 2017). In addition, scoring all responses is not an option if the focus of a study is the level of responses (Silvia et al., 2009).

A more sophisticated approach that reduces the burden on raters is to rely on planned missing data designs (Graham, 2009) which allow for unbiased estimation of creativity scores based on the response level. In fact, these designs take advantage of a reduced amount of information per rater without relying on a single data point by treating each response for what it is: an individual behavioral outcome to a given task. In the case of creativity research, these can be referred to as products. Notably, there are applications of planned missing data designs in creativity research (Barbot, 2020; Fürst, 2020; Primi et al., 2019). For example, Barbot (2020) merged different datasets in which some measures were in common and added another dataset, as a type of a linking sample to increase covariance coverage. Hence, his approach of integrative data analysis was highly similar with a planned missing data design. In addition, Fürst (2020) employed a planned missing data design for a more efficient, yet comprehensive, assessment of creative potential. While Barbot (2020) and Fürst (2020) used structural equation modeling and full information maximum likelihood estimation to handle missing data, Primi et al. (2019) examined simulated missing data patterns in the context of many facet Rasch modeling.

Psychometric Modeling of Human Ratings

Judge response theory (JRT) refers to the adaptation of polytomous item response theory models [e.g., the graded response model (Samejima, 1969) or the generalized partial credit model (Muraki, 1992)] for human ratings in the context of creativity research (Myszkowski, 2021; Myszkowski & Storme, 2019). JRT explicitly models differences in the rating behavior of human judges as reflected by severity (or leniency) effects and differences between raters with respect to their discrimination parameter. The considered unit of measurement in JRT is situated at the product level. Products must be conceptualized very broadly for the context of the current work. For example, one might consider more complex products such as newly designed electronic devices or very simple expressions of thought (e.g., a response generated within the context of creative thinking testing). Following Plucker et al. (2004) and Runco et al. (2001), we use the term 'product' in its broadest sense referring to something that is perceptible. With availability of ratings for each of the products in a given dataset, JRT as implemented in the R package jrt (Myszkowski, 2021) is a powerful tool for estimation of rater parameters and reliability. The jrt package uses the mirt package (Chalmers, 2012) for multidimensional item response theory modeling as estimation engine. The mirt package provides fast estimation algorithms (also when missing data are present) and it also includes highly efficient functions for simulation studies.

Specifically, the default setting of the jrt() function (i.e., the main function of the jrt package) fits various unidimensional polytomous item response theory models to the rating data and chooses the best fitting model (Myszkowski, 2021). The best fitting model is chosen based on the Akaike information criterion (AIC; Akaike, 1973) which combines a model's likelihood and a penalty for model complexity to emphasize that a useful model should be parsimonious and fit to the data. Smaller AIC values imply better model fit, while model parsimony is also taken into account. Consequently, in the jrt() function the model with the lowest AIC is chosen. In addition, it provides classical inter-rater agreement statistics such as intra-class correlations and also model-based empirical reliability estimates along with estimates of the latent creative quality for each of the products in the datasets. Here, reliability refers to the estimated squared correlation between the estimated latent creative quality [i.e., the factor scores provided by the jrt() function] and true latent creative quality. The square-root of this estimate is also known as factor determinacy index (FDI) in the literature on factor analysis. Hence, the FDI is an estimate of the correlation between the factor scores obtained for the rated products and the true factor scores. For both types of indices (i.e., reliability and FDI), cut-offs for research and practical assessment contexts exist and such cut-offs should be considered when planning a study involving creativity ratings. For a research project, the rating design could be planned towards a target FDI of .80 (Ferrando & Lorenzo-Seva, 2018), for example. Although we appreciate that such cut-offs are frequently subject of debate, and rightfully so, they offer a pragmatic benchmark that can be easily tested against.

In general, the recruitment of human raters is hard, and especially so in the case of rating creativity data, because rating creativity data is less of a trivial task as compared to other rating ventures. Further, the recruitment of human judges might be limited due to monetary reasons, for example, because the available project budget might be already exhausted. Regardless, researchers will be interested in getting the job done in the most pragmatic way:

with the least amount of time and money spent, while still obtaining high-quality data (i.e., reliable ratings that approximate the true ability of a participant).

Such a dataset that needs rating might for example include 5000 products (e.g., divergent thinking test responses). If we assume 500 responses can be rated (without rushing) in 1 hour (i.e., on average 7.2 seconds per response) and each rater receives $10 per hour of work, this would equal $100 per rater for rating the full response set. Obviously, costs increase with the number of raters that are employed for the task, and as the raters will usually not work in full synchronicity, the temporal aspect need not be forgotten either.

The number of raters depends, as mentioned above, on various considerations and, in a best-case scenario, should be chosen on the basis of empirical reasoning. For example, an arguably well-defined criterion would be to adhere to the afore-mentioned .80 FDI cut-off (i.e., FDI > .80), in order to ensure sufficient reliability for further analysis. If the most likely model and typical model parameters for the target rater population and the target product population are known (at least reasonable “guesstimates” are needed, which could also rely on experiences with similar data), it is possible to leverage mirt’s simulation functionality to answer such questions of rater design (i.e., how many raters are needed?).

Based on a simulation study, an expected FDI can be obtained. In the same vein, we obtain information regarding the number of raters and how many responses are rated by how many raters. Hence, the expected measurement precision and how much it would cost can be determined a priori. For example, based on the above calculations, it might be that a cut-off of .80 for the FDI will be surpassed with 3 raters which implies costs of $300. However, it might take 10 raters to surpass a cut-off of .90 and a much higher budget of $1000 would then be needed. Although the costs for the respective number of raters could have been easily calculated before, the simulation study extends the provided information by an estimate regarding the FDI, and thus the reliability of the obtained ratings. This information provides real value to researchers, as this enables them to consider trade-offs between monetary and temporal costs, and measurement precision.

To further refine planning of rater design, it is also highly useful that mirt models can be estimated for planned missing data designs (e.g., Fürst, 2018). For example, it is possible to simulate data that account for specific levels of missingness (e.g., one rating less for 20% of all responses). This functionality can thus be used in a pragmatic way to further reduce costs without sacrificing too much measurement precision. By means of planned missing data designs, rating designs can be efficiently and effectively planned towards both a target level of measurement precision and an available budget. Similarly, missing data designs could be used for planning towards a given due date at which the ratings must be available for further data analysis.

The Present Research

The goal of the current work is to introduce simulation-based planning of rating designs that (a) incorporate planned missingness designs, and (b) allow for an effective outweighing of a target level of measurement precision and monetary costs (or costs in terms of time, when approaching a due date). We argue that this work will be helpful for researchers studying creativity, as it provides a practical example on how to use planned missing rating designs for their own purposes. To this end, we illustrate the usefulness of this approach based on a realistic planning scenario when layperson ratings are to be used for scoring of an Alternate Uses Task (e.g., Hass et al., 2018).

Method

The empirical part of this work comprises of (a) an initial analysis of rating data and (b) a simulation study to inform planning of a missing data design. The first part was needed to derive a realistic simulation model and ranges for model parameters (e.g., discrimination parameters), whereas the second part was needed to see which planned missing data designs work well with respect to psychometric as well as cost-effectiveness criteria. Importantly, the empirical part should be understood as providing proof-of-concept on how to implement the approach for planning a missing data design for human ratings. As such, all reported findings are limited to the rater population that was sampled and the Alternate Uses Task as a measure of creative thinking, for example. Thus, we strongly recommend caution when interpreting the current findings. Especially, for other rater populations, other populations from which participants are sampled from, and other creativity measures, we recommend to contextualize and redo all steps outlined in this work.

Dataset

We first tested different rater models on an available dataset. The dataset included 3236 responses generated by $N = 209$ participants on two different Alternate Uses Tasks (using the words *box* and *rope,* respectively). Participants had two minutes to complete each of the tasks and were instructed to be creative. Each response was rated by three raters (undergraduate students majoring in psychology) using the subjective scoring method guidelines for divergent thinking (https://osf.io/vie7s). Following these guidelines, the raters used a 5-point Likert-scale. According to Cicchetti's criteria (Cicchetti, 2001), inter-rater reliabilities were fair in terms of absolute agreement (ICC = .42, 95%-CI: [.02, .63]) and consistency (ICC = .56, 95%-CI: [.53, .58]). The study was approved by the Institutional Review Board of The Pennsylvania State University. All participants gave informed consent to participate in the study.

Obtaining a Realistic Rater Model

Before a simulation can be set up for planning of a rating design, a reasonable simulation model and a realistic range of rater parameters (i.e., parameters related to raters' severity and discrimination between products) ought to be found. We used the jrt() function from the jrt package (Myszkowski, 2021) which is implemented in the statistical software R (R Core Team, 2021). This way, the best fitting polytomous IRT model was determined based on the Akaike Information Criterion and AIC-based model weights (Wagenmakers & Farrell, 2004). The parameter estimates from the best fitting model were then used to construct an empirically justified simulation setup. Model comparison results of all computed models as to the default function of jrt() can be found in Table S1 in the online supplemental material (https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70).

The best fitting model was the generalized partial credit model (Muraki, 1992). As this model will be used for the simulation below, it is worthwhile to consider its model equation

$$P(X \;=\; k \mid \theta_i, \alpha_j, d_j) = \frac{\exp[ak_{k-1}(a*\theta)+d_{k-1}]}{\sum_{v=1}^{K}\exp[ak_{v-1}(a*\theta)+d_{v-1}]}\,, \tag{1}$$

with θ_i being the latent score for response *i*, α_j being the discrimination parameter of rater *j*, d_j being the intercept vector of rater *j*, and ak_k being constraint to 0, 1,…, *K*-1 (with *K* being the number of response categories). In addition, *I* refers to the number of responses in the context of this work, and *J* to the number of raters. This parameterization of the GPCM is implemented in the mirt package (Chalmers, 2012) and commonly referred to as the slope-intercept parameterization (Matlock et al., 2018). The model is further identified by assuming that the latent response scores are *N*(0,1) distributed. Given that the data at hand had five response categories (i.e., *K* = 5), there were *K*-1 = 4 intercept parameters for each of the three raters (i.e., d_{1j}, d_{2j}, d_{3j}, and d_{4j} with *j* = 1,…, *J*), and three discrimination parameters (i.e., α_1, α_2, and α_3). The estimated model parameters for each rater are shown in Table 1. For example, Rater 2 was found to have a much higher discrimination parameter as compared to Rater 1 and Rater 3 which means that this rater was much better in distinguishing highly creative responses from less creative responses. In addition, for Rater 2 by far the lowest intercept parameters were obtained which means that this rater was the most severe during the rater process. Rater 1 and Rater 3 were much more lenient in their ratings with Rater 1 being the most lenient one in this rater sample (c.f., Table 1).

Table 1

Rater Parameter Estimates based on the Generalized Partial Credit Model

Parameter	Rater 1	Rater 2	Rater 3
α_j	$\alpha_1 = 0.72$	$\alpha_2 = 1.83$	$\alpha_3 = 0.77$
d_{1j}	$d_{11} = 3.84$	$d_{12} = -1.66$	$d_{13} = 1.67$
d_{2j}	$d_{21} = 3.53$	$d_{22} = -3.76$	$d_{23} = 1.03$
d_{3j}	$d_{31} = 1.92$	$d_{32} = -7.52$	$d_{33} = -0.89$
d_{4j}	$d_{41} = -1.16$	$d_{42} = -13.42$	$d_{43} = -3.79$

Note. α = discrimination parameter, *d* = intercept. The response category curves for all three raters can be found in Figure S1 in the online supplemental material (https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70).

Simulations

Construction of Planned Missing Data Matrices

In this work, we focus on matrix planned missing data designs (Silvia et al., 2014) because they allow equal distribution of work across available raters. We constructed the designs the following way:

1. We obtained all possible combinations of raters based on the overall number of raters and the target number of ratings per response. The combinations were calculated by means of the CombSet() function from the DescTools R package (Signorell, 2021). For example, for three available raters and two ratings per response there would be three possible combinations of raters: {{Rater 1, Rater 2}, {Rater 1, Rater 3}, {Rater 2, Rater 3}}.
2. We determined how many rows in the planned missing data matrix should be rated by each of the combinations obtained from Step 1. This was obtained by the floor function of the ratio of overall number of responses and the number of combinations obtained from Step 1. If the number of responses exceeded this number the last combinations were randomly sampled with replacement from all possible combinations.
3. The matrix planned missing data designs obtained from Step 2 were further reduced or increased as a final optional step. A reduced design was obtained by randomly setting a planned rating to a planned missing value for a fixed percentage of responses. Analogously, an increased design was obtained by randomly setting a planned missing value to a planned rating for a fixed percentage of responses. Responses were also chosen randomly for both types of designs.

Design

In our simulation design, we varied the number of raters (2 vs. 3 vs. 4 vs. 5) and the number of responses rated by each rater (2 vs. 3 vs. 4 vs. 5) resulting in 10 possible design cells (i.e., 5 ratings were only possible with 5 raters; see also Figure 1 below). The number of possible raters adheres to numbers of raters usually used for research purposes. In addition, we crossed this design with different percentages (20% vs. 40% vs. 60% vs. 80%) to reduce the number of ratings needed which resulted in 40 additional design cells. Decreasing a design with 3 ratings per response by 20%, for example, means that 20% of the responses will receive only 2 ratings per response. The responses and the rater who would not rate the response anymore were chosen randomly. Analogously, we crossed the design with different percentages (20% vs. 40% vs. 60% vs. 80%) to increase the number of ratings needed which resulted in another 24 additional design cells. Thus, here we did not combine cells with increasing percentages in which the number of raters equals the number of ratings (i.e., all complete designs). It should be noted, however, that increasing a complete design with two raters by 20% would also result from reducing a complete three rater design by 80% (which is already included). Increasing a design with 3 ratings per response by 20%, for example, means that 20% of the responses will receive 4 ratings per response. The responses and the rater who would rate this additional response were also chosen randomly. Thus, overall 74 different design cells were simulated.

Data Generation

We used the simdata() function from the mirt package (Chalmers, 2012) for data generation. First, we sampled latent response scores from a *N*(0, 1) distribution. Discrimination parameters were sampled from a *U*(0.72, 1.83) distribution (i.e., the range was taken from the estimates reported in Table 1). The intercept parameters were sampled as follows: first, we calculated the average across each rater's intercept parameters to reflect rater easiness. Then, we sampled from a *U*(-6.59, 2.03) to reflect rater easiness. Next, we subtracted each rater's easiness from their four intercept parameters for centered intercept parameters. Each of the four centered intercept parameters was averaged across raters and used to construct a sampling rationale for the four intercept parameters. The d_1 parameter was sampled from *U*(2.97, 4.93) with 2.97 being the average centered d_1 parameter across raters and 4.93 being the maximum of the centered d_1 parameters. The d_2 parameter was sampled from $U(1.95, 2.97)$ and the d_3 parameter from $U(-0.48, 1.94)$ with the lower bounds here being the average centered intercept parameters, respectively. Finally, the d_4 parameter was sampled from $U(-6.83, -0.49)$ with -6.83 being the minimum of the centered d_4 parameters. The sampled easiness and the sampled centered intercept parameters were added up to yield the intercept parameters for data generation. For each cell we simulated 500 replications of 1000 AUT responses (e.g., approximating an assessment context in which for $n = 100$ participants ten responses are to be expected on average). The average correlation across replications between the estimated latent response scores (based on the expected a-posteriori method; EAP) and the true latent response scores was our main dependent variable in this simulation. We further obtained the standard deviations and the standard errors of the correlations as an indicator of sampling variability. The R code to reproduce all reported results in this work is openly available via the Open Science Framework (https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70).

Results and Discussion

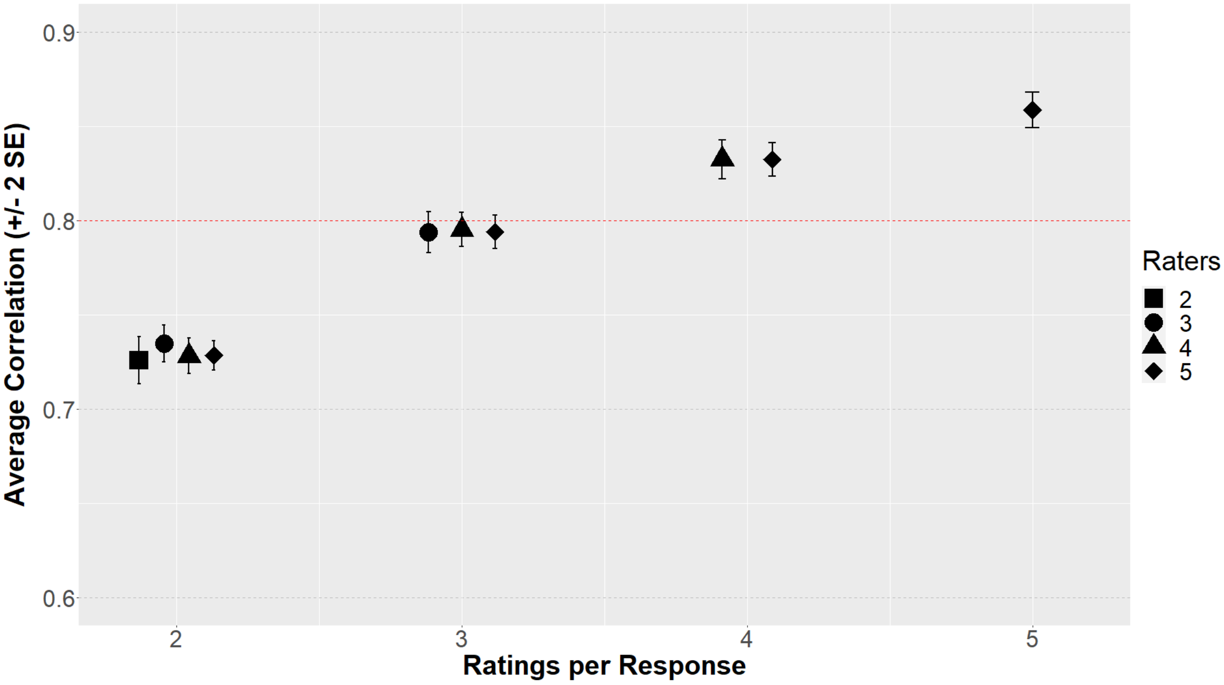
Simulation-Based Planning of a Rater Design

The reported findings from our simulation study serve the purpose of making readers familiar with interpreting findings obtained by the proposed approach for planning of missing data designs. In addition, we report the findings quite comprehensively so that interested researchers get an impression of how one might adjust simulation-based planning (e.g., by decreasing or increasing the number of rated responses for a proportion of raters) in ways that improve initially unsuccessful designs. For example, a design could be considered as unsuccessful when reliability is far above a target cut-off (the efficiency of the design can still be improved) or still below such a target cut-off (the expected psychometric quality must be improved).

First, we present the results of the full-data simulations, where either all raters rated all responses, or all responses were rated by n-1 raters in Figure 1. We observed a clear main effect for the number of ratings per response. Although the confidence intervals for all simulations supposing three ratings per response include the defined target level (*r* = .8) of the correlation between latent score estimates and their true values, on average this specific target level is not surpassed under this condition (i.e., three ratings per response). This holds independent of the number of raters that were specified. In order to exceed the defined target level of *r* = .8, at least four ratings per response would be needed, which corresponds to employing at least four independent raters.

Figure 1

Results of Full Design Simulations



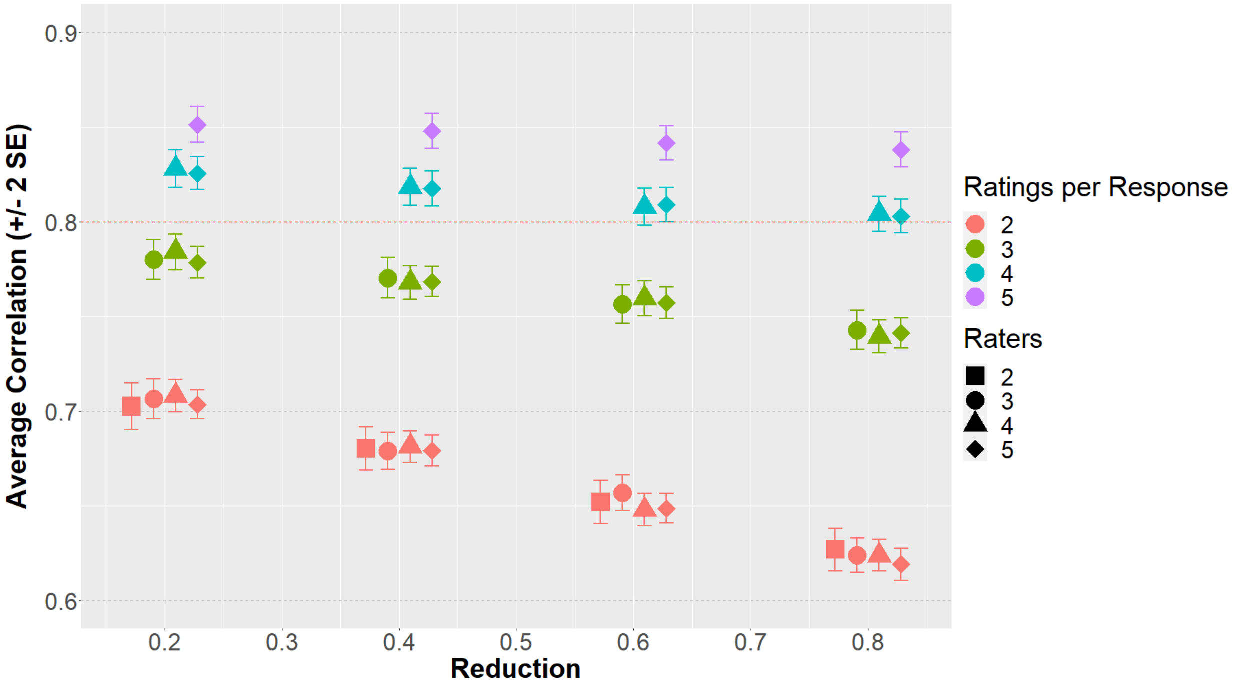
Note. Each point is based on 500 replications and 1000 responses in each replication. The red dotted line at .80 on the y-axis refers to the common cut-off for the correlation between latent score estimates and their true values. When the correlation surpasses this cut-off, latent score estimates display high enough measurement precision for research purposes (Ferrando & Lorenzo-Seva, 2018).

Next, we present the results of the planned-missingness data simulations, where in each simulation the ratings per response of the full dataset were reduced by either 20%, 40%, 60%, or

80% (Figure 2). Again, decreasing a design with 3 ratings per response by 20%, for example, means that 20% of the responses will receive only 2 ratings per response. Again, we observed a clear main effect for the number of ratings per response. However, although again at least four ratings per response yield the best results in terms of surpassing the a priori defined correlation of .80, further reducing the relative amount of responses that need to be rated at least four times, does not impair the estimated correlations very much. On the contrary, reducing the responses needed to be rated by all four raters by 60% still leaves enough information in the data to surpass the target level of *r* = .80. This finding can be readily translated to a monetary advantage, as not all raters have to rate all of the responses, but sufficient reliability is still achieved.

Figure 2

Results of Reduced Design Simulations

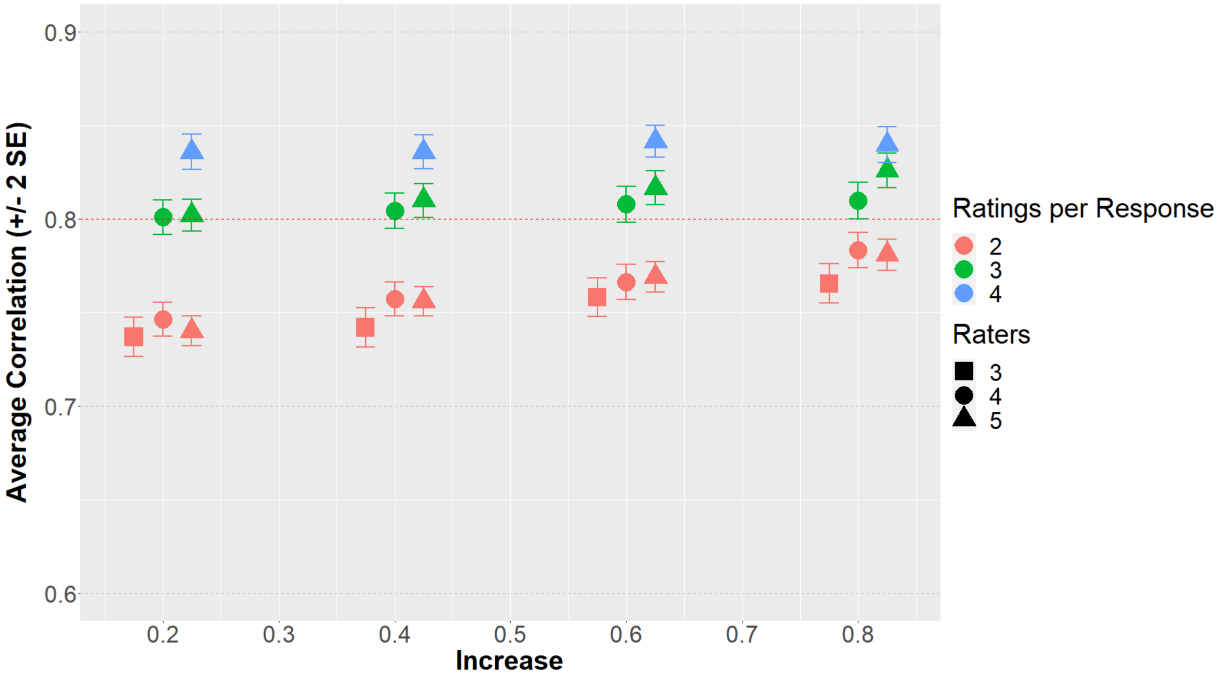


Note. Reduction = % of total responses that are rated by *n* - 1 raters. The red dotted line at .80 on the y-axis refers to the common cut-off for the correlation between latent score estimates and their true values. When the correlation surpasses this cut-off, latent score estimates display high enough measurement precision for research purposes (Ferrando & Lorenzo-Seva, 2018).

Lastly for this section, we show the results of the planned-missingness data simulations, where in each simulation the ratings per response of the full dataset were increased by either 20%, 40%, 60%, or 80% (Figure 3). In these simulations, the previously observed main effect of number of ratings per response remained. We were not able to identify any substantial effects that go beyond this main effect; the confidence intervals of all remaining simulation cells were overlapping. Increasing the responses needed to be rated by a given set of raters slightly increases the observed correlation, but according to the here provided data the differences might be negligible.

Figure 3

Results of Increased Design Simulations



Note. Increase = % of total responses that are rated by *n* + 1 raters. The red dotted line at .80 on the y-axis refers to the common cut-off for the correlation between latent score estimates and their true values. When the correlation surpasses this cut-off, latent score estimates display high enough measurement precision for research purposes (Ferrando & Lorenzo-Seva, 2018).

Cost-Effectiveness Calculations

In this section, we provide some insights into possible cost-effectiveness calculations, that is, considerations regarding a trade-off between measurement precision and monetary costs. To do so, we first provide a set of assumptions for our calculations: We assume that a layperson rater, who is properly trained via a short, written instruction regarding what is expected of them, can rate about 500 responses per hour. This equals 7.2 seconds per response, but this estimate seems reasonable given that raters will usually accelerate the rating process over time and with every response. For sake of the argument, we also assume that raters are paid $10 per every hour of work. In the current case, we further assume that instructing a rater does not count as time spent working; after all, we would like to provide relatively pure estimations only regarding the rating process itself. In addition to that, we base but not constrain our calculations to the assumption that any given number of raters can work perfectly parallel to each other. Although this assumption will be rarely met in reality, it will help to illustrate the inherent advantages of using certain planned-missingness rater designs. For our first example, we will further assume that four human raters can be appointed to rating data of an Alternate Uses Task with 1000 responses in total. We aim at illustrating the process that can be applied to decide for one or the other planned-missingness rater design.

In Table 2, we provide a complete overview of all relevant parameters important for deciding for a rater design. Each row of the table refers to a unique (planned-missingness) rater design. We explicitly report the number of total raters; the given ratings per response; whether the full, an increased, or a reduced dataset was used; how many responses were assigned to each rater; what the mean and the standard deviation of the obtained correlation was; how much money a specific design translates; and the estimated rating time in total and per rater (which would also equal the total rating time for all four raters, if all of them would be working perfectly parallel).

Table 2

Example of Cost-Effectiveness Calculations

N_{Raters}	Ratings per Response	Condition	$\text{Range}_{\text{Responses per Rater}}$	M_r	SD_r	Estimated Costs	Estimated $\text{Time}_{\text{total}}$ in h	Estimated $\text{Time}_{\text{per Rater}}$ in h
4	4	Full	1000	.832	.116	$80.00	8.00	2.00
4	4	reduction (20%)	940-957	.828	.113	$76.00	7.60	1.90
4	4	reduction (40%)	891-913	.819	.109	$72.00	7.20	1.80
4	4	reduction (60%)	831-864	.808	.109	$68.00	6.80	1.70
4	4	reduction (80%)	782-809	.804	.105	$64.00	6.40	1.60
4	3	increase (20%)	758-802	.801	.105	$62.12	6.21	1.55
4	3	increase (40%)	776-860	.804	.106	$64.20	6.42	1.61
4	3	increase (60%)	779-914	.808	.107	$66.50	6.65	1.66
4	3	increase (80%)	804-956	.810	.109	$68.70	6.87	1.72

Note. Reduction = % of total responses that are rated by *n* - 1 raters. Increase = % of total responses that are rated by *n* + 1 raters. An extended version of this table including much more simulated conditions can be found in Table S2 in the online supplemental material file in the OSF repository (https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70).

For example, while having the full data set rated by all four raters (i.e., 1000 responses per rater) would cost $80 and, on average, yield a correlation of .83 between latent score estimates and their true scores; using a design that supposes only 3 ratings per response, with an increase of one more rating per response for only 20% of the data, would reduce the total estimated costs by > ⅕ (i.e., ~22.5%), and still yields a correlation of *r* = .80. This reduction in monetary costs is obviously also reflected in the time that is needed to obtain all necessary ratings; that is, instead of 8h of scoring for the full data, implementing the planned missingness design of the provided example results in a total time of 6.2h.

It can be argued that this reduction of monetary and temporal costs by 22.5% could be understood as both a relative and an absolute increase of cost-effectiveness. Whereas in our example with 1000 responses, the absolute cost reduction of the planned missingness rating design can seem negligible in the light of huge research grants, or when researchers only plan on rating one creativity task like the Alternate Uses Task, the inherent benefit of these planned designs becomes clearer, when a larger scale is considered.

For example, consider a large online-panel study assessing creativity by means of a two-item Alternate Uses Task with 1000 participants. If we assume that each participant, on average, provides 10 responses per item, a huge dataset with 20,000 responses would be obtained. Appointing four raters to rate all of the responses would result in costs of $1,600 (40h of work per rater) and take a considerable amount of time, as rating creativity responses is usually not a full-time job and moreover exhausting for the raters (fatigue). If the design mentioned above would be applied to this situation (3 ratings per response, with an increase of one more rating per response for only 20% of the data), the relative decrease of costs would of course remain the same, but in terms of absolute numbers the cost decrease would add up to $360, which sometimes is the price of attending a conference to present the results of a study. In addition, of course the time for each rater working on their rating would decrease considerably (9h - which is longer than the time spent working in an ordinary 9-5 job).

Summary and Recommendations

In this work we proposed a simulation-based approach for effective planning of rater designs with missing data. We demonstrated in an empirical proof of concept illustration how a reasonable simulation model can be obtained from existing data, how simulations can be used to fine tune the planned design, and how based on these simulations cost-benefit analysis can be done when project budget and/or time are limited resources. Specifically, we used available rating data for responses on the Alternate Uses task and found by means of the the jrt package (Myszkowski, 2021) that the GPCM fitted these data best. Hence, we used the GPCM and the obtained parameter estimates for informing simulation-based planning. Then, our simulation implies strategies that are useful for fine-tuning the planned missing values design: run simulations with full data designs and varying numbers of raters, identify the full data designs that are closest to a target level of the correlation between estimated and true factor scores (e.g., .80), and finally increase or decrease the number of ratings per response for a certain proportion of randomly chosen responses.

Importantly, we have shown that even in situations in which ratings might not be too expensive in terms of monetary costs, enough money could be saved that allows a doctoral student, for example, to go to a conference. Clearly, in case that experts are needed as raters for a study the planning approach outlined in this paper is expected to result in even greater savings, because expert raters are much more expensive; for example, architects that would be hired to rate construction designs provided by participants of a study on architectural creativity.

We strongly recommend that researchers use this approach—adapted to the context of their studies—for the case that that human ratings of creative products are involved to ensure the quality of final scores based on planned missing data designs. We provide the needed R code for simulation-based planning in an openly accessible repository (https://osf.io/7b9z5/?view_only=902f015df3304dfbae60a4c06eb66c70) to facilitate this step for researchers who are not yet familiar with the software used in this work. However, having planned a missing data rating design implies that further steps are needed.

As a final step, one would reevaluate the best fitting model of the obtained ratings by means of the jrt package. Of course, the more is known about the target rater population (e.g., laypersons for rating divergent thinking responses), the unlikelier it will be that the JRT model fitting the data best will deviate from the anticipated model in the planning phase. However, as in our illustration here one might have only three raters available for setting up a reasonable simulation model (or even no data at all). In such situations the final data could better fit to a different model which should then be used for deriving final latent scores. Furthermore, also the finally achieved reliability of the scores should be reevaluated to check if the rating process resulted in the anticipated level of measurement precision and/or if the level of measurement precision is high enough for the purpose of measurement (Ferrando & Lorenzo-Seva, 2018). We recommend to check the square-root of empirical reliability of the final scores which provides an estimate of the correlation between estimated latent response scores and the true responses.

Limitations and Future Directions

The dataset we used in our study for illustration and for a hypothetical planning scenario might not have been comprehensive. Additional complexities are expected to arise when, for example, model parameters for each rater differ as a function of the task for which the ratings are needed. For example, in the used dataset, participants generated responses for two different AUT objects (i.e., *box* and *rope*) and we ignored that discrimination and intercept parameters in the GPCM could differ between both objects. Such differences could be considered during the simulation by means of using a multiple group model with as many groups as there are tasks in the planned study. However, such a more complex simulation would only make sense when enough empirical evidence for a mostly non-overlapping parameter range between the tasks is available. Of course, this knowledge can only be gained if such differences in parameters are evaluated and this can be nicely done at the stage of reevaluating model fit and reliability of the final ratings.

The outlined empirical example is further limited to a range of matrix planned missing data designs (Silvia et al., 2014). This design type is attractive as it will likely result in a well linked sample which guarantees unbiased parameter and latent score estimation. However, there are other designs that might be as attractive for a rating study. For example, Fürst (2020) used a design in which two raters (out of five for one of the tasks) rated all responses, whereas three other raters provided two ratings per response in a full matrix design. Of course, such designs come with their own disadvantages, namely that at least some raters are experiencing the full burden of the rating task. With that being said, we recommend researchers choosing their rating design also based on the expected work load of the single raters and take into consideration the experience of their raters. In addition, it is arguably a good idea to check a simulation even in situations that allows all raters to rate all responses, just to make sure study planning is sound. The open material we provide along with this paper can be easily extended to such other designs.

Furthermore, the designs considered in this work can be easily extended by anticipating other missing value issues (e.g., missing values because of study drop-out of participants). For example, when assuming that a certain proportion of responses will be missing completely at random, it is possible to incorporate this in the simulation to identify a “safe” rater design. Furthermore, it is important that not all studies will focus on comparably large numbers of products to be rated. Some studies may require only very few (or at least much fewer) ratings instead. For such situations one might further consider technical issues such as problems with model convergence, for example. In such situations the target model might not be estimable and it would be very useful to anticipate approaches to deal with such issues. For example, one could increase the number of iterations, focus on less complex models (i.e., models that require less parameters to be estimated), or use Bayesian estimation with somewhat informative priors. As a final remark it should be noted that there could be a trade-off between the number of Likert-points used by the raters and model complexity. While more scale points provide more information and potentially increase reliability, the estimated models would incorporate more parameters (e.g., intercept parameters in the GPCM) to be estimated.

Conclusion

Human ratings are ubiquitous in creativity research which makes running studies a laborious endeavor. In this work, we have demonstrated how information obtained from JRT and simulations can be used for a fine-tuned planning of missing data designs that reduce the amount of work needed for reliable scoring. We have further shown how such a careful planning further translates into cost-effectiveness considerations. Hence, we anticipate that the outlined approach will be of great practical value for the field and invite interested researchers to explore and use the material we made available. This way, research money—and a lot of time—will be saved for all of us.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Hrsg.), *2nd International Symposium on Information Theory* (S. 267–281). Akadémiai Kiadó.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Barbot, B. (2020). Creativity and Self-esteem in Adolescence: A Study of Their Domain-Specific, Multivariate Relationships. *The Journal of Creative Behavior*, *54*(2), 279–292. https://doi.org/10.1002/jocb.365

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., & Silvia, P. (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & Cognition*, *41*(2), 255–267.

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

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2022). The Machines Take Over: A Comparison of Various Supervised Learning Approaches for Automated Scoring of Divergent Thinking Tasks. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.559

Chalmers, R. P. (2012). **mirt**: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Christensen, A. P., Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2018). Clever people: Intelligence and humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 136–143. https://doi.org/10.1037/aca0000109

Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology*, *53*(2), 82–88. https://doi.org/10.1037/h0045461

Cicchetti, D. V. (2001). Methodological Commentary The Precision of Reliability and Validity Estimates Re-Visited: Distinguishing Between Clinical and Statistical Significance of Sample Size Requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 695–700. https://doi.org/10.1076/jcen.23.5.695.1249

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

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*. https://doi.org/10.1016/j.tsc.2016.12.005

Fürst, G. (2020). Measuring Creativity with Planned Missing Data. *The Journal of Creative Behavior*, *54*(1), 150–164. https://doi.org/10.1002/jocb.352

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the Dependability and Feasibility of Layperson Ratings of Divergent Thinking. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01343

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. Understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 332–340. https://doi.org/10.1037/a0034809

Kleinkorres, R., Forthmann, B., & Holling, H. (2021). An experimental approach to investigate the involvement of cognitive load in divergent thinking. *Journal of Intelligence*, *9*(1). https://doi.org/10.3390/jintelligence9010003

Kornilov, S. A., Kornilova, T. V., & Grigorenko, E. L. (2016). The Cross-Cultural Invariance of Creative Cognition: A Case Study of Creative Writing in U.S. and Russian College Students. *New Directions for Child and Adolescent Development*, *2016*(151), 47–59. https://doi.org/10.1002/cad.20149

Long, H. (2014). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. *Thinking Skills and Creativity*, *13*, 183–194. https://doi.org/10.1016/j.tsc.2014.05.002

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, *15*, 13–25. https://doi.org/10.1016/j.tsc.2014.10.004

Matlock, K. L., Turner, R. C., & Gitchel, W. D. (2018). A Study of Reverse-Worded Matched Item Pairs Using the Generalized Partial Credit and Nominal Response Models. *Educational and Psychological Measurement*, *78*(1), 103–127. https://doi.org/10.1177/0013164416670211

Mouchiroud, C., & Lubart, T. (2001). Children's Original Thinking: An Empirical Examination of Alternative Measures Derived From Divergent Thinking Tasks. *The Journal of Genetic Psychology*, *162*(4), 382–401. https://doi.org/10.1080/00221320109597491

Muraki, E. (1992). A GENERALIZED PARTIAL CREDIT MODEL: APPLICATION OF AN EM ALGORITHM. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Myszkowski, N. (2021). Development of the R library "jrt": Automated item response theory procedures for judgment data and their application with the consensual assessment technique. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 426–438. https://doi.org/10.1037/aca0000287

Myszkowski, N., & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 167–175. https://doi.org/10.1037/aca0000225

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 231–241. https://doi.org/10.1037/aca0000086

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research. *Educational Psychologist*, *39*(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1

Primi, R. (2014). Divergent productions of metaphors: Combining many-facet Rasch measurement and cognitive psychology in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 461–474. https://doi.org/10.1037/a0038055

Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 176–186. https://doi.org/10.1037/aca0000230

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.1.2). R Foundation for Statistical Computing.

Reinig, B. A., & Briggs, R. O. (2013). Putting Quality First in Ideation Research. *Group Decision and Negotiation*, *22*(5), 943–973. https://doi.org/10.1007/s10726-012-9338-y

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101–138.

Runco, M. A., Plucker, J. A., & Lim, W. (2001). Development and Psychometric Integrity of a Measure of Ideational Behavior. *Creativity Research Journal*, *13*(3–4), 393–400. https://doi.org/10.1207/S15326934CRJ1334_16

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/BF03372160

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

Signorell, A. (2021). *DescTools: Tools for Descriptive Statistics* (R package version 0.99.44).

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, *46*(1), 41–54. https://doi.org/10.3758/s13428-013-0353-y

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*(2), 79–85. https://doi.org/10.1016/j.tsc.2009.06.005

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. 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

Stevenson, C. E., Smal, I., Baas, M., Dahrendorf, M., Grasman, R., Tanis, C., Scheurs, E., Sleiffer, D., & van der Maas, H. (2020). *Automated AUT scoring using a Big Data variant of the Consensual Assessment Technique*.

Storme, M., Myszkowski, N., Çelik, P., & Lubart, T. (2014). Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges. *Learning and Individual Differences*, *32*, 19–25. https://doi.org/10.1016/j.lindif.2014.03.002

Taylor, C. L., & Barbot, B. (2021). Dual pathways in creative writing processes. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000415

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*(5), 362–370. https://doi.org/10.1037/h0060857