Applying many-facet Rasch modeling in the assessment of creativity

By: Ricardo Primi, Paul J. Silvia, Emanuel Jauk, and Mathias Benedek

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

Abstract:

Creativity assessment with open-ended production tasks relies heavily on scoring the quality of a subject’s ideas. This creates a faceted measurement structure involving persons, tasks (and ideas within tasks), and raters. Most studies, however, do not model possible systematic differences among raters. The present study examines individual rater differences in the context of a planned-missing design and its association with reliability and validity of creativity assessments. It applies the many-facet Rasch model (MFRM) to model and correct for these differences. We reanalyzed data from 2 studies (Ns = 132 and 298) where subjects produced metaphors, alternate uses for common objects, and creative instances. Each idea was scored by several raters. We simulated several conditions of reduced load on raters where they scored subsets of responses. We then compared the reliability and validity of IRT estimated scores (original vs. IRT adjusted scores) on various conditions of missing data. Results show that (a) raters vary substantially on the lenient-severity dimension, so rater differences should be modeled; (b) when different combinations of raters assess different subsets of ideas, systematic rater differences confound subjects’ scores, increasing measurement error and lowering criterion validity with external variables; and (c) MFRM adjustments effectively correct for rater effects, thus increasing correlations of scores obtained from partial with scores obtained with full data. We conclude that MFRM is a powerful means to model rater differences and reduce rater load in creativity research.

Keywords: creativity | assessment | raters | many-facet Rasch model | planned missing data

Article:

The creativity of an idea or product is necessarily subjective (Sawyer, 2006), so researchers commonly assess creativity by asking judges to assign creativity scores. This seemingly simple approach, however, raises complex methodological issues. How should reliability be quantified for faceted designs, such as several judges rating several items for several creativity tasks per participant? How can differences in leniency or severity between the judges be measured and corrected? Must every judge score every response for every participant?

The methods commonly applied to subjective ratings solve some but not all of these problems. In this article, we describe many-facet Rasch models (MFRM; Eckes, 2011; Linacre, 1994), an extension of Rasch modeling for faceted data. We illustrate how they can handle three issues in subjective scoring: (a) they yield a holistic measure of reliability; (b) they quantify and adjust for each rater's severity; and (c) they afford efficient planned-missing-data designs that greatly reduce rater burden. Using two illustrative data sets, we describe how to conduct and interpret many-facet Rasch models in R (R Core Team, 2016) and examine the influence of missing ratings on reliability.

Subjective Ratings in Creativity Research

Subjective ratings of ideas and products are widespread in creativity research. Divergent thinking tasks—perhaps the most popular lab tasks—are often scored for overall creativity (Silvia et al., 2008) or for dimensions like novelty, appropriateness, remoteness, or realism (Christensen, Guilford, & Wilson, 1957; Cropley & Kaufman, 2012; Diedrich, Benedek, Jauk, & Neubauer, 2015). Humor production tasks ask participants to come up with funny ideas and then have raters judge them for funniness (Nusbaum, Silvia, & Beaty, 2017). The Consensual Assessment Technique (CAT; Amabile, 1982), probably the best-known approach to subjective scoring, asks domain experts to rate the creative quality of products according to personal definitions of creativity.

Although popular, subjective ratings bring some underappreciated statistical and methodological problems. First, studies using ratings typically have a faceted data structure. For example, a sample of 150 participants will complete three divergent thinking tasks, and four raters will score every participant's response to every task. As a result, evaluating the reliability of the assessment design is not straightforward. Researchers typically will report the internal consistency of the three tasks, omitting variation due to raters, or report the agreement of the raters, omitting variation due to tasks.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) can provide a holistic estimate of reliability for faceted designs, and has been useful in creativity research (e.g., Long & Pang, 2015). But generalizability theory doesn't solve the second challenge: estimating and correcting for rater differences. Not surprisingly, raters can vary substantially when asked to judge how creative, funny, or useful an idea is. No matter how extensively they are trained, some raters will always be more lenient or severe than others (Eckes, 2011). Generalizability theory can estimate the amount of variance due to a rater facet, but it will not scale each rater on an underlying severity dimension or correct a participant's estimated trait score.

And third, subjective scoring methods impose a heavy burden as sample sizes increase (Forthmann et al., 2017). Virtually all studies use complete-data rating designs: all raters rate all responses for all participants. The practical issues of rating thus constrain the sample sizes researchers can collect. If participants complete four tasks where they generate on average five responses that three raters score, for example, then adding 100 participants creates 6000 more judgments. If researchers reduce this burden by having different raters score different subsets of responses, rater differences in leniency-severity confound the ratings.

Many-Facet Rasch Models (MFRM)

Many-facet Rasch models (Eckes, 2011; Linacre, 1994) are a potential solution. Many-facet Rasch models for polytomous ratings can be written, following Eckes (2011), as:

$$log(P_{nijk}/P_{nij(k-1)}) = \theta_n - \beta_i - \alpha_j - \tau_k \qquad (1)$$

This equation expresses the log of the odds of person *n* receiving a rating *k* rather than *k*-1 on item *i* by rater *j*. If a 4-point rating scale is used (0 to 3), this equation compares the odds of receiving 3 versus 2, 2 versus 1 and 1 versus 0. The interesting feature of this formulation is that the general odds of receiving any possible high rating—as compared to an immediately lower one—is a linear comparison of the subject *n* latent ability theta versus the "difficulty" of the task, represented by a joint combination of the task difficulty β*i*, rater *j* severity α*j*, and the general difficulty of receiving *k* rating as compared to *k*-1. With four possible ratings (0 to 3) there are three thresholds τk representing the difficulty of ratings 1 versus 0 (τ1), 2 versus 1 (τ2), and 3 versus 2 (τ3). These three parameters related to the task have a negative sign compared to theta, which has a positive sign. This means that the likelihood of having an idea that will be scored 3, for instance, is a fundamental comparison of subjects' creative ability versus the general difficulty of the task, the severity of the particular rater, and the global difficulty of score 3.

This MFRM formulation is an extension of the Rasch-Andrich Rating Scale Model called the three-facet rating scale model (Wright & Masters, 1982). MFRM extends this model by including the effect of differences in raters' severity. The latent trait parameter theta is thus not biased by rater differences.

Correcting for Rater Severity

In many-facet Rasch models, one can compute "fair average scores" (Eckes, 2011) that are a transformation of a subject's theta back to the original metric (in this case, a 0 to 3 scale). This is an adjusted expected score of a particular subject as if he or she had answered an average task that was scored by an average rater. When performing this adjustment, formula 1 is written as:

$$log(P_{nijk}/P_{nij(k-1)}) = \theta_n - \beta_0 - \alpha_0 - \tau_k, \qquad (2)$$

where β0 is the mean difficulty for items and α0 is the mean of the raters' severity. A transformation of Equation 2 can produce expected category response probabilities for subject *n* on each rating *r* up to a maximum rating of 3 (0 to 3) denoted as *Pns*. Then the adjusted average score is calculated as:

$$M_{fair\,n} = \sum_{r=0}^{s} rP_{ns} \qquad (3)$$

So each score *r* is weighted by the expected probability given subject *n*'s ability (and the difficulty of receiving a particular score), adjusting the other facets to an average rater and item, and then summed to give an average adjusted score for each subject.

Accommodating Missing Data

When using incomplete data, subjects will have their responses scored by different combinations of raters, so the unadjusted average scores of a subgroup of subjects will be affected by the particular rater who happened to score those subjects' responses. When two lenient raters are combined, for example, subjects' unadjusted score averages will tend to be higher compared to the whole group—this difference in rater behavior will be confounded with the subjects' ability.

The most interesting feature of the adjusted scores from MFRM is that rater severity is explicitly modeled and therefore is no longer confounded with the examinees' ability. Therefore, it solves the problem of the comparability of subjects' scores when analyzing incomplete data. Subjects' theta scores are thus comparable across *booklets*, which represent different combinations of raters for specific subsets of responses. This opens the possibility of reducing rater workload because not all raters need to rate all ideas—and not all tasks need to be answered by all subjects.[1]

The Present Research

The MFRM has a long history of successful application in various research areas (Engelhard & Wind, 2018; Linacre, 2018) but is relatively unknown in creativity research. Some recent studies have applied MFRM to investigate how characteristics of novice raters affect ratings for the CAT (Tan et al., 2015). Other studies have applied it to ratings of children's creative writing (Barbot, Tan, Randi, Santa-Donato, & Grigorenko, 2012), metaphors (Primi, 2014b), and studies of rater bias (Hung, Chen, & Chen, 2012). We believe that MFRM can improve current methods of assessing creativity and deserves serious consideration from researchers who collect subjective ratings. We thus reanalyzed data using MFRM from two studies (Jauk, Benedek, & Neubauer, 2014; Silvia & Beaty, 2012) in which raters judged ideas produced in ideation tasks. For both studies, we fit a MFRM model and explored the model parameters, especially raters' parameters. We also compared original scores with latent scores (i.e., the adjusted "fair average" scores) corrected for raters' differences. Finally, we introduced missing data by simulating a situation where not all raters rated all responses. We then explored the correlations of adjusted versus original scores with the scores computed with full data set, as well as effects of missing data on validity coefficients with external variables.

Our main questions were (a) is there a noticeable difference in terms of leniency versus severity between raters? If it exists, does it affect the quality of measurement? (b) Is there a practical advantage, in terms of improved reliability and validity, of using MFRM to model differences between raters? (c) Can we reduce the load on raters by doing an optimized missing data plan and then using MFRM to achieve a common metric between raters? How much can rater burden

[1] Recently, Fürst (2018) tested the feasibility of planned missing design in creativity assessment and showed that it is possible to estimate convergent and discriminant validity while reducing by a third the working load from participants and raters. This work differs from the MFRM approach because it is done in the context of structural equation modeling (SEM) estimated via full information maximum likelihood, which is a modern way to deal with missing data. Fürst and colleagues were mainly interested in estimation of correlations between latent factors. Our work uses an IRT approach to estimate unbiased scores for subjects while controlling for systematic rater effects. The IRT approach requires an explicit emphasis on defining a common metric disentangling rater effects from subjects' latent traits (linking and equating) that may not be the case for SEM.

be reduced without reducing validity and reliability? Our final aim is to provide a practical tutorial in MFRM using FACETS and R (see online supplemental materials).

Study 1: Metaphor Ratings

This study reanalyzes data from a study in which 132 participants generated two metaphors that were rated by three raters on a 5-point scale (1 = *not at all creative*, 5 = *very creative*) and completed 6 inductive reasoning tasks as measures of Gf and a measure of Big-5 personality structure (for details, see Silvia & Beaty, 2012).

Design and Statistical Modeling

Each participant gave two responses that were rated by three raters, so we have six scored responses per participant, resulting in 786 data points (132 subjects × 6 responses). We used FACETS (Linacre, 2018), which implements several MFRM models, using joint maximum-likelihood estimation method (JMLE), to compute the model parameters of Formula 1 and scores of average creativity using Formula 3. Online supplemental material shows how to run MFRM with this data in FACETS and in R.

We ran models three times: (a) the benchmark with complete data, (b) then introducing 33% of missing data, keeping 67% of the total available data points (g67), and (c) then introducing 57% of missing data, keeping 43% of the total available data points (g43). For the g67, we eliminated data from one of the three raters. For the g43 we reproduced a common situation where a set of common responses is scored by all raters and then each subject is scored by only one rater for the remaining dataset. We randomly selected 20 subjects to be the common responses for which we preserved the complete data. For the remaining 112 subjects we selected scored responses of only one rater, rotating raters to have equal number of subjects scored by each rater. This group had a total of 342 data points. We compared reliability and validity indices across these levels of incompleteness.

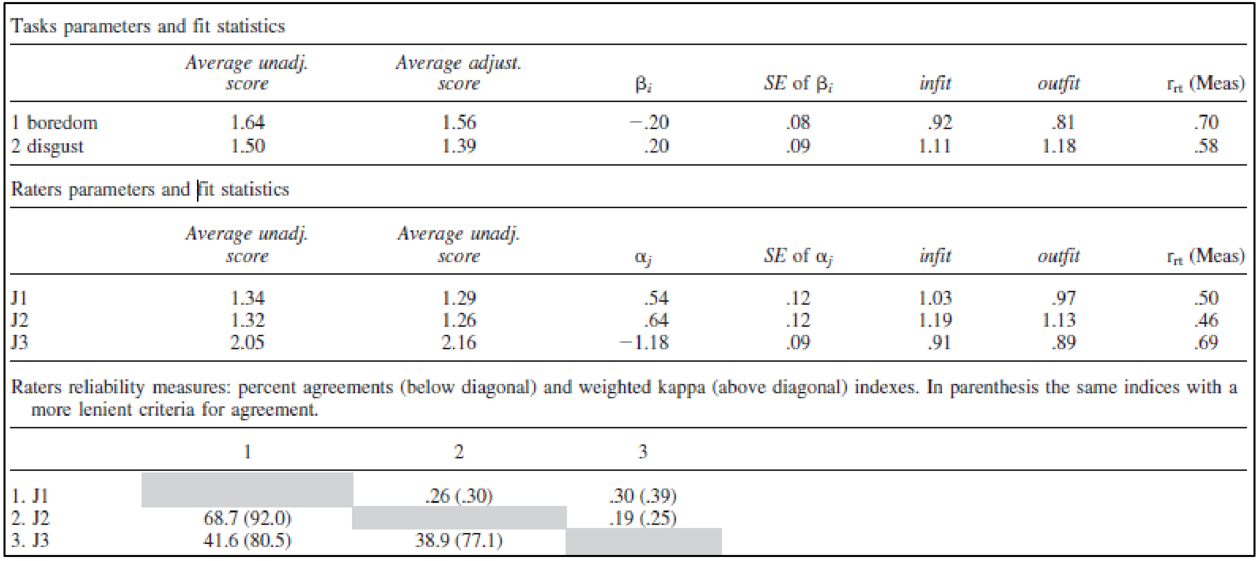
Results and Discussion

We first explored the model parameters of tasks and raters using the full dataset. The upper part of Table 1, in the first two lines, shows average scores on each task, the average adjusted score—the observed score corrected for the other facets—beta parameters of task difficulty (and their standard errors), fit statistics (*infit* and *oufit*)[2], and item-theta (latent score) correlations. The next

[2] After the MFRM model parameters are estimated, an expected score E is calculated for each subject's n response to item i scored by rater j. A residual r score is calculated comparing the expected score with the actual response x: $r = x - E$. These residuals are squared and standardized by dividing them by the modeled variance of the residuals. The variance of the residual accounts for the fact that for some situations the model is much more certain about a particular expected response, such as when person ability is far apart from the difficulty of task or the severity of the rater. These squared residuals are accumulated over items and persons and then averaged to calculate summaries of item and person *misfit*. This average is called *outfit.* It is more sensitive to unexpected residuals found in situations in which person ability is far apart from item difficulty/rater severity. Another information-weighted index, *infit*, weighs heavily residuals that occur in items/raters whose difficulties are closer to the ability of the person. For large samples the range of reasonable values for fit indices is 0.8–1.2. Indices higher than 1.4 indicate many responses in unexpected directions. Indices below .80 could indicate unmodeled dependences, that is, a facet

three lines show the same information for raters. Fit indexes are directly proportional to residuals—the difference between observed versus expected scores. Values below 1.4 usually indicate acceptable fit (Linacre, 2018). The lower part of Table 1 shows classical interrater reliability measures based on absolute agreement (consensus measures): the percent agreement and the weighted Kappa index.

Table 1. Basic Statistics and Psychometric Information on Metaphor Tasks and Raters



Tasks parameters and fit statistics

	Average unadj. score	*Average adjust. score*	β_i	*SE* of β_i	*infit*	*outfit*	r_{rt} (Meas)
1 boredom	1.64	1.56	−.20	.08	.92	.81	.70
2 disgust	1.50	1.39	.20	.09	1.11	1.18	.58

Raters parameters and fit statistics

	Average unadj. score	*Average unadj. score*	α_j	*SE* of α_j	*infit*	*outfit*	r_{rt} (Meas)
J1	1.34	1.29	.54	.12	1.03	.97	.50
J2	1.32	1.26	.64	.12	1.19	1.13	.46
J3	2.05	2.16	−1.18	.09	.91	.89	.69

Raters reliability measures: percent agreements (below diagonal) and weighted kappa (above diagonal) indexes. In parenthesis the same indices with a more lenient criteria for agreement.

	1	2	3
1. J1		.26 (.30)	.30 (.39)
2. J2	68.7 (92.0)		.19 (.25)
3. J3	41.6 (80.5)	38.9 (77.1)	

Note. Average adjusted score is calculated as demonstrated in Equation 3 considering the rater facet. β_i, task *i* difficulty parameter, α_j rater *j* parameter. r_{rt} (Meas) is the correlation of item scores and rater total scores that combine other items or raters. In the lower part of the matrix presented at the bottom of the table we show inter-rater reliability index calculated from standard formula and, in parenthesis, the results considering a more lenient criterion for agreements tolerating a difference of one point. In the calculation of weighted kappa we supplied the following weights for the perfect to worst disagreement: 1 for a distance of 0, 1 for a distance of 2, 4 for a distance of 3, 8 for a distance of 4. All inter-rater reliability calculations made on R package irr (Gamer, Lemon, Fellows, & Singh, 2015).

We observe that task difficulty isn't very dispersed (β's from −.20 to .20) in comparison to raters, who had a wider range of severity (α's from −1.18 to .64). Fit statistics suggest a good fit: there were no marked deviations from the expected model responses when we analyze each rater's or task's residual separately. The rater-total (measure) and task-total (measure) correlation is high, above .50. The Rasch average reliability, which is a measure of internal consistency based on estimated parameters of Rasch model, was *rtt* = .46. This is low when compared with expected benchmarks of .70 as acceptable level of reliability. But this should be interpreted relative to the number of data points per subject (only six) and also considering that Rasch reliability is an estimate of the lower-bound of the true reliability value (Linacre, 2004). Raters' classical reliability indexes—agreement and weighted kappa—are presented in the lower part of Table 1 (see Cohen, 1968; Stemler, 2004). These indexes show that in general raters have fair agreement (.25 to .39). Only J2 versus J3 had a lower agreement when considering exact scores. In general, we observed that raters' scoring behavior showed an acceptable level of agreement.

is more discriminating than expected by the Rasch model (see more details here https://www.rasch.org/rmt/rmt103a.htm and here https://www.rasch.org/rmt/rmt82a.htm).

Figure 1 presents the distributions of all model parameters (subjects' ability, items difficulties, raters' parameters, and scale thresholds for score points). Since all parameters are on a common scale, we can compare parameters with each other. A first noticeable result is that the distribution of creativity is positively skewed: it is denser at the lower end (−1 to −4.5). At this level the most likely score, given by the right most figure, is 1 "*not at all creative*." We noted also that the two items were equivalent in terms of difficulty, and that one rater was substantially more lenient than the other two.

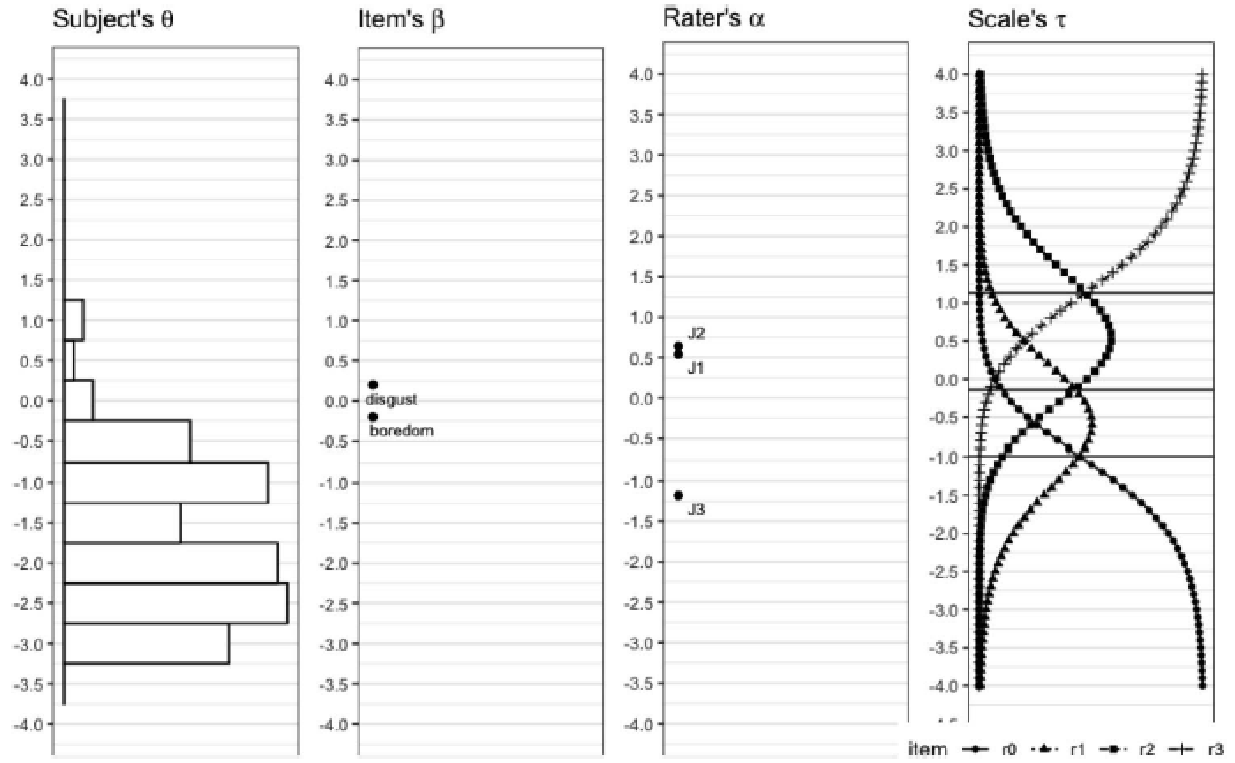
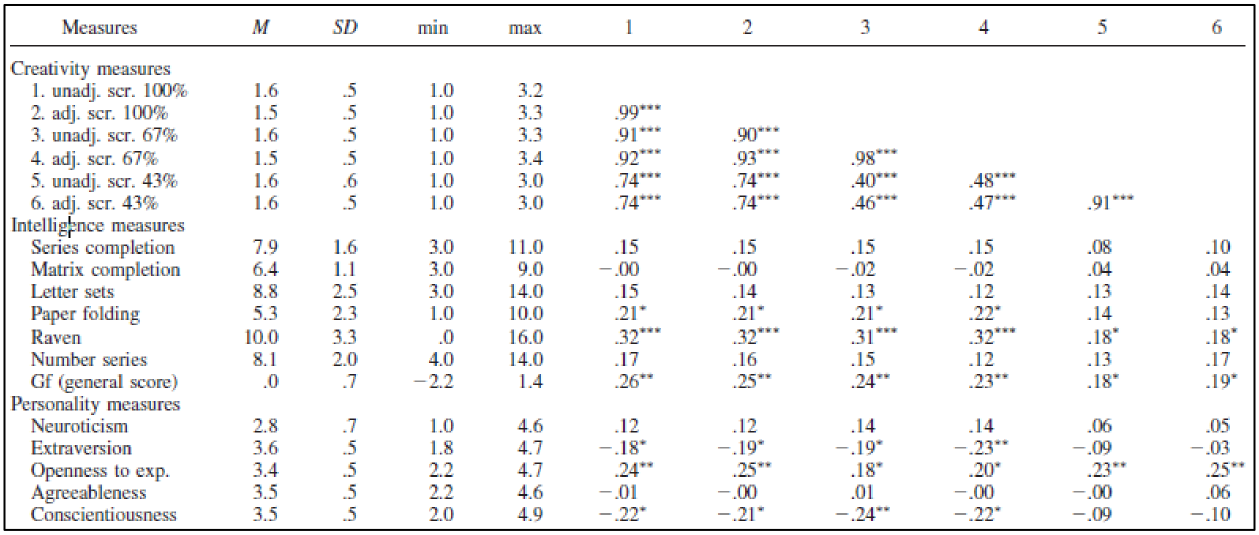


Figure 1. Construct map showing the distributions of subject's abilities, item's difficulties, rater's severities and category thresholds.

We next examined how creativity scores estimated with MFRM in two conditions of missingness compare with scores obtained from the full dataset. Table 2 shows the correlations under the three conditions of incomplete data (full 100%, 67%, and 43%) and adjustment: original (unadj.) versus equated (adj.). It also shows score correlations with criterion measures. We considered *adj. scr. 100%.* as a benchmark because it contains all data and it is adjusted for eventual differences in rater's leniency-severity. The full data set (*adj. scr. 100%*) correlates *r* = .93 with *adj. scr. 67%*, which is slightly above its correlation (*r* = .90) with original *unadj. scr. 67%*. Moreover, the full data set condition (*adj. scr. 100%*) correlates *r* = .74 with *adj. scr. 43%* and *unadj. scr. 43%*.

Looking to criterion validity, we observe that the pattern of correlations with the full dataset of creativity and external measures (intelligence and personality) remained largely the same if we calculate creativity scores that preserved up to 67% of data. With 43% of the full data set, the magnitude of the correlations tended to be slightly lower. We didn't observe differences in validity coefficients when comparing adjusted full data versus original scores where the raters' leniency-severity is not controlled for.

Table 2. Correlations of Scores Under Various Conditions of Completeness and Adjustment With Criterion Measures



Measures	*M*	*SD*	min	max	1	2	3	4	5	6
Creativity measures										
1. unadj. scr. 100%	1.6	.5	1.0	3.2						
2. adj. scr. 100%	1.5	.5	1.0	3.3	$.99^{***}$					
3. unadj. scr. 67%	1.6	.5	1.0	3.3	$.91^{***}$	$.90^{***}$				
4. adj. scr. 67%	1.5	.5	1.0	3.4	$.92^{***}$	$.93^{***}$	$.98^{***}$			
5. unadj. scr. 43%	1.6	.6	1.0	3.0	$.74^{***}$	$.74^{***}$	$.40^{***}$	$.48^{***}$		
6. adj. scr. 43%	1.6	.5	1.0	3.0	$.74^{***}$	$.74^{***}$	$.46^{***}$	$.47^{***}$	$.91^{***}$	
Intelligence measures										
Series completion	7.9	1.6	3.0	11.0	.15	.15	.15	.15	.08	.10
Matrix completion	6.4	1.1	3.0	9.0	−.00	−.00	−.02	−.02	.04	.04
Letter sets	8.8	2.5	3.0	14.0	.15	.14	.13	.12	.13	.14
Paper folding	5.3	2.3	1.0	10.0	$.21^{*}$	$.21^{*}$	$.21^{*}$	$.22^{*}$	.14	.13
Raven	10.0	3.3	.0	16.0	$.32^{***}$	$.32^{***}$	$.31^{***}$	$.32^{***}$	$.18^{*}$	$.18^{*}$
Number series	8.1	2.0	4.0	14.0	.17	.16	.15	.12	.13	.17
Gf (general score)	.0	.7	−2.2	1.4	$.26^{**}$	$.25^{**}$	$.24^{**}$	$.23^{**}$	$.18^{*}$	$.19^{*}$
Personality measures										
Neuroticism	2.8	.7	1.0	4.6	.12	.12	.14	.14	.06	.05
Extraversion	3.6	.5	1.8	4.7	$-.18^{*}$	$-.19^{*}$	$-.19^{*}$	$-.23^{**}$	−.09	−.03
Openness to exp.	3.4	.5	2.2	4.7	$.24^{**}$	$.25^{**}$	$.18^{*}$	$.20^{*}$	$.23^{**}$	$.25^{**}$
Agreeableness	3.5	.5	2.2	4.6	−.01	−.00	.01	−.00	−.00	.06
Conscientiousness	3.5	.5	2.0	4.9	$-.22^{*}$	$-.21^{*}$	$-.24^{**}$	$-.22^{*}$	−.09	−.10

Note. On creativity measures: *unadj. scr*: average scores not adjusted for rater effects, *adj. scr.*: average scores adjusted for rater effects, that is, equated scores. The amount of data used to calculate scores varied across three levels: full dataset *100%*, *67%* and *43%*.
* $p < .05$. ** $p < .01$. *** $p < .001$.

It is important to emphasize the utility of construct maps, a byproduct of the application of MFRM, to visualize all model parameters on a common metric. This enhances the understanding of the latent variable metric, making it less arbitrary and more meaningful (Embretson, 2006; Primi, 2014b; Wyse, 2013). It is interesting to note that the distribution of creative potential as measured by metaphor tasks is positively skewed. For instance, when we see theta distributions in relation to the actual response categories thresholds (τ) we learn that most subjects' ideas receive an overall score as "not at all creative." This indicates that that creative idea generation is quite difficult (cf., Primi, 2014a). Besides adjusting scores with respect to rater differences, MFRM can be valuable in applied creativity measurement. Construct maps and task and rater parameters (difficulty and consistency) can help understand raters' behavior in terms of difficulty-leniency as well as reliability. Task parameters can shed light on their usefulness and consistency. Finally, category threshold maps and parameters can assist in the development of scoring rubrics. In summary, MFRM produced good fit to the data. The tasks' β difficulty parameters didn't differ substantially, but the raters' α's, on the other hand, were more dispersed, justifying the need to correct for rater differences. The adjusted "fair average" scores performed well despite missing data. In general, these results indicate that you can have similar scores with about 1/3 workload reduction on raters.

Study 2: Ratings for Alternate Uses and Instances Tasks

This study is a reanalysis of a study in which 298 participants generated responses to six divergent thinking tasks that were rated for creativity on a 4-point scale (0 = *not creative* to 3 = *very creative*) by four raters and completed several measures of creative achievement, personality, and intelligence (for details, see Diedrich et al., 2018; Jauk, Benedek, Dunst, & Neubauer, 2013; Jauk et al., 2014). The primary goal in this reanalysis is to fit a MFRM model, to compare the reliability and validity of original scores versus adjusted scores corrected for rater differences, and to do so under two conditions of simulated missing data that reflect different patterns of rater assignments.

Design and Statistical Modeling

The DT data encompasses the responses of 298 people to six DT tasks. The average number of responses to one DT task was 12 (*SD* = 6.13), and the total number of responses to all 6 tasks was 22,064. The total number of nonredundant responses that were evaluated by the raters was 9,013 (45%). The final total data set comprising all responses with ratings from four raters thus included 88,256 data points.

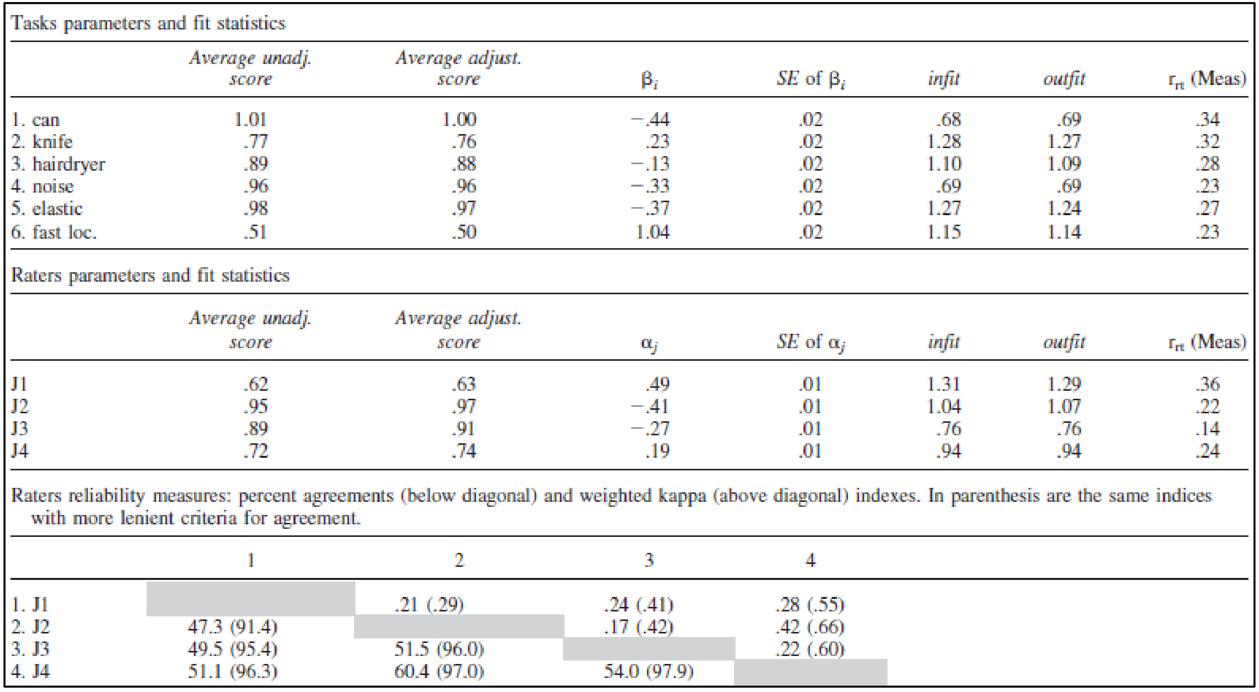
To evaluate the effects of missing data, we computed average creativity scores based on complete data and compared them to scores computed with incomplete data (50% and 25%). For the complete data set there were on average 74 responses per subject (across six DT tasks), with a minimum of 19 and maximum of 144. Since each response was scored by four raters there were, on average, 296 data points per subject to calculate the average creativity scores. For incomplete data scores, we employed a Balanced Incomplete Block (BIB) design. The 50% group had ratings from combinations of two of the four raters for all six tasks. The 25% group had ratings from combinations of two of the four raters, but for only three tasks. Appendix A has detailed information on the combinations of raters and tasks used. The BIB was an essential part of this design to guarantee the connectedness of the dataset.

Results and Discussion

We again focused our analysis on exploring the model parameters of MFRM and on comparing subject scores estimated by MFRM with the full dataset versus two conditions of missingness. Examining Table 3 we can observe, first, that fit indices were acceptable since none of the *infit* and *outfit* indices were 1.4 or higher. Second, the task and rater parameters vary systematically. αj varies from −.44 (can) to 1.04 (fast locomotion), and βi varies from −.41 (J2) to .49 (J1), a difference of .90 logits. The ratio of the raters' alpha parameters differences between any two raters divided by their standard errors would reach statistical significance. This signals a potential problem of confounding subject abilities with rater characteristics.

Third, the kappa consensus measures of interrater reliability indicate fair to moderate agreement. If part of the interrater disagreements is due to differences in levels of severity, the index should increase when we recalculate kappa with relaxed criteria for agreement. We thus recalculated indices allowing that a difference of one point will still be considered agreement. Indeed, the numbers in parentheses in the lower part of Table 3 shows that interrater reliability goes up when we consider a more liberal criterion. Absolute agreement reaches levels higher than 90%, and weighted kappas are now in in the range of fair to substantial agreement (in kappa calculations we supplied the following weights for the perfect to worst disagreement: 0 for a distance of 1, 1 for a distance of 2, 4 for a distance of 3, 8 for a distance of 4). This result again reinforces the point that raters vary in their leniency.

Table 3. Basic Statistics and Psychometric Information on Tasks and Raters



Tasks parameters and fit statistics

	Average unadj. score	*Average adjust. score*	β_i	*SE* of β_i	*infit*	*outfit*	r_{rt} (Meas)
1. can	1.01	1.00	−.44	.02	.68	.69	.34
2. knife	.77	.76	.23	.02	1.28	1.27	.32
3. hairdryer	.89	.88	−.13	.02	1.10	1.09	.28
4. noise	.96	.96	−.33	.02	.69	.69	.23
5. elastic	.98	.97	−.37	.02	1.27	1.24	.27
6. fast loc.	.51	.50	1.04	.02	1.15	1.14	.23

Raters parameters and fit statistics

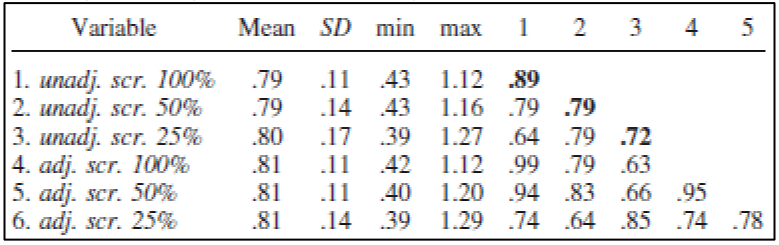
	Average unadj. score	*Average adjust. score*	α_j	*SE* of α_j	*infit*	*outfit*	r_{rt} (Meas)
J1	.62	.63	.49	.01	1.31	1.29	.36
J2	.95	.97	−.41	.01	1.04	1.07	.22
J3	.89	.91	−.27	.01	.76	.76	.14
J4	.72	.74	.19	.01	.94	.94	.24

Raters reliability measures: percent agreements (below diagonal) and weighted kappa (above diagonal) indexes. In parenthesis are the same indices with more lenient criteria for agreement.

	1	2	3	4
1. J1		.21 (.29)	.24 (.41)	.28 (.55)
2. J2	47.3 (91.4)		.17 (.42)	.42 (.66)
3. J3	49.5 (95.4)	51.5 (96.0)		.22 (.60)
4. J4	51.1 (96.3)	60.4 (97.0)	54.0 (97.9)	

Note. Average adjusted score is calculated as demonstrated in Equation 3 considering the rater facet. β_i, task *i* difficulty parameter, α_j rater *j* parameter. r_{rt} (Meas) is the correlation of item scores rater total scores that combine other items or rater score. In the lower part we present inter-rater reliability index calculated from standard formula and, in parenthesis, the results considering a more lenient criteria for agreements tolerating a difference of one point. In the calculation of weighted kappa we supplied the following weights for the perfect to worst disagreement: 1 for a distance of 0, 1 for a distance of 2, 4 for a distance of 3, 8 for a distance of 4. All inter-rater reliability calculations made on R package *irr* (Gamer, Lemon, Fellows, & Singh, 2015).

Table 4. General Descriptive Statistics and Intercorrelations Among the Six Original Observed Average Ratings of Ideas (Unadjusted) Versus Corrected Averages by the MFRM (Adjusted)



Variable	Mean	*SD*	min	max	1	2	3	4	5
1. *unadj. scr. 100%*	.79	.11	.43	1.12	**.89**				
2. *unadj. scr. 50%*	.79	.14	.43	1.16	.79	**.79**			
3. *unadj. scr. 25%*	.80	.17	.39	1.27	.64	.79	**.72**		
4. *adj. scr. 100%*	.81	.11	.42	1.12	.99	.79	.63		
5. *adj. scr. 50%*	.81	.11	.40	1.20	.94	.83	.66	.95	
6. *adj. scr. 25%*	.81	.14	.39	1.29	.74	.64	.85	.74	.78

Note. On the Diagonal We Present a Global Index of Reliability Calculated from Standard Errors of the Latent Scores That We Estimated for Each Subject. unadj. scr: average scores not adjusted for rater effects, adj. scr.: average scores adjusted for rater effects, that is, equated scores. The off-diagonal correlation coefficients are all significant at $p < .05$.

But to what extent do rater differences impact scores? Does MFRM correct for the potential differences that may arise? Examining Table 4 we see that for the full dataset, unadjusted (*unadj. scr. 100%*) and adjusted scores (*adj. scr. 100%*) are practically the same, $r = .99$. This result is expected since every rater scores all subjects in every task and therefore there is no confounding of rater/tasks with subjects. This constitutes the best-case scenario where we observed all data points and used adjusted scores. With incomplete data we see a different picture. The worst-case scenario occurs when we compare adjusted full dataset scores (*adj. scr. 100%*) with unadjusted 25%-dataset (*unadj. scr. 25%*). In this case the correlation drops to *r* = .64. This constitutes a lower benchmark of examining the effect of reducing data points—and not adjusting scores of combinations of raters and tasks that vary in levels of difficulty/severity. This contrast suggests that lower numbers of observations and, consequently, different pairings of raters that vary in their level of severity, introduces substantial error and reduces the correlation with the full dataset.

One important question is what happens when we adjust scores for raters effects using MFRM? These adjustments control rater and task effects only. The correlation of the adjusted full dataset score (*adj. scr. 100%*) with adjusted 25%-dataset scores (*adj. scr. 25%*) increases to *r* = .74 (as compared with *r* = .64). Interestingly, we observe the same effect of the model adjustments with the 50%-dataset. Adjusted full dataset scores (*adj. scr. 100%*) correlates *r* = .79 with unadjusted 50%-dataset scores (*unadj. scr. 50%*). But it correlates *r* = .95 with adjusted 50%-dataset scores (*adj. scr. 50%*). It is not a surprise to see that correlations increase if we have more data, but when using MFRM to adjust scores, we can practically obtain similar results with using half the data.

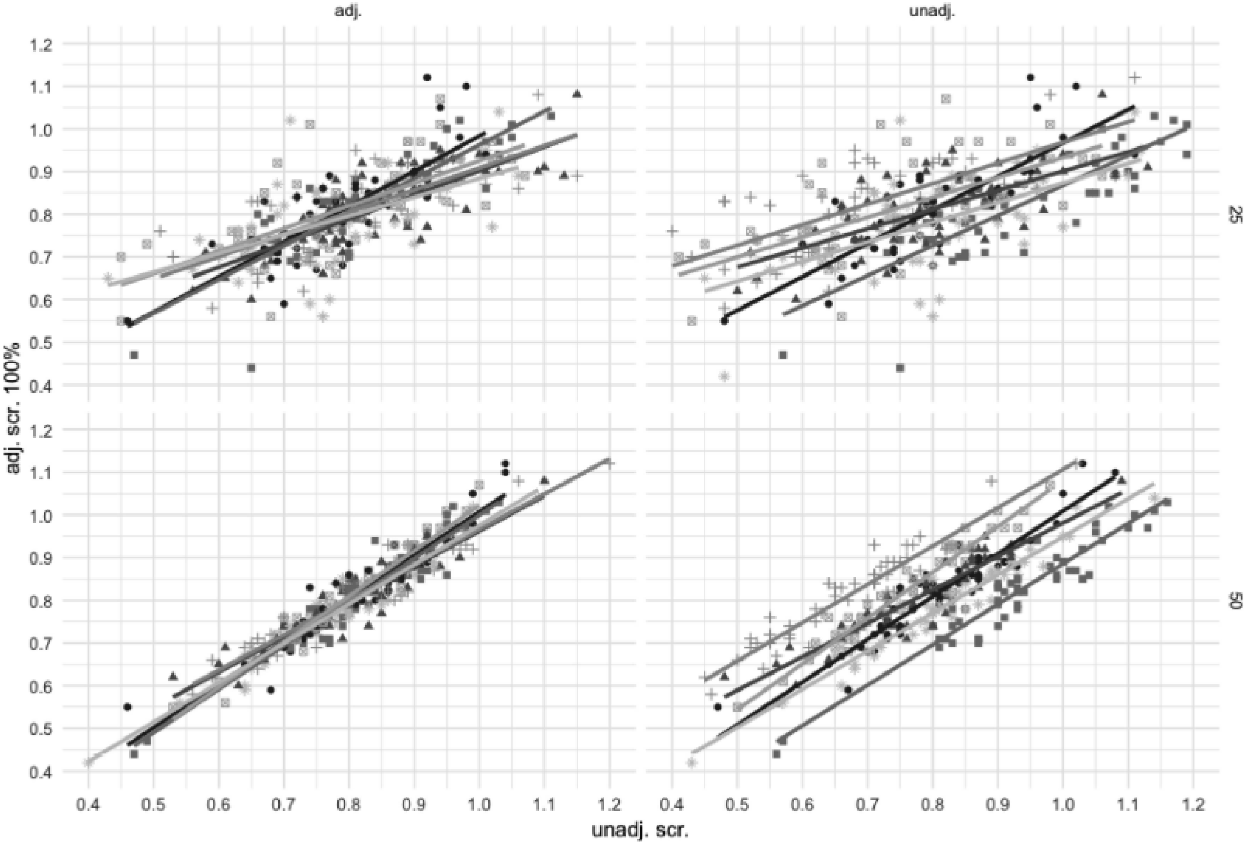


Figure 2. Comparing benchmark scores (adjusted score with full dataset adj. scr. 100%; Y-axis) with scores in four conditions of completeness and adjustment: adjusted 25%-dataset (upper-left), unadjusted 25%-dataset (upper-right), adjusted 50%-dataset (lower-left), unadjusted 50%-dataset (lower-right). The different shades and shape of the points indicate the six possible pairwise combinations of raters scoring the ideas (the six booklets). Six regression lines predict benchmark scores from incomplete data conditioned on particular pairwise combinations of raters.

Figure 2 illustrates the effects of incomplete data and score adjustments. It shows four scatterplots where points represent subjects' scores comparing benchmark scores with scores calculated with incomplete data. It can be seen that in the figures of the right column—the unadjusted scores—that the regression lines have different intercepts, suggesting that different combinations of raters map on different levels on the benchmark scores. In the figures of the left column—the adjusted scores—we see that these differences in intercept practically disappear, especially on the 50%-dataset where we have more data points. These figures show clearly that the improvements in approximating incomplete data scores to the benchmark scores is due to correction of varying levels of severity that follows when we combine different pairings of raters.

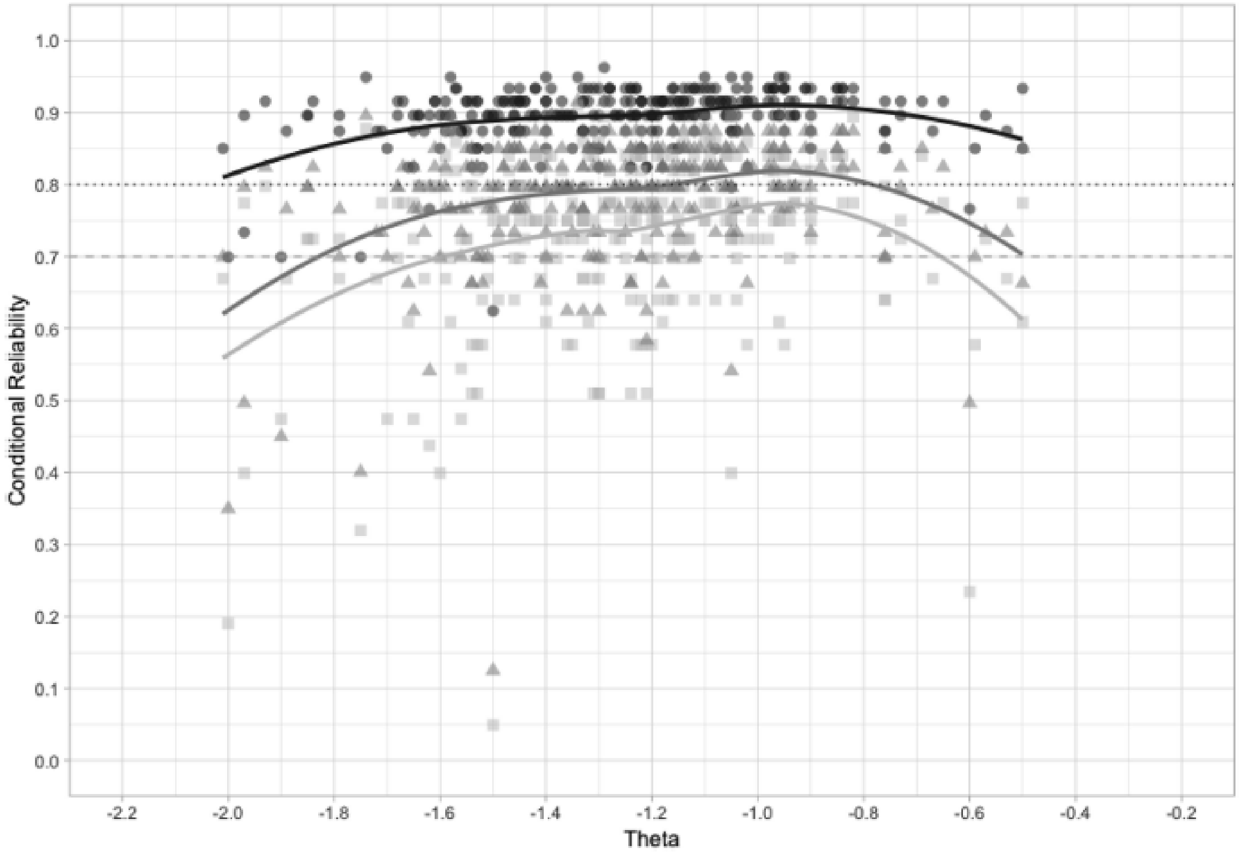
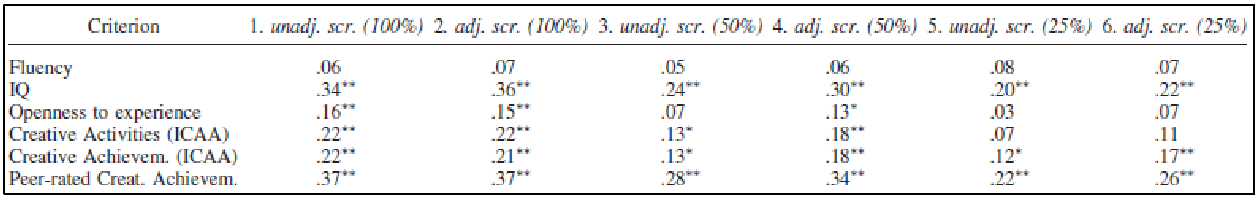


Figure 3. Conditional reliability of latent scores (theta) under three conditions of completeness (black: full data-set, dark-gray: 50%-dataset and light-gray: 25%-dataset). We rescaled results of the information function to the standard reliability metric of 0 to 1, and plotted smoothed average lines of points (i.e., subjects) in these three conditions.

How reliable are a subject's latent scores with incomplete data? In IRT models, reliability coefficients vary as function of latent ability because the particular mix of items may measure specific sectors of the latent scale better than others. In Figure 3 we show the information function plotting reliability coefficients (Y-axis) as a function of the latent scale (X-axis). Again, the full-dataset group is a benchmark to compare the other two conditions of incomplete data. Median reliabilities were good: full-dataset .90 (75% of points above .87), 50%-dataset .80 (75% of points above .77), and 25%-dataset .75 (75% of points above .70). Since the full dataset many data points per subject (*M* = 296) estimates of latent reliability are very high, but with half the data points we still reach high reliability. With one quarter of the data points the reliability is lower but isn't unacceptable.

Is test-criterion validity compromised when using shorter tests (fewer tasks and raters)? Do rater effects compromise validity? Table 5 presents correlations of unadjusted versus adjusted scores on three conditions of completeness with several criterion measures.

Table 5. Correlations of Scores Under Various Conditions of Completeness and Adjustment With Criterion Measures



Criterion	1. *unadj. scr. (100%)*	2. *adj. scr. (100%)*	3. *unadj. scr. (50%)*	4. *adj. scr. (50%)*	5. *unadj. scr. (25%)*	6. *adj. scr. (25%)*
Fluency	.06	.07	.05	.06	.08	.07
IQ	.34**	.36**	.24**	.30**	.20**	.22**
Openness to experience	.16**	.15**	.07	.13*	.03	.07
Creative Activities (ICAA)	.22**	.22**	.13*	.18**	.07	.11
Creative Achievem. (ICAA)	.22**	.21**	.13*	.18**	.12*	.17**
Peer-rated Creat. Achievem.	.37**	.37**	.28**	.34**	.22**	.26**

Note. Columns presents creativity measures: unadj. scr: average scores not adjusted for rater effects, adj. scr.: average scores adjusted for rater effects, that is, equated scores. The amount of data used to calculate scores varied across three levels: full dataset 100%, 50% and 25%. ICAA = Inventory of Creative Activities and Achievements (Diedrich et al., 2018).
* *p* < .05. ** *p* < .01. *** *p* < .001.

As expected, validity coefficients are lower when we have fewer data points. The more marked decrease occurs when we have a quarter of the original data points. But the decrease is higher for unadjusted scores as compared with adjusted scores. Specifically, with adjusted scores with 50% of data, the decrease in validity coefficients is minimal. In general, MFRM correction helps to maintain the criterion validity in a manner that is unaffected by potential differences in raters' severity.

In summary, this second study further explored reliability and validity of unadjusted and adjusted scores in conditions of missing data. We found a similar pattern of raters' α parameters variability and substantial variability in tasks' β difficulty parameters as in Study 1. MFRM adjustment helped to recover information that is obtained with a more extensive sampling of behavior, and criterion validity was higher for adjusted scores than unadjusted scores. When using half of the data available, the criterion validity coefficients were very similar to the coefficients calculated with full dataset. Finally, the reliability coefficients were good even with only half the data available.

General Discussion

Creativity assessment relies on scoring the quality of people's ideas. This creates a complex measurement structure involving persons, tasks (and ideas within tasks), and raters. Most studies, however, do not model possible systematic differences among raters. The present study examines the impact of individual rater differences in the reliability and validity of creativity assessments, and it explains how an MFRM approach can model and correct for these differences, especially in the context of incomplete data.

MFRM reveals systematic differences among raters that will bias subjects' creativity scores if not accounted for. Studies 1 and 2 showed that differences in leniency versus severity between raters and differences in task difficulty can be quite pronounced. Nevertheless, there was virtually no difference between observed and adjusted measures when using the full data (i.e., when all raters score all ideas). On the other hand, rater differences were influential when using incomplete data because combinations of raters will co-occur with subsets of subjects, thus confounding rater severity with subjects' scores. MRFM was an efficient method for controlling for rater differences and adjusting scores, thus removing this systematic error, making scores more convergent, and maintaining validity.

The benefits of MFRM become particularly apparent when using incomplete data designs. Removing 50% of data (Study 2) substantially affects reliability and validity evidence for unadjusted scores, whereas adjusted scores appear largely unaffected. This effect is similar but less pronounced for designs with 33% missing data (Study 1). However, when introducing missing data beyond 50% (Study 1: 43%, Study 2: 75%), deviations from benchmark scores are substantial. The lack of convergence between benchmark scores and incomplete data scores could be due to two main reasons: (a) a general effect of a reduced sample of behavior, like a test with fewer items (Primi, 2012), and (b) raters' systematic errors of measurement confounded with subjects' scores. Our findings suggest that both factors play a role but adjustments for variability in rater and task facets by means of MFRM partly compensate for the loss due to incomplete data.

MFRM also offers new ways of quantifying reliability. In divergent thinking tasks, we usually have a sample of a subject's ideas prompted by a task. If we consider each idea as an "item," since we have a variable number of ideas per subject, we have a situation as if subjects had answered different tests composed by a variable number of items. Then we have raters who score each idea. The usual way to quantify reliability is to calculate some sort of interrater reliability (Stemler, 2004). This method quantifies only one facet of this complex design, namely, how similar the scores of two independent raters of a given idea are. But the final quantity of interest is the subject's estimated creativity score, which is based in a large amount of data points from the aggregation of rater, task, and response facets. MFRM uses the general IRT concept of information function (de Ayala, 2009) to compute model-based reliabilities for all elements within facets.

Thus, MFRM provides a way to estimate reliability in creativity assessment for the facet we are most interested in, that is, the subject's ability parameter. In Study 2, for instance, we found that interrater reliability ranged from fair to moderate (Kappa indices ranged from .17 to .42). But we found that the Rasch reliability of subjects' creativity scores was excellent, around .90. These two coefficients address different sources of error in the global process of measurement. Kappa quantifies the extent to which two raters will give exactly same score for the same set of ideas. It is a consensus interrater reliability coefficient (Stemler, 2004). Rasch subjects' reliabilities indicate the confidence (inverse of the error of measurement) of the point estimate of a given subject's estimate of creative ability. Because each subject provides several ideas, we have an extended sample of behavior that are scored by raters. Therefore, when aggregating rater scores to estimate a person's creative ability, the reliability of this subject's point estimate will be built from raters' agreement. If we had high kappa coefficients, we could have achieved the same level of Rasch reliability with fewer data points (fewer ideas and less rater effort). Because of that, MFRM yields higher reliability estimates as compared to the reliability of raters alone.

Limitations and Future Directions

The findings suggest that it is feasible to reduce subjects' and raters' load, but our study was based on post hoc simulation of missingness. One open question is how our findings generalize to conditions where we actually have a reduced sample of responses and raters. It seems possible that reducing the amount of ratings per rater may reduce fatigue effects (Forthmann et al., 2017) and thereby even yield more reliable ratings. Another limitation is that we haven't considered interaction effects between facets. For instance, the interaction of rater by rubric (rating scale) examines if a particular rater uses the scoring rubric in same way as other raters, that is, addresses response styles. Another important interaction is rater by group, that is, rater differential functioning with respect to a particular group. These interaction effects need to be explored, especially in situations where we have planned missing data designs.

Supplemental Materials

Supplemental materials are available at https://doi.org/10.1037/aca0000230.

References

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

Barbot, B., Tan, M., Randi, J., Santa-Donato, G., & Grigorenko, E. L. (2012). Essential skills for creative writing: Integrating multiple domain-specific perspectives. *Thinking Skills and Creativity*, *7*, 209–223. 10.1016/j.tsc.2012.04.006

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. 10.1037/h0026256

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: Non-expert raters and the Creative Solution Diagnosis Scale. *The Journal of Creative Behavior*, *46*, 119–137. 10.1002/jocb.9

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York, NY: Guilford Press.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*, 35–40. 10.1037/a0038688

Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018). Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts*, *12*, 304–316. 10.1037/aca0000137

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main, Germany: Peter Lang. 10.3726/978-3-653-04844-5

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, *61*, 50–55. 10.1037/0003-066X.61.1.50

Engelhard, E., Jr., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York, NY: Routledge.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139. 10.1016/j.tsc.2016.12.005

Fürst, G. (2018). Measuring creativity with planned missing data. *The Journal of Creative Behavior*. Advance online publication. 10.1002/jocb.352

Gamer, G., Lemon, J., Fellows, I., & Singh, P. (2015). *irr: Various Coefficients Interrater Reliability and Agreement (R package version 0. 84)* [Computer software]. Retrieved from http://cran.r-project.org/web/packages/irr/

Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, *24*, 345–357. 10.1080/10400419.2012.730331

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, *41*, 212–221. 10.1016/j.intell.2013.03.003

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The road to creative achievement: A latent variable model of ability and personality predictors. *European Journal of Personality*, *28*, 95–105. 10.1002/per.1941

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. Chicago, IL: Mesa Press.

Linacre, J. M. (2004). KR-20 or Rasch reliability: Which tells the “truth”?*Rasch Measurement Transactions*, *11*, 580–581.

Linacre, J. M. (2018). *Facets computer program for many-facet Rasch measurement (version 3.81.0)*. Beaverton, OR: Winsteps.com.
Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, *15*, 13–25. 10.1016/j.tsc.2014.10.004

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*, 231–241. 10.1037/aca0000086

Primi, R. (2012). Psicometria: Fundamentos matemáticos da Teoria Clássica dos Testes [Psychometrics: Mathematical foundations of classical test theory]. *Avaliação Psicológica*, *11*, 297–307.

Primi, R. (2014a). 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*, 461–474. 10.1037/a0038055

Primi, R. (2014b). Developing a fluid intelligence scale through a combination of Rasch modeling and cognitive psychology. *Psychological Assessment*, *26*, 774–788. 10.1037/a0036712

R Core Team. (2016). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. Retrieved from https://CRAN.R-project.org/package=TAM

Robitzsch, A., & Steinfeld, J. (2017). *Immer: Item response models for multiple ratings*. Retrieved from https://CRAN.R-project.org/package=immer

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, *60*, 101–139.

Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation*. New York, NY: Oxford University Press.

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence*, *40*, 343–351. 10.1016/j.intell.2012.02.005

Silvia, P. J., Winterstein, B. B., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., . . .Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85. 10.1037/1931-3896.2.2.68

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*, 1–11. http://PAREonline.net/getvn.asp?v=9&n=4

Tan, M., Mourgues, C., Hein, S., MacCormick, J., Barbot, B., & Grigorenko, E. (2015). Differences in Judgments of creativity: How do academic domain, personality, and self-reported creativity influence novice judges' evaluations of creative productions?*Journal of Intelligence*, *3*, 73–90. 10.3390/jintelligence3030073

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. Boston, MA: O'Reilly Media.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA.

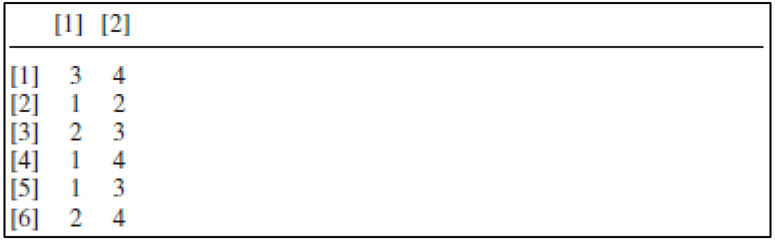
Wyse, A. E. (2013). Construct maps as a foundation for standard setting. *Measurement*, *12*(1–2), 62–68.

APPENDIX A. The Balanced Incomplete Block Design Combinations of Raters and Tasks

50%-BIB

Considering six raters we have six combinations that we call booklets. In each booklet each rater is combined with another. In Table A1 each row corresponds to booklet and each cell contains the number of the particular rater on the booklet. Subjects were randomly assigned to one of these six booklets, that is, to one of these combinations of two raters.

Table A1. Booklets 1. Six (Booklets) Combinations of Two Out of Four Raters



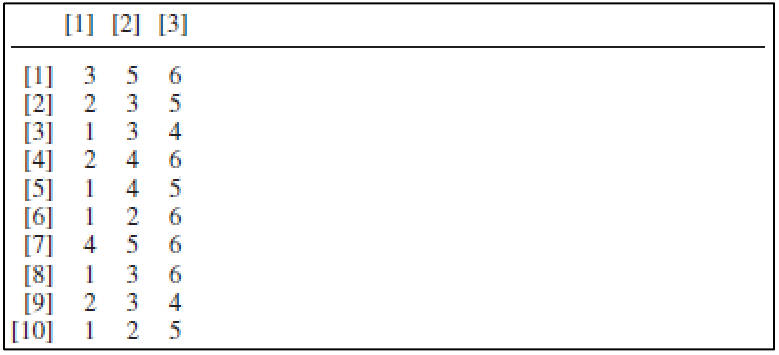
```
     [1]  [2]
[1]   3    4
[2]   1    2
[3]   2    3
[4]   1    4
[5]   1    3
[6]   2    4
```

25%-BIB

We further construed in the same way 10 combinations of three out of six tasks. In Table A2 each row corresponds to a booklet in which three out of six tasks are combined. Now each pairwise combination of any two tasks appear repeated in two booklets (for instance tasks 3 and 5 appear in booklets 1 and 2, tasks 5 and 6 in booklets 1 and 7 and so on).

BIB 25 were formed combining Booklets 1 and 2, resulting in 60 booklets. This is formed from six (combinations of raters) times 10 combinations (combinations of tasks). Therefore in 25%BIB each subject was randomly assigned to one of these combinations. Therefore in this BIB each subject will be scored by two raters on only three tasks.

Table A2. Booklets 2. Ten (Booklets) Combinations of Three Out of Six DT-Tasks



```
      [1] [2] [3]
 [1]    3   5   6
 [2]    2   3   5
 [3]    1   3   4
 [4]    2   4   6
 [5]    1   4   5
 [6]    1   2   6
 [7]    4   5   6
 [8]    1   3   6
 [9]    2   3   4
[10]    1   2   5
```