JOURNAL OF
Creative Behavior
Publication of the Creative Education Foundation™

NILS MYSZKOWSKI
MARTIN STORME

Accounting for Variable Task Discrimination in Divergent Thinking Fluency Measurement: An Example of the Benefits of a 2-Parameter Poisson Counts Model and its Bifactor Extension Over the Rasch Poisson Counts Model

ABSTRACT

Fluency tasks are among the most common item formats for the assessment of certain cognitive abilities, such as verbal fluency or divergent thinking. A typical approach to the psychometric modeling of such tasks (e.g., *Intelligence*, 2016, 57, 25) is the Rasch Poisson Counts Model (RPCM; *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960), in which, similarly to the assumption of (essential) τ-equivalence in Classical Test Theory, tasks have equal discriminations—meaning that, beyond varying in difficulty, they do not vary in how strongly they are related to the latent variable. In this research, we question this assumption in the case of divergent thinking tasks, and propose instead to use a more flexible 2-Parameter Poisson Counts Model (2PPCM), which allows to characterize tasks by both difficulty and discrimination. We further propose a Bifactor 2PPCM (B2PPCM) to account for local dependencies (i.e., specific/nuisance factors) emerging from tasks sharing similarities (e.g., similar prompts and domains). We reanalyze a divergent thinking dataset (*Psychology of Aesthetics, Creativity, and the Arts*, 2008, 2, 68) and find the B2PPCM to significantly outperform the 2PPCM, both outperforming the RPCM. Further extensions and applications of these models are discussed.

Keywords: fluency, divergent thinking, latent variable models, psychometrics.

A number of tasks used to measure cognitive abilities require examinees to generate as many productions (words, ideas, instances of a category, etc.) as possible, out of a (practically) infinite pool, in a fixed amount of time: These tasks are generally referred to as fluency tasks. Although the most famous instances of such tasks are probably verbal fluency tasks (Thurstone, 1938), fluency tasks can actually be found in several other domains. Notably, fluency tasks have been for a long time used as measures of general divergent thinking ability (Kim, 2011; Torrance, 1966; Wallach & Kogan, 1965) and have been adapted to more applied fields—for example to managerial (Myszkowski, Storme, Davila, & Lubart, 2015) and social creativity (Mouchiroud & Bernoussi, 2008). While a number of researchers (Hass, 2017a; Plucker, Qian, & Wang, 2011; Silvia et al., 2008) have discussed how alternative divergent thinking responses (e.g., the statistical rarity of the responses as a measure of originality, subjective evaluations of the productions, combinations of scoring approaches, semantic distance between the responses) could be used, we here focus on fluency scores—the count of productions—and their psychometric modeling.

In creativity research, fluency scores are counts of the number of ideas generated by participants across divergent thinking tasks. Consequently, modeling fluency scores commands a psychometric approach that is appropriate for count variables. Therefore, researchers (Forthmann et al., 2016) have suggested to use the Rasch Poisson Counts Model (RPCM; Rasch, 1960). Although the RPCM represents a major advance in the modeling of fluency in divergent thinking tasks, we aim here to question one of its limitations: The fact that it only accounts for item differences in difficulty, but not in discrimination. Discrimination parameters are more commonly referred to as loadings, weights, or slopes in the factor analysis tradition and are used in psychometric models to represent the strength and direction of the relation between an item and the latent

variable. The assumption that all items have the same discrimination is generally referred to as the assumption of (essential) τ-equivalence in Classical Test Theory (CTT). By using the RPCM, researchers assume that two items that are equally difficult have the same distribution conditional on person ability.

Our first aim with the present research is to discuss how this assumption is questionable from the point of view of item construction. Our second aim is to propose to use instead a 2-Parameter Poisson Counts Model (2PPCM)—an instance of Generalized Latent Trait Modeling (Moustaki & Knott, 2000; Rabe-Hesketh, Skrondal, & Pickles, 2004)—which allows to characterize items by both difficulty and discrimination. Third, we aim to demonstrate with an empirical example that including a discrimination parameter can improve the psychometric modeling of fluency scores in divergent thinking. Finally, we discuss situations where divergent thinking items are clustered—notably when some items have similar prompts (e.g., alternate uses, consequences)—and how to model them with a bifactor extension of the 2PPCM.

FLUENCY SCORES IN CREATIVITY RESEARCH

Since the seminal work of Guilford (1967), creativity researchers have regarded ideational fluency as one of the pillars of creative ability. Numerous empirical studies have shown indeed that there is a relation between the number of productions and their quality, whether it is at the level of the career of an eminent creator (Simonton, 2010), in group brainstorming tasks (Briggs & Reinig, 2010) or in divergent thinking tasks (Silvia et al., 2008). In the 1960s, Osborn (1963, p. 131) summarized one of most commonly accepted reasons why fluency should facilitate creativity: "the more ideas we produce, the more likely we are to think up some that are good."

The appeal of using fluency scoring as an indicator of creativity also lies in the fact that it is easy to operationalize from a methodological viewpoint. Researchers simply need to count the number of ideas or productions generated by a participant or a group during the task. The double advantage of a count is that (a) the implementation is straightforward and inexpensive—no need, for example, to recruit and train judges—and (b) it is objective—contrary to some other divergent thinking scoring methods that for some of them require raters.

Nevertheless, one cannot simply reduce creativity to fluency. This is the reason why creativity researchers have investigated alternate scoring strategies for divergent thinking tasks, beyond fluency scoring. For example, it has been recently advanced that creativity may be better captured through the semantic distance between an individual's responses (Hass, 2017b; Heinen & Johnson, 2018), or by using holistic ratings of all the ideas generated by an individual for a task (Silvia, Martin, & Nusbaum, 2009). However, many scoring methods might be contaminated by fluency (Clark, Griffing, & Johnson, 1989). For example, flexibility scores—that is, the count of the different conceptual categories from which ideas are generated—are usually criticized because they are intrinsically linked to fluency scores. The same is true for uniqueness scores—that is, the number of statistically rare ideas generated by a participant. The difficulty to disentangle fluency, flexibility, and uniqueness in divergent thinking is probably another reason for the success of fluency scoring in the study of divergent thinking.

WHY USING AN ACCURATE MEASUREMENT MODEL MATTERS

Because fluency scoring, albeit questioned, is an important approach to the measurement of creativity, using appropriate psychometric approaches to fluency scores is highly beneficial to the measurement and study of creativity, for various reasons. First, by using a measurement model that allows for items varying in difficulty and discrimination separately, we can better study the relative importance of items. In other words, using such a model, the researcher is informed about the items that represent best the latent variable measured (controlling for difficulty) in a given item set. Second, using a better psychometric model improves a researcher's accuracy in evaluating an examinee's ability. In other words, better measurement models allow to better achieve the measurement of individual differences. Finally, better measurement models can be more practical, and, as we will later show in our supplementary analysis, allowing divergent thinking items to differ in discrimination notably allows to build models that account for specific/nuisance factors—which in our example dataset, we use to account for different types of prompts (alternate uses, consequences, etc.).

THE 2-PARAMETER POISSON COUNTS MODEL (2PPCM)

In contrast with models for binary, ordinal, and normally distributed responses, for which a wealth of item response models and extensions have been developed (Shao, Janse, Visser, & Meyer, 2014), the leading

response model used for fluency scores for modeling fluency scores remains (Baghaei & Doebler, 2019; Forthmann, Celik, Holling, Storme, & Lubart, 2018; Forthmann et al., 2016) the Rasch Poisson Counts Model (RPCM; Rasch, 1960). However, because the RPCM is a particular case of the 2-Parameter Poisson Counts Model (2PPCM)—the model which we propose as a better alternative—we will first introduce the 2PPCM for clarity.

Probability distribution

Perhaps, the simplest way of introducing the 2PPCM is to discuss it as a Generalized Linear Item Response Theory model (Mellenbergh, 1994), in which the link function is the natural logarithm function and the item response distribution is a Poisson distribution of rate parameter λ_{ij}. More specifically, the 2PPCM—like its special case the Rasch Poisson Counts Model and its extension the Bifactor 2PPCM (both later discussed)—formulates that, for a person i and an item j, the probability $P(X_{ij} = k)$ that a fluency score X_{ij} is equal to k (a non-negative integer), is a function of the rate parameter λ_{ij} (comprised between 0 and $+\infty$), such as:

$$P(X_{ij} = k) = \frac{\lambda_{ij}^{k} e^{-\lambda_{ij}}}{k!} \tag{1}$$

Unlike the Normal distribution used in linear models, which has a continuous unbounded support, the support of the Poisson distribution consists of non-negative integer numbers, and therefore matches the observable modalities of fluency item scores. In addition, unlike in linear models where the variance of the Gaussian errors is assumed to be constant (homoscedasticity), the Poisson distribution allows for a degree of heteroscedasticity often observed in count data (since larger counts are often associated with larger error variance). More specifically, in the Poisson distribution, the variance is equal to the rate (equidispersion).

Response model

In the 2PPCM, the rate λ_{ij} is modeled as a function of the person's latent ability θ_i—which, in this parametrization is of variance fixed to 1 and mean fixed to 0—the item's difficulty parameter b_j (freely estimated) and the item's discrimination parameter a_j (freely estimated), as given in the following equation:

$$\lambda_{ij} = e^{b_j + a_j \theta_i} \tag{2}$$

Equation 2 can also be written as a generalized linear model using the natural logarithm as the link function, which means that the 2PPCM can be called a log-linear model. It can be noted that, at constant discrimination and for a given person, a higher b_j implies higher expected counts. For this reason, we may refer to b_j as an easiness parameter rather than a difficulty parameter, as is often suggested. At constant easiness, a higher item discrimination a_j implies that an increase in the latent ability θ_i results in larger increases in expected counts. The discrimination parameter therefore represents the strength of association between the latent ability and the expected count (for a given item difficulty). Finally, let us note that the formulation in Equation 2 is simplified to consider that all items have the same exposure (or *offset*), meaning that, in the context of fluency items, all items have the same time limit. The 2PPCM can however easily be adjusted for situations where items have different exposure parameters (in this case, different time limits), through a variable exposure term, in a manner similar to the RPCM (see Baghaei & Doebler, 2019).

It should be noted that the 2PPCM is not per se a new model, in that it can notably be considered as an instance of Generalized Linear Item Response Theory (GLIRT; Mellenbergh, 1994) or of Generalized Linear Latent Mixed Models (GLAMM; Rabe-Hesketh & Skrondal, 2016; Rabe-Hesketh et al., 2004). In addition, the use of factor structures in Poisson counts modeling has been previously discussed in a framework proposed by Wedel, Böckenholt, and Kamakura (2003), under which the 2PPCM is subsumed. However, no formal name has been proposed for this model. Since it is traditional in IRT to refer to models by how many item parameters they comprise (e.g., the 2-parameter logistic model), we propose to refer to the model presented in equation 2 as the 2-Parameter Poisson Counts Model (2PPCM).

Estimation

Unfortunately, only a few software packages allow to estimate the 2PPCM. Indeed, on one hand, many Generalized Linear Mixed Models (GLMM) packages, such as "lme4" (Bates, Mächler, Bolker, & Walker, 2015), cannot estimate discrimination parameters, and therefore cannot estimate the 2PPCM ; on the other

hand, many general purpose Structural Equation Modeling (SEM) software packages—for example, "lavaan" (Rosseel, 2012)—can estimate discrimination parameters, but do not allow for Poisson distributed responses, thus not allowing to fit the 2PPCM. Fortunately, the 2PPCM can still be estimated by the few (Generalized) SEM packages that allow for Poisson counts models—such as Mplus (Muthén & Muthén, 1998) and Stata (StataCorp, 2017)—and the few GLMM packages that allow to estimate discrimination parameters—such as the "NLMIXED" package for SAS (see Sheu, Chen, Su, & Wang, 2005), and the packages "PLmixed" (Jeon & Rockwood, 2018) and "brms" (Bürkner, 2017) for R.

Special case: The Rasch Poisson Counts Model (RPCM)

The Rasch Poisson Counts Model (RPCM) can be seen as a special case of the 2PPCM, where all discrimination parameters are constrained to be equal. By constraining all discrimination parameters to be equal, we obtain an equation for the rate λ_{ij} that is similar to that of Equation 2, except that there is now only one discrimination a for all items:

$$\lambda_{ij} = e^{b_j + a\theta_i} \tag{3}$$

An alternative identification consists in freeing the latent variance while constraining the common slope parameter a to 1 (so as to practically remove it from the equation). While Generalized SEM software like Mplus or Stata allow for both formulations, this alternative formulation allows to estimate the RPCM using most GLMM software packages, including "lme4" (Baghaei & Doebler, 2019). Another alternative for parametrization concerns the latent mean, which can be freely estimated instead of fixed, provided that a constraint is added on the b_j difficulty parameters. Besides being estimable using more software packages, the RPCM also presents characteristic advantages of Rasch modeling that are lost when using variable discrimination parameters, such as specific objectivity and allowing for the use of person total scores as sufficient statistics for the estimation of ability (Masters & Wright, 1984; Rasch, 1960).

Extension: The bifactor 2PPCM (B2PPCM)

The 2PPCM (like the RPCM) is based on an assumption typical to latent trait models, known as the assumption of local (or conditional) independence. This assumption states that the item scores should not be related beyond them being indicators of the same latent trait (here divergent thinking fluency). In the case of divergent thinking tasks, it is however frequent to use tasks that share similarities beyond being fluency tasks. More specifically, some tasks may share similarities in their prompts (e.g., instances of a category or alternate uses of an object) or domains (e.g., figural or verbal), resulting in local dependencies between tasks that share such similarities.

Because of these potential violations, we propose a bifactor extension of the 2PPCM, which we refer to as the Bifactor 2PPCM (B2PPCM), to account for local dependencies. This model, based on the bifactor modeling approach (Holzinger & Swineford, 1937; Reise, 2012), differs in structure from the previously discussed (unidimensional) 2PPCM, in that, in this model, in addition to the general latent ability, additional specific factors are specified to represent communalities between tasks (e.g., prompts, domains). In the B2PPCM, for a person i and an item j, the rate parameter of the Poisson distribution λ_{ij} is modeled as a function of the person's latent attribute on the general factor θ_i, the person's latent attribute on the specific factor θ'_i, the difficulty/intercept item parameter b_j, the general discrimination/slope parameter a_j, and the specific discrimination/slope parameter a'_j, as showed in Equation 4. All correlations between latent factors (general and specific) are fixed to 0.

$$\lambda_{ij} = e^{b_j + a_j\theta_i + a'_j\theta'_i} \tag{4}$$

The B2PPCM is a structural extension of the 2PPCM that can be used to represent task clusters. We provide an example use in this paper where we use 3 specific factors to represent clusters of tasks that share similar prompts (alternate uses, instances of a category and consequences).

WHY WOULD WE IGNORE ITEM DISCRIMINATION DIFFERENCES IN DIVERGENT THINKING TASKS?

The RPCM is undoubtedly the most discussed response model for count data. However, in the RPCM, there are no (free) item discrimination parameters, meaning that items are only characterized by one parameter, representing their easiness. Why would creativity researchers rely on this assumption in the case of

Poisson measurement models? We here tentatively advance several reasons, and discuss their legitimacy in the case of divergent thinking tasks.

Wrong reason #1: It is traditional to focus on difficulty only

The IRT tradition primarily emphasizes how item difficulty should be first considered in the modeling of item responses and the estimation of psychological constructs. This translates into how some modeling strategies only consider 1-parameter Rasch models, and into how most model comparisons strategies start with these models, before sequentially adding parameters—generally starting with a slope/discrimination parameter (Birnbaum, Lord, & Novick, 1968).

Still, while Rasch models have undoubtedly revolutionized psychometric research, the fact that Rasch models are historically prominent is not a valid argument, especially if we consider that, for binary items, 1-parameter models are regularly empirically outperformed by models that are more flexible, especially models that account for item differences in discrimination (e.g., Storme, Myszkowski, Baron, & Bernard, 2019). Further, in the case of more traditional (linear) measurement models, the assumption that items do not vary in discrimination/loadings is generally regarded as simply not realistic (Trizano-Hermosilla & Alvarado, 2016), as illustrated by the common use of factor analysis.

Wrong reason #2: Most GLMM estimation packages do not allow otherwise

As we previously discussed, the RPCM can be estimated using most GLMM estimation packages (Baghaei & Doebler, 2019), while the 2PPCM can be estimated by fewer packages. Over the years, GLMM estimation packages have become increasingly popular and available, notably with the R package "lme4" (Bates et al., 2015), making Rasch models more commonly implemented. Still, the lack of availability (and ease of use) of packages, although problematic, should not limit the development or use of response models. Besides, as previously discussed, a growing number of packages are capable of estimating Poisson counts models with a discrimination parameter.

Wrong reason #3: Items are equally discriminant

The Rasch measurement tradition does not focus so much on fitting the data, but more so on building instruments consistent with Rasch models—in the case of the RPCM, this implies creating fluency tasks with equal discrimination. Certainly, in some situations, it is reasonable to assume that different fluency tasks could be equally discriminant theoretically, because we can clearly define the latent variable being measured and affirm that the items are equally distant to it. For example, if one creates a test aiming to measure verbal fluency, it can be theoretically sustained that generating words that start with the letter *c* is equally related to verbal fluency as generating words that start with the letter *g*, and thus that the two tasks, by construction, do not differ in discrimination—only in difficulty —thereby building an instrument that meets the assumptions of Rasch modeling.

However, is this a realistic expectation of tasks like divergent thinking tasks? If one creates a test aiming to measure divergent thinking fluency, then one would likely use various prompts, such as prompts for alternate uses of an object and prompts for instances of a category. These variations in types of prompts have been discussed as tapping into different cognitive processes (Hass & Beaty, 2018), making it unlikely that two items with different prompts would reflect equally a common latent variable. Further, even with the same prompt, there are important differences between tasks. If we take the example of two tasks prompting for alternate uses of a knife or a brick (used in the present paper), the two tasks may tap into different domains of expertise—speculatively, masonry for the brick and cooking or camping for the knife. Further, these domains of expertise may be more or less present for different tasks—following our example, perhaps the expertise effect of masonry could be less than the effect of cooking, since bricks could have fewer uses in masonry than knives in cooking. In addition to this, the two tasks may engage different personality traits. Using our previous brick and knife examples, individuals with higher social inhibition may see the expression of their creativity limited with the knife prompt (they may, for example, refrain from responding with violent uses of a knife), but less so with the brick prompt. As a consequence, we argue that there are so many substantial item specificities (from the domain, the type of prompt, etc.) in divergent thinking tests and variations in how different tasks engage psychological attributes (expertise, personality, cognitive abilities), that it is unrealistic to affirm (without empirical investigation) that a set of divergent thinking fluency tasks equally reflect a common latent variable—in other words, that they are have equal discriminations. Therefore, we argue that researchers should use a model that does not make this

assumption—such as the 2PPCM—or at least examine whether this assumption is supported empirically—which can be done formally by comparing the fit of the RPCM and the 2PPCM.

The aim of this paper

To provide an empirical example of the benefits of a 2-Parameter Poisson Counts Model (2PPCM) over the RPCM, we reanalyzed a publicly available prototypical dataset with several fluency task scores—that were not necessarily assumed to differ (or to not differ) in discrimination—and a reasonably large sample size. In this dataset, we compare the fit of various models to the data, including a Poisson baseline model where items are perfectly interchangeable/parallel, the RPCM and the 2PPCM. We hypothesized that the 2PPCM would outperform the RPCM, which would itself outperform the baseline model. In addition, we discuss how the B2PPCM can be used to detect and account for local item dependencies.

METHOD

DATASET

We used the data from a previous research effort on creative cognition (Silvia, 2008a, 2008b; Silvia et al., 2008), reused with permission from the author. The dataset consisted of the responses of 242 college students to 6 divergent thinking fluency tasks. The participants were successively asked to provide as many creative (a) uses for a brick, (b) round objects, (c) effects if people no longer had to sleep, (d) uses for a knife, (e) things that make a noise, and (f) consequences if everyone shrank to be 12-inches tall. The participants were encouraged to give as many creative responses as possible and had 3 minutes for each task.

STATISTICAL ANALYSIS

The aim of this paper being to demonstrate how a 2-parameter extension of the RPCM is a more accurate item response modeling strategy than existing alternatives, the statistical analysis for this paper consisted in fitting different item response models and comparing their fit to the data. These models were (a) a baseline Poisson model—which essentially ignores item differences, since in this model all items are constrained to have the same easiness and discrimination—(b) the RPCM—in which the items are allowed to vary in easiness but not discrimination—and (c) the 2PPCM—in which items are allowed to vary in both easiness and discrimination.

In addition, we explored for the presence of specific/nuisance factors using the B2PPCM. The B2PPCM was initially specified with 3 specific factors, representing similarities between prompts (alternate uses, instances, and consequences). Based on the estimates of the B2PPCM, a simpler bifactor model—referred to as the B2PPCM'—was estimated, with only one specific factor corresponding to instances.

As is traditionally done in IRT modeling, the variance of θ_i (and specific factors θ'_i for the bifactor models) was constrained to 1 to identify all models. Although we previously discussed alternative parametrizations, this choice was motivated by the fact that it allows to interpret θ_i on the standard Normal scale—as one would interpret a z-score.

MODEL ESTIMATION

All models were fit to the data using Mplus 8.4 (Muthén & Muthén, 1998). The syntax for fitting the models is provided as Appendix S1, along with the data set formatted in wide format, and prepared for direct analysis.

MODEL COMPARISONS

The models were compared using Likelihood Ratio Tests as well as information criteria—we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for which a smaller value indicates a better fit. These procedures for model fit comparisons are typically used in Item Response Theory modeling, including for Poisson counts models (Baghaei & Doebler, 2019; Forthmann, Gühne, & Doebler, 2019).

Based on the AIC and BIC of all models, we also computed the difference between the smallest AIC (and BIC) and the AIC (and BIC) of each model (often referred to as Δ AIC and ΔBIC), as well as the AIC weights (also known as Akaike Weights) and BIC weights (also known as Schwartz weights) for further interpretation. Although these procedures are more extensively discussed in the multimodel inference literature (Burnham & Anderson, 2004; Wagenmakers & Farrell, 2004), ΔAIC and ΔBIC values below 2 indicate substantial support for the model, values between 2 and 4 indicate strong support for the model, values

between 4 and 7 indicate considerably less support for the model, and values above 10 indicate essentially no support. The model with the smallest AIC (or BIC) has a trivial ΔAIC (or ΔBIC) of 0. AIC (and BIC) weights sum to 1, and give an estimate of the relative evidence in favor of each model in the candidate set of models—values close to 1 indicating strong support for the model, and values close to 0 indicating that the model is not supported. Finally, as recommended by Baghaei and Doebler's (2019) for the RPCM, we also used boxplots of the Pearson residuals for the RPCM and the 2PPCM for each item, good item-model fit being indicated by that residuals roughly contained between −2 and +2.

MODEL INTERPRETATION

Plotting item response functions is central to the understanding of item response models (Muraki, 1993). Item response function plots essentially graph the expected response—in binary logistic IRT, it is typically the expected probability of passing an item, but here, it is the expected count of ideas—as a function of the latent trait estimate—which, as we pointed, is typically scaled similarly to a z-score. We plotted item response functions for all the items, overlaying the RPCM and 2PPCM for comparison.

EQUIDISPERSION

In Poisson models, it is assumed that the variance and the expectation are equal, an assumption referred to as *equidispersion*. When the error variance is smaller than the expectation—a phenomenon called *underdispersion*—then the standard errors and confidence intervals of the model are conservative (Forthmann et al., 2019). A bigger—and more actively debated—problem occurs when there is more variability than expected—a phenomenon called *overdispersion*. When overdispersion occurs, then the standard errors and confidence intervals of the model are not conservative enough—in other words, the estimates appear more reliable than they actually are. To assess dispersion, we computed the ϕ coefficient, which is the ratio of the model implied variance to the expectation of the mean, using the formula provided in Baghaei and Doebler's (2019) paper on estimating the RPCM. A value of 1 indicates equidispersion; a value lower than 1 indicates underdispersion; a value higher than 1 indicates overdispersion. As suggested by the authors, we also used scatterplots of the Pearson residuals as a function of the scores predicted by the models. As reported previously regarding the RPCM (Doebler & Holling, 2016), we also computed bootstrapped percentile 95% confidence intervals (with 5,000 resamples) for the ϕ coefficient.

RELIABILITY

A key difference between CTT and IRT is that IRT conceptualizes reliability as conditional upon the latent attribute θ_i. A corollary of this is that different persons have different reliability estimates. Using the person estimates and their standard errors, we computed person-reliability ρ_{θ_i} and group-level reliability ρ_θ (computed as the average of the observed person-reliabilities ρ_{θ_i}). We then plotted the conditional reliability of the RPCM and 2PPCM person estimates as a function of the latent trait. Similar to previous research regarding IRT reliability estimates (Myszkowski & Storme, 2017, 2018), and similar to what was done in the present study regarding overdispersion, we computed bootstrapped percentile 95% confidence intervals (with 5,000 resamples) for group-level reliability estimates. Since reliability is based on which model is selected, we also computed a model-averaged reliability estimate, consisting in an average of the reliability estimates, weighted by the model (Akaike and Schwartz) weights previously computed. We refer to these estimates as the Akaike model-averaged reliability and the Schwartz model-averaged reliability.

RESULTS

MODEL FIT

The AIC and BIC of all models, along with their associated weights, are reported in Table 1. As hypothesized, the RPCM fit significantly better than the baseline model ($\chi^2(5) = 485.59$, $p < .001$). Because the only difference between the baseline model and the RPCM is that the RPCM accounts for item differences in difficulty, this result implies that items significantly differed in difficulty. Boxplots of the distribution of the Pearson residuals per item are presented in Figure 1. For the RPCM, residuals were for the most part between −2 and +2, indicating adequate fit.

The item parameters for the RPCM are presented in Table 2. The significance of the item differences in difficulty is also reflected in the non-overlap between the confidence intervals of certain intercept estimates —for example the confidence intervals of the intercepts of items 1 and 2 do not overlap. In this case, it means that item 2 was significantly easier than item 1.

TABLE 1. Model fit statistics

Model	Log-likelihood	AIC	ΔAIC	*w*(AIC)	BIC	ΔBIC	*w*(BIC)
Baseline	−3952.629	7909.26	585.87	0.00	7916.24	544.00	0.00
RPCM	−3709.835	7433.67	110.28	0.00	7458.09	85.86	0.00
2PPCM	−3657.525	7339.05	15.66	0.00	7380.92	8.69	0.01
B2PPCM	−3647.695	7331.39	8.00	0.02	7394.19	21.96	0.00
B2PPCM'	−3647.693	7323.39	0	0.98	7372.23	0	0.99

AIC: Akaike Information Criterion of the model considered; ΔAIC : Difference between the AIC of the model considered and the smallest AIC; w(AIC) : Akaike weight of the model considered (rounded); BIC : Bayesian Information Criterion of the model considered; ΔBIC : Difference between the BIC of the model considered and the smallest BIC; w(AIC) : Schwartz weight of the model considered (rounded).

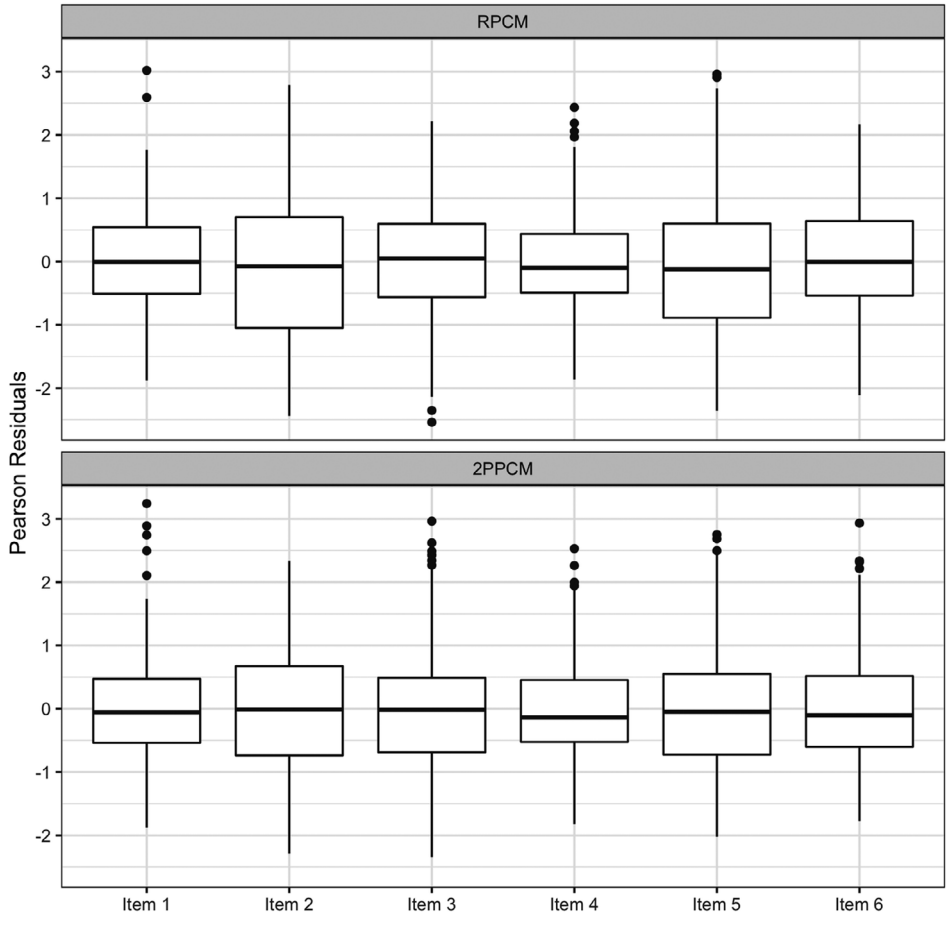

FIGURE 1. Distribution of the residuals of the RPCM and 2PPCM.

TABLE 2. Parameter estimates of the RPCM

Item	Parameter	Estimate	*SE*	*z*	*p*	95% CI
1	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	1.919	0.032	59.422	<.001	[1.856;1.982]
2	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	2.140	0.030	70.170	<.001	[2.080;2.199]
3	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	1.847	0.033	56.030	<.001	[1.782;1.911]
4	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	1.887	0.033	57.914	<.001	[1.823;1.951]
5	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	2.364	0.029	81.680	<.001	[2.307;2.420]
6	Slope (all constrained equal)	0.333	0.018	18.131	<.001	[0.297;0.369]
	Intercept	1.820	0.033	54.822	<.001	[1.755;1.886]

As hypothesized, whether the Likelihood Ratio Tests or the information criteria were considered, the 2PPCM significantly outperformed the RPCM – $\chi^2(5) = 104.62$, $p < .001$. Because the only difference between the two models is that, in addition to differences in difficulty, the 2PPCM accounts for item differences in discrimination, this model comparison can be used as a formal test of the RPCM assumption that items are equally discriminant. Here, its result implies that this assumption is violated—in other words, items significantly differed in discrimination. For the 2PPCM, residuals were distributed similarly to those of the RPCM and were also for the most part between −2 and + 2, indicating good fit.

The item parameters for the 2PPCM are presented in Table 3. The significance of the item differences in discrimination is also reflected in the non-overlap between the confidence intervals of certain slope estimates.

ITEM RESPONSE FUNCTIONS

The item response functions, which present the expected fluency scores as a function of the latent ability, are presented in Figure 2. They show the differences between the RPCM and the 2PPCM. As can be expected due to the exponential form of the two models, the largest differences are observed at high levels of the latent trait. For example, for a person with a θ level of 3 (as previously mentioned, due to how the models were identified, this can be interpreted like a z-score of 3), the 2PPCM predicts that the person would produce 39 ideas for item 5, while the RPCM would predict 29 ideas. For the same latent level, the 2PPCM would predict on average 12 ideas for item 3, while the RPCM would predict on average 19 ideas.

TABLE 3. Parameter estimates of the 2PPCM

Item	Parameter	Estimate	*SE*	*z*	*p*	95% CI
1	Slope	0.255	0.028	9.024	<.001	[0.199;0.310]
	Intercept	1.943	0.030	65.352	<.001	[1.884;2.001]
2	Slope	0.467	0.033	14.171	<.001	[0.402;0.531]
	Intercept	2.088	0.039	54.230	<.001	[2.013;2.164]
3	Slope	0.199	0.028	7.004	<.001	[0.143;0.255]
	Intercept	1.883	0.028	66.225	<.001	[1.827;1.939]
4	Slope	0.332	0.030	10.986	<.001	[0.273;0.392]
	Intercept	1.888	0.033	56.382	<.001	[1.823;1.954]
5	Slope	0.450	0.031	14.745	<.001	[0.390;0.510]
	Intercept	2.320	0.036	64.426	<.001	[2.249;2.390]
6	Slope	0.211	0.028	7.418	<.001	[0.156;0.267]
	Intercept	1.854	0.029	63.629	<.001	[1.797;1.911]

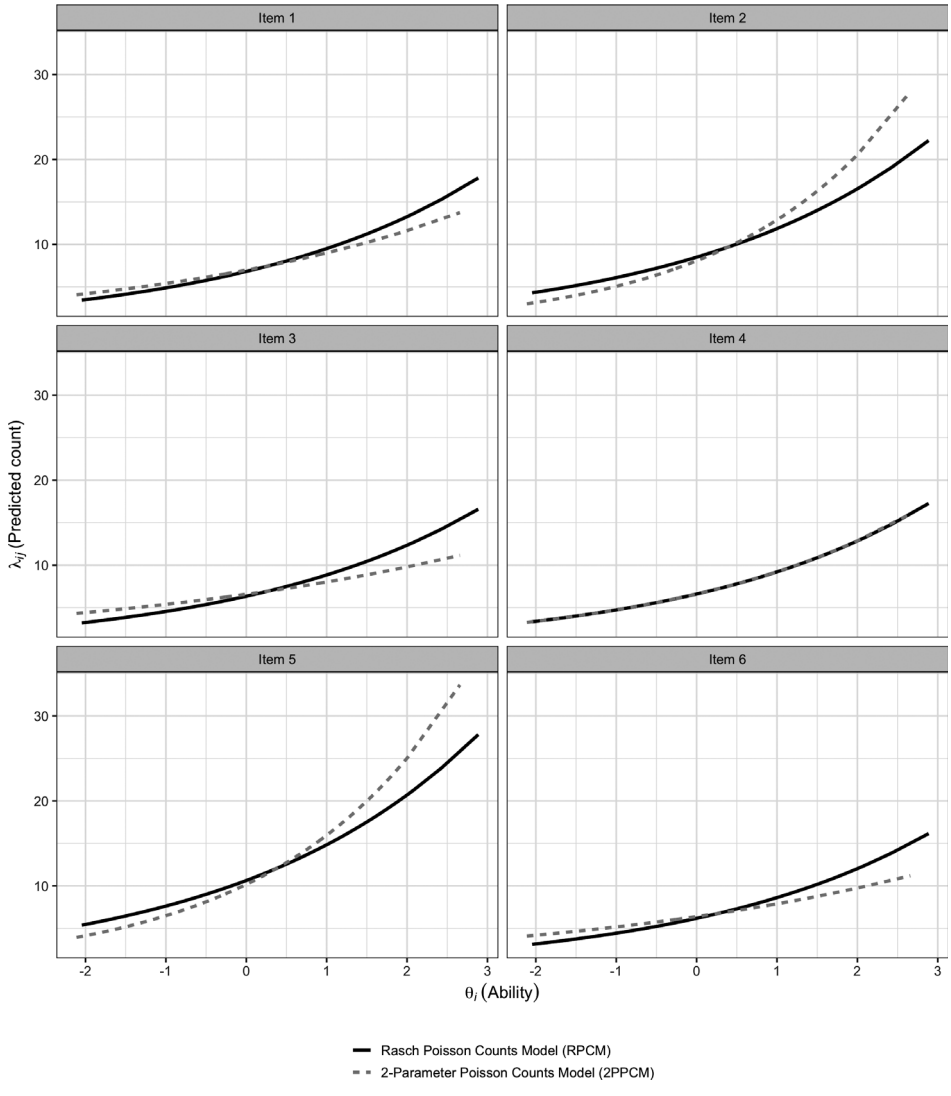

FIGURE 2. Item response functions (predicted counts) of the RPCM and the 2PPCM.

EQUIDISPERSION

The baseline model was overdispersed ($\phi_{baseline} = 1.27, 95\%CI[1.15, 1.40]$), while both the RPCM ($\phi_{RPCM} = 0.85, 95\%CI[0.79, 0.91]$) and the 2PPCM ($\phi_{2PPCM} = 0.78, 95\%CI[0.72, 0.84]$) showed underdispersion. Figure 3 shows the Pearson residuals as a function of the predicted for each item in both the RPCM and 2PPCM, and confirms this finding. This underdispersion implies that both the reliability estimates of the RPCM and 2PPCM are inaccurate, but likely conservative.

RELIABILITY

Overall, both models yielded reliable person estimates, with a slightly higher reliability for the 2PPCM (ρ_θ=.813, 95% CI [.806, .820]) as opposed to the RPCM (ρ_θ = .795, 95% CI [.789, .801]). In Figure 4, we present a plot of the reliability estimates for the RPCM and 2PPCM models.

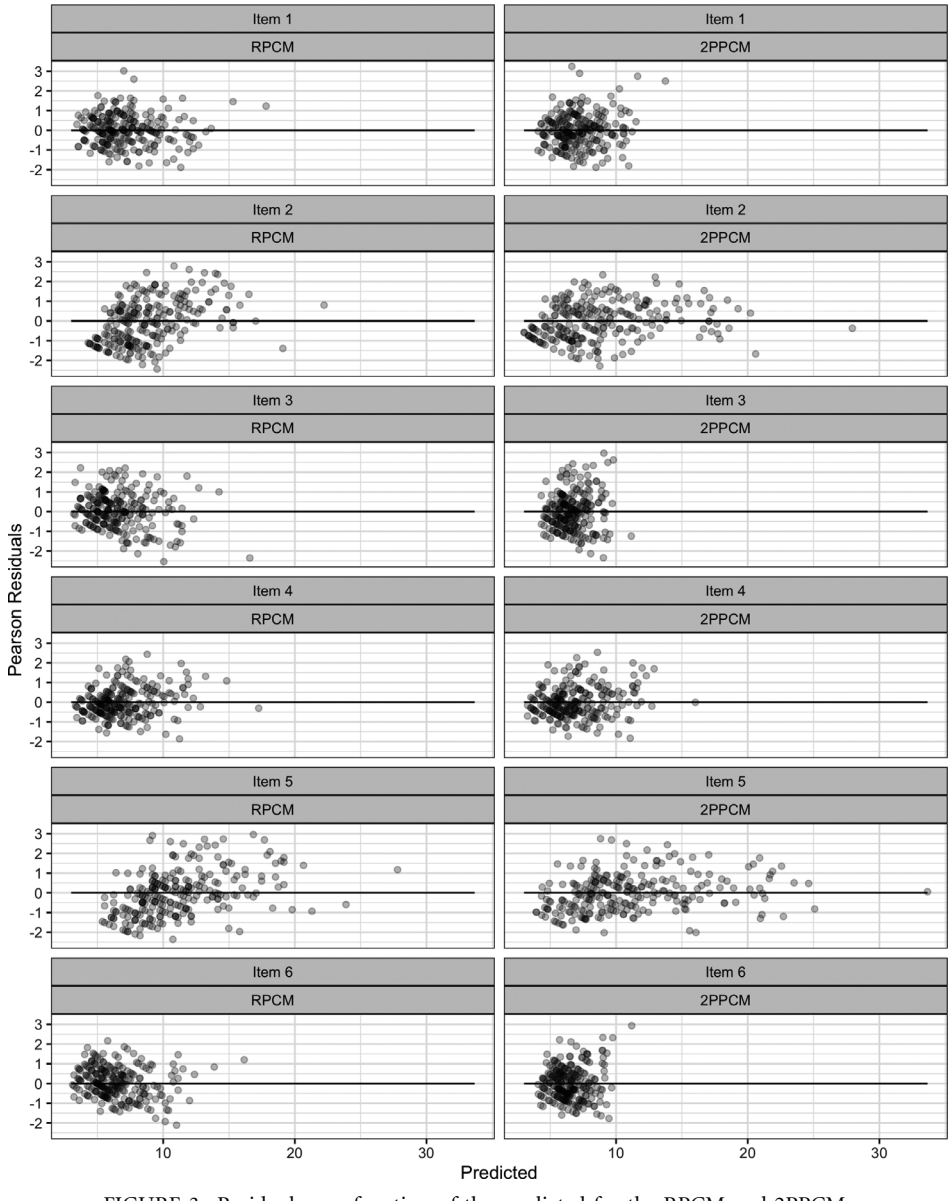

FIGURE 3. Residuals as a function of the predicted for the RPCM and 2PPCM.

ACCOUNTING FOR TASK CLUSTERING WITH A BIFACTOR MODEL

The B2PPCM significantly outperformed the 2PPCM $-\chi^2(6) = 19.66$, $p = .003$. The a' parameter estimates (which are the specific factor loadings) were very close to zero for the alternate uses and the consequences factor and were only significant for the instances factor. Therefore, to better stabilize estimation and more parsimoniously fit the data, we removed from the model the alternate uses and consequences specific factors, making a model referred to as the B2PPCM'. Like the B2PPCM, the B2PPCM' significantly outperformed the 2PPCM $-\chi^2(2) = 19.66$, $p < .001$. As confirmed by the B2PPCM not outperforming the simpler

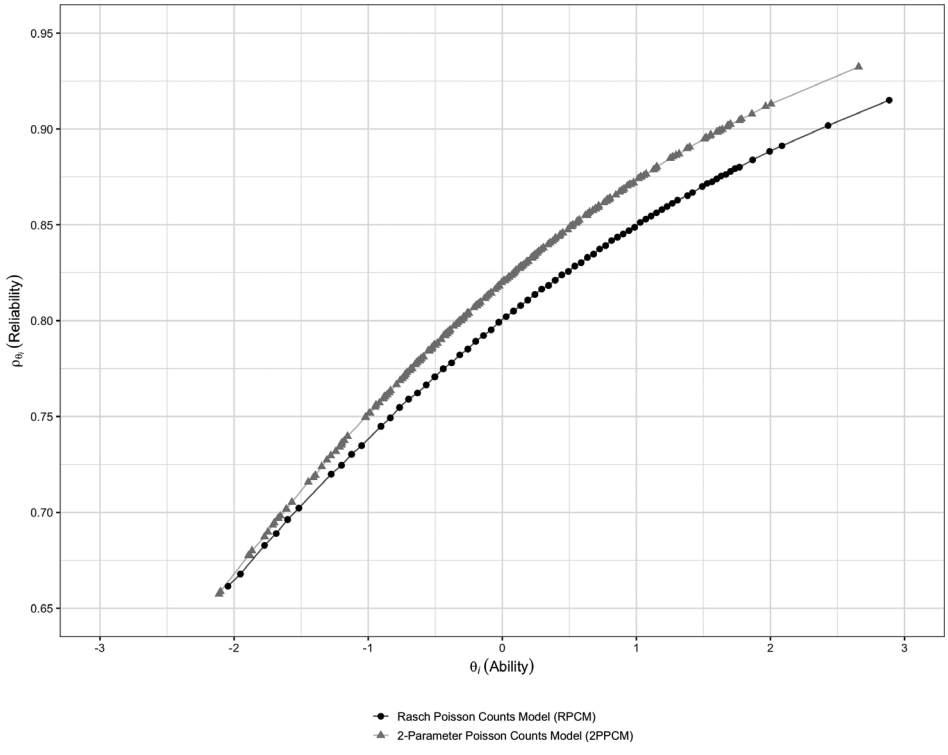

FIGURE 4. Conditional reliability for the RPCM and 2PPCM.

B2PPCM'—$\chi^2(4) = 0.00$, $p > .999$—and by the AIC and BIC of the two models, the B2PPCM' appeared as the model with the best fit on this data among all the tested models.

These results suggest that local dependencies due to similar prompts may arise in some situations (the instances in this case) but not all (the alternate uses and consequences here), and thus that the use of a bifactor structural model to account for these hypothetical local dependencies may or may not be useful depending on types of task used. The estimates of the B2PPCM and the B2PPCM' are reported respectively in Table 4 and Table 5. Like previously, we report the boxplots of the residuals per item for both models in Figure 5.

Like the RPCM and the 2PPCM, the B2PPCM and B2PPCM' were underdispersed ($\phi_{2BBPCM} = 0.65, 95\%\text{CI}[0.61, 0.70], \phi_{2BBPCM'} = 0.65, 95\%\text{CI}[0.61, 0.70]$), implying that their reliability estimates were also conservative. This was corroborated by the visual inspection of the residuals as a function of the predicted scores for both models, presented in *Figure 6*. The average reliability (for the general factor) was .694 for the B2PPCM (95% CI [.687, .700]) and .693 for the B2PPCM' (95% CI [.686, .699]).

The Akaike model-averaged reliability was .693 and the Schwartz model-averaged reliability was .694. The difference between these estimates and the reliability estimates of the RPCM and 2PPCM suggests that, when the assumption of local independence is violated (like here), using these models to estimate reliability could lead to misestimating reliability quite substantially. As a side note, Cronbach's α, which is the procedure that researchers often use as a default to estimate reliability, would provide an estimate of reliability of .81—which is also considerably different from the model-averaged reliability estimates obtained here.

DISCUSSION

In certain situations, fluency tasks can be so similar that one can assume that all fluency scores are equally related to the latent construct. Nevertheless, we argue that divergent thinking fluency scores do not fall in this category, in that, beyond their variation in difficulty, different tasks could reflect divergent

TABLE 4. Parameter estimates of the B2PPCM

Item	Parameter	Estimate	*SE*	*z*	*p*	95% CI
1	Slope (general)	0.274	0.029	9.597	<.001	[0.218;0.330]
	Slope (alternate uses)	0.002	0.056	0.033	.974	[−0.107;0.111]
	Intercept	1.937	0.031	63.329	<.001	[1.877;1.997]
2	Slope (general)	0.413	0.040	10.449	<.001	[0.336;0.491]
	Slope (instances)	0.235	0.052	4.499	<.001	[0.132;0.337]
	Intercept	2.084	0.039	53.347	<.001	[2.008;2.161]
3	Slope (general)	0.220	0.030	7.223	<.001	[0.160;0.280]
	Slope (consequences)	0.027	0.206	0.133	0.894	[−0.376;0.431]
	Intercept	1.877	0.030	63.574	<.001	[1.820;1.935]
4	Slope (general)	0.339	0.030	11.209	<.001	[0.280;0.399]
	Slope (alternate uses)	0.002	0.053	0.031	.975	[−0.102;0.106]
	Intercept	1.884	0.034	55.651	<.001	[1.818;1.951]
5	Slope (general)	0.387	0.038	10.289	<.001	[0.313;0.461]
	Slope (instances)	0.258	0.043	5.963	<.001	[0.174;0.343]
	Intercept	2.313	0.037	62.535	<.001	[2.241;2.386]
6	Slope (general)	0.233	0.029	7.939	<.001	[0.175;0.290]
	Slope (consequences)	0.014	0.113	0.125	.900	[−0.208;0.237]
	Intercept	1.849	0.030	61.668	<.001	[1.790;1.907]

TABLE 5. Parameter estimates of the B2PPCM'

Item	Parameter	Estimate	*SE*	*z*	*p*	95% CI
1	Slope (general)	0.275	0.029	9.605	<.001	[0.219;0.331]
	Intercept	1.937	0.031	63.356	<.001	[1.877;1.997]
2	Slope (general)	0.412	0.039	10.497	<.001	[0.335;0.489]
	Slope (instances)	0.237	0.051	4.656	<.001	[0.137;0.337]
	Intercept	2.086	0.039	53.369	<.001	[2.009;2.162]
3	Slope (general)	0.221	0.029	7.557	<.001	[0.164;0.278]
	Intercept	1.878	0.029	64.224	<.001	[1.821;1.935]
4	Slope (general)	0.339	0.030	11.213	<.001	[0.280;0.399]
	Intercept	1.885	0.034	55.699	<.001	[1.819;1.952]
5	Slope (general)	0.386	0.037	10.346	<.001	[0.313;0.461]
	Slope (instances)	0.260	0.043	6.111	<.001	[0.177;0.343]
	Intercept	2.314	0.037	62.571	<.001	[2.242;2.387]
6	Slope (general)	0.233	0.029	8.053	<.001	[0.177;0.290]
	Intercept	1.849	0.030	61.713	<.001	[1.790;1.908]

thinking to a different degree (e.g., they may tap into different cognitive processes, engage different personality traits, or involve specific domain knowledge). Because items may not be equally discriminant, the RPCM may be inappropriate. The proposed alternative—a 2-Parameter Poisson Counts Model—allows discrimination to vary per item. In our example dataset (Silvia, 2008a, 2008b; Silvia et al., 2008), it significantly outperformed the RPCM, indicating that discrimination was indeed variable by item. In addition, we introduced a bifactor extension of the 2PPCM, which allows to identify and account for item clustering (or local dependencies), and which outperformed both the RPCM and the 2PPCM in this example.

From the results obtained in this example, we suggest that, before using the RPCM, researchers question if a model with variable discrimination, such as the 2PPCM, is more appropriate. One can decide to use the RPCM rather than the 2PPCM if the divergent thinking tasks are assumed to be equivalent by design, but such an assumption should at least be clearly discussed. In the example at hand, justifying theoretically that all tasks reflect divergent thinking fluency to the same degree is hardly defensible, but in other cases (for

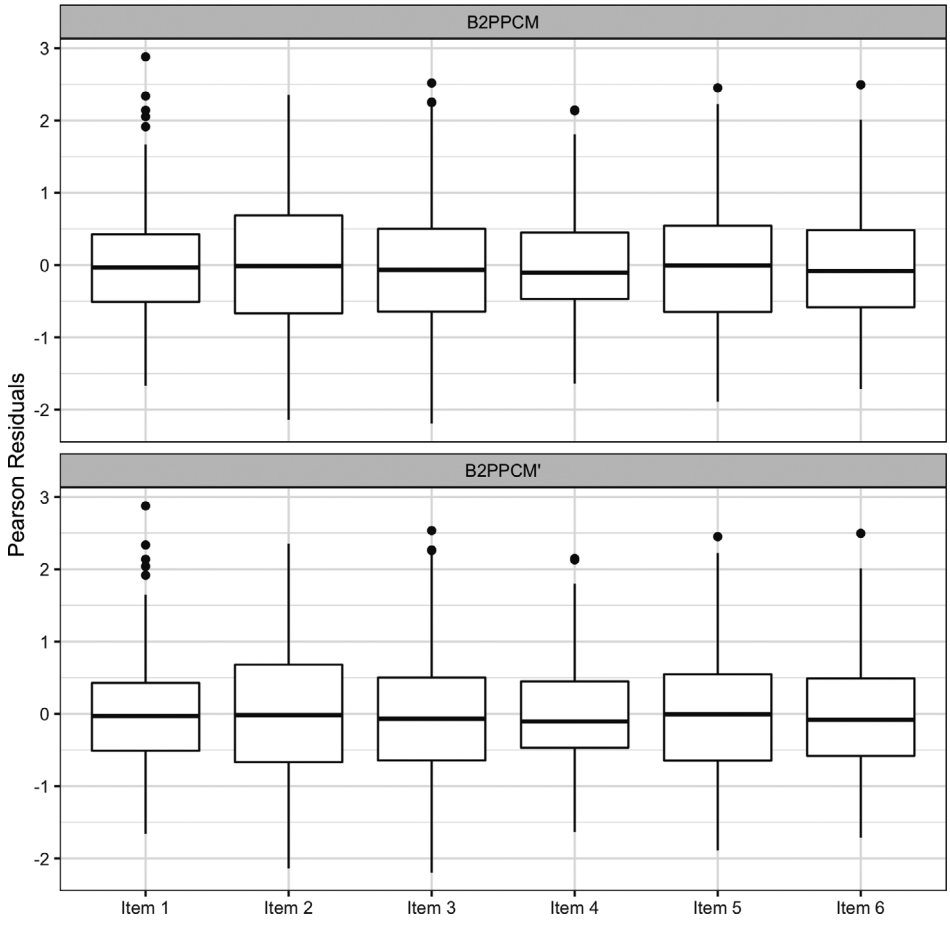

FIGURE 5. Distribution of the residuals of the B2PPCM and B2PPCM'.

example, for verbal fluency tasks), there could be substantive reasons to prefer the RPCM. Alternatively, a straightforward way to empirically test the assumption that items are equally discriminant is to compare the 2PPCM and the RPCM. If the 2PPCM outperforms the RPCM, then the assumption is violated and a 2PPCM is preferable.

Both the RPCM and the 2PPCM make the assumption that items are locally independent. In this example, we found that this assumption was violated, as bifactor extensions of the 2PPCM were able to identify a specific factor corresponding to the instances tasks. Further, we found that failing to account for such local dependencies resulted in overestimating reliability. Overall, the results suggest that local dependencies should be inspected upon fitting the RPCM or the 2PPCM, especially when communalities are suspected between tasks (e.g., some tasks sharing a similar prompt or subdomain). If such local dependencies are found, a bifactor 2PPCM is a more appropriate modeling approach.

FUTURE DIRECTIONS

The present research could find several extensions. First, a limitation for this study is that we only presented the 2PPCM and its bifactor extension, and applied them to an example dataset, but we did not investigate how these models perform in various conditions—notably for different sample sizes. Further simulation studies may focus on how the 2PPCM and its bifactor extension can be accurately estimated under various conditions of sample size, number of items, and dimensionality.

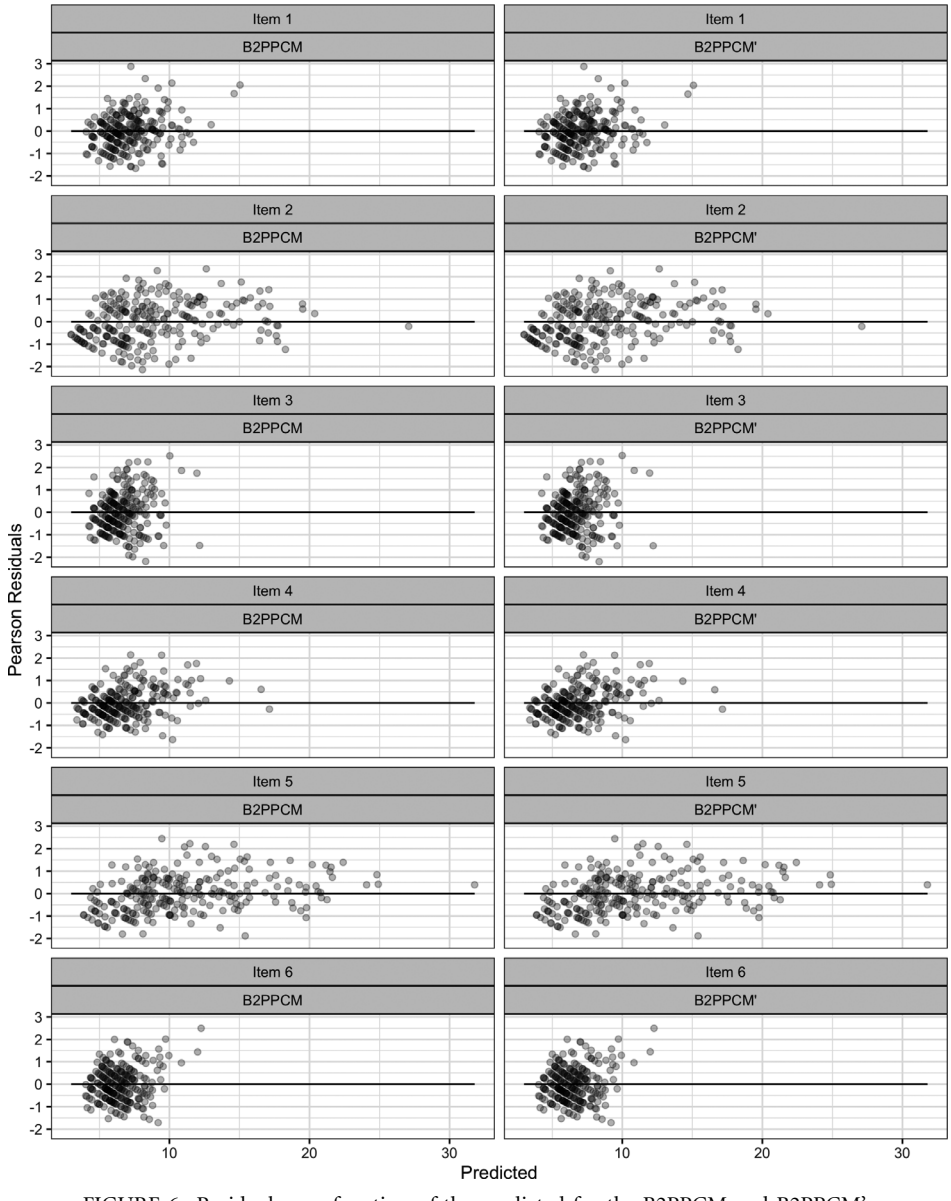

FIGURE 6. Residuals as a function of the predicted for the B2PPCM and B2PPCM'.

Like the RPCM, the 2PPCM and the B2PPCM are models that can directly be used to obtain person estimates (which can therefore replace sum/average scores) and to estimate reliability—therefore replacing more traditional techniques like Cronbach's α. However, a practical limitation of this study is the lack of statistical packages that allow to estimate the 2PPCM and the B2PPCM. While the RPCM can be estimated by most GLMM packages—such as "lme4" for R—it is not the case for the 2PPCM, which requires packages that are both capable of fitting Poisson counts models and factor structures. For the 2PPCM, researchers may turn to the "NLMIXED" package for SAS, or the packages "PLmixed" (Jeon & Rockwood, 2018) and

"brms" (Bürkner, 2017) for R, which allow for multilevel Poisson counts models with factor structures. A recent paper (Bürkner, 2020) presents how to use "brms" for item response models with discrimination parameters, which can probably be repurposed to fit the 2PPCM. Still, unfortunately and to the best of our knowledge, these packages currently would not allow to fit the B2PPCM, implying that researchers would have to use commercial software capable of estimating generalized structural equation models like Mplus or Stata. Further research may focus on how to make the estimation of the 2PPCM and the B2PPCM feasible and practical using various software packages, by comparing the capabilities and performance of different software under different conditions (such as different sample sizes and dimensionality).

The 2PPCM and B2PPCM are here discussed as better fitting alternatives to the RPCM. However, other models than the RPCM have been suggested for count responses. Further research may focus on comparing the 2PPCM and B2PPCM with other developments of Poisson response models, such as logistic Poisson models (Doebler, Doebler, & Holling, 2014), Conway–Maxwell–Poisson models (Forthmann et al., 2019) and Zero-Inflated Poisson models (Wang, 2010), in terms of empirical fit, estimation strategies, and interpretability.

In addition, we centered our work on how to account for item differences in discrimination when modeling and estimating the latent construct of fluency, but the 2PPCM and B2PPCM could further be used to analyze the items themselves, for example for test construction purposes. Researchers may notably be interested in using such a framework to identify tasks (or categories of tasks) that are particularly indicative of divergent thinking fluency. To do so, they may also be interested in ruling out specific factors due to item clustering using the bifactor approach.

The RPCM, 2PPCM, and B2PPCM showed underdispersion in our example dataset, as showed by the confidence intervals of the ϕ coefficient not including 1. Although reliability estimates tend to be more conservative in the presence of underdispersion, this result shows that the assumption of equidispersion was not met in the present study, which indicates that the absolute fit of all the models tested remained problematic. Replications in other datasets could reveal whether this issue is specific to this dataset or general to fluency scores in divergent thinking tasks.

Another limitation to overcome in the future is that the 2PPCM and B2PPCM, like the RPCM, assume a constant rate of responding for a given person and item. This assumption may be further discussed, especially as it could be pointed that, in divergent thinking tasks notably, examinees may have fluctuations of response rate, due to phenomena such as fatigue, attention fluctuation or variations of emotional states (Barbot, 2018). Challenging this assumption could lead to the development of Poisson count models with dynamic rate. Related to the issue of variable rate, another important assumption of the 2PPCM and RPCM is that events—in our example, reporting an idea—are independent (for a given individual and item). This assumption may also be challenged, especially in the case of idea generation: It may be that ideas tend to come in batches—an idea possibly increasing the rate of the next few ideas—rather than as independent events.

Finally, another possible extension of the model is the use of collateral information in estimation. The count of ideas is in some cases only one of the sources of information about the individual. In the case of divergent thinking tasks, for example, it is frequent practice to evaluate the originality or creativity of the ideas rating them. Other scoring techniques have been developed, such as asking the respondents to rank their own ideas (Silvia, 2008b), using the semantic distance between ideas (Hass, 2017b; Hass & Beaty, 2018; Heinen & Johnson, 2018) or using an overall scoring of all ideas of an examinee altogether (Silvia et al., 2009). Further research may focus on how fluency scoring with the 2PPCM may integrate with other scoring procedures.

CONCLUSION

We propose the 2-Parameter Poisson Counts Model and its Bifactor extension as more flexible alternatives to the traditionally used RPCM to model divergent thinking fluency scores. Certainly, the 2PPCM may not be estimable by most GLMM software like the RPCM is (Baghaei & Doebler, 2019), but the RPCM makes a strong assumption about the items—that their only differences lie in their difficulties—an assumption that we argue to be highly questionable in the context of divergent thinking tasks, and which we show to be violable empirically.

REFERENCES

Baghaei, P., & Doebler, P. (2019). Introduction to the rasch poisson counts model: An R Tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577.

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, *9*, 2529. https://doi.org/10.3389/fpsyg.2018.02529.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Birnbaum, A., Lord, F.M., & Novick, M.R. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores* (pp. 397–472). IAP.

Briggs, R.O., & Reinig, B.A. (2010). Bounded Ideation Theory. *Journal of Management Information Systems*, *27*(1), 123–144. https://doi.org/10.2753/MIS0742-1222270106.

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2020). Analysing standard progressive matrices (SPM-LS) with bayesian item response models. *Journal of Intelligence*, *8*(1), 5. https://doi.org/10.3390/jintelligence8010005.

Burnham, K.P., & Anderson, D.R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Clark, P.M., Griffing, P.S., & Johnson, L.G. (1989). Symbolic play and ideational fluency as aspects of the evolving divergent cognitive style in young children. *Early Child Development and Care*, *51*(1), 77–88. https://doi.org/10.1080/0300443890510107.

Doebler, A., Doebler, P., & Holling, H. (2014). A Latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, *38*(8), 587–598. https://doi.org/10.1177/0146621614543513.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts model. *Learning and Individual Differences*, *52*, 121–128. https://doi.org/10.1016/j.lindif.2015.01.013.

Forthmann, B., Celik, P., Holling, H., Storme, M., & Lubart, T. (2018). Item response modeling of divergent-thinking tasks: A comparison of rasch's poisson model with a two-dimensional model extension. *International Journal of Creativity and Problem Solving*, *28*(2), 83–95.

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32. https://doi.org/10.1016/j.intell.2016.03.005.

Forthmann, B., Gühne, D., & Doebler, P. (2019). Revisiting dispersion in count data item response theory models: The Conway–-Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology*, https://doi.org/10.1111/bmsp.12184.

Guilford, J.P. (1967). Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x.

Hass, R.W. (2017a). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y.

Hass, R.W. (2017b). Semantic search during divergent thinking. *Cognition*, *166*, 344–357. https://doi.org/10.1016/j.cognition.2017.05.039.

Hass, R.W., & Beaty, R.E. (2018). Use or consequences: Probing the cognitive difference between two measures of divergent thinking. *Frontiers in Psychology*, *9*, 2327. https://doi.org/10.3389/fpsyg.2018.02327.

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

Holzinger, K.J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965.

Jeon, M., & Rockwood, N. (2018). PLmixed: an R package for generalized linear mixed models with factor structures. *Applied Psychological Measurement*, *42*(5), 401–402. https://doi.org/10.1177/0146621617748326.

Kim, K.H. (2011). Proven reliability and validity of the Torrance Tests of Creative Thinking (TTCT). *Psychology of Aesthetics, Creativity, and the Arts*, *5*(4), 314–315. https://doi.org/10.1037/a0021916.

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*(4), 529–544. https://doi.org/10.1007/BF02302590.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. https://doi.org/10.1037/0033-2909.115.2.300.

Mouchiroud, C., & Bernoussi, A. (2008). An empirical study of the construct validity of social creativity. *Learning and Individual Differences*, *18*(4), 372–380. https://doi.org/10.1016/j.lindif.2007.11.008.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411. https://doi.org/10.1007/BF02296153.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351–363. https://doi.org/10.1177/014662169301700403.

Muthén, L.K., & Muthén, B.O. (1998). *Mplus User's Guide*. Muthén & Muthén.

Myszkowski, N., & Storme, M. (2017). Measuring "Good Taste" with the visual aesthetic sensitivity test-revised (VAST-R). *Personality and Individual Differences*, *117*, 91–100. https://doi.org/10.1016/j.paid.2017.05.041.

Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence*, *68*, 109–116. https://doi.org/10.1016/j.intell.2018.03.010

Myszkowski, N., Storme, M., Davila, A., & Lubart, T. (2015). Managerial creative problem solving and the Big Five personality traits. *Journal of Management Development*, *34*(6), 674–684. https://doi.org/10.1108/JMD-12-2013-0160.

Osborn, A.F. (1963). *Applied imagination: Principles and procedures of creative problem-solving*. New York, NY: Scribner.

Plucker, J.A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. *The Journal of Creative Behavior*, *45*(1), 1–22. https://doi.org/10.1002/j.2162-6057.2011.tb01081.x.

Rabe-Hesketh, S., & Skrondal, A. (2016). Generalized linear latent and mixed modeling. In W.J. van der Linden (Ed.), *Handbook of item response theory, Volume One: Models* (Vol. *1*, pp. 503–526). CRC Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/10.1007/BF02295939.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shao, Z., Janse, E., Visser, K., & Meyer, A.S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00772

Sheu, C.-F., Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods*, *37*(2), 202–218. https://doi.org/10.3758/BF03192688.

Silvia, P.J. (2008a). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual Differences*, *44*(4), 1012–1021.

Silvia, P.J. (2008b). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139–146. https://doi.org/10.1037/1931-3896.2.3.139.

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., ... 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.

Simonton, D.K. (2010). Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews*, *7*(2), 156–179. https://doi.org/10.1016/j.plrev.2010.02.002.

StataCorp. (2017). *Stata statistical software: Release 15*. College Station, TX: StataCorp LLC.

Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *Journal of Intelligence*, *7*(3), 1–17. https://doi.org/10.3390/jintelligence7030017.

Thurstone, L.L. (1938). *Primary mental abilities. Psychometric Monograph*. Chicago: University of Chicago Press.

Torrance, E.P. (1966). *The torrance tests of creative thinking–Norms—Technical Manual Research Edition—Verbal Tests, Forms A and B—Figural Tests, Forms A and B*. Personnel Press.

Trizano-Hermosilla, I., & Alvarado, J.M. (2016). Best alternatives to cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/fpsyg.2016.00769.

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.

Wallach, M.A., & Kogan, N. (1965). *Modes of thinking in young children*. New York, NY: Holt, Rinehart and Winston.

Wang, L. (2010). IRT-ZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, *35*(6), 671–692.

Wedel, M., Böckenholt, U., & Kamakura, W.A. (2003). Factor models for multivariate count data. *Journal of Multivariate Analysis*, *87*(2), 356–369. https://doi.org/10.1016/S0047-259X(03)00020-4.

Nils Myszkowski, Pace University

Martin Storme, IESEG School of Management, LEM-CNRS 9221

Correspondence concerning this article should be addressed to Nils Myszkowski, Department of Psychology, Pace University, Room 1315, 41 Park Row, New York, NY 10038. E-mail: nmyszkowski@pace.edu

AUTHOR NOTE

The data used in this study are from a previously published dataset (Silvia, 2008a, 2008b; Silvia et al., 2008), made available by the authors of the original papers. For convenience, a version of this dataset, reshaped for the analysis made in the present article, is also made available as Appendix S1.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Mplus code for model estimation.