9-9-2021

Taking Inventory of the Creative Behavior Inventory: An Item Response Theory Analysis of the CBI

Rebekah M. Rodriguez

Paul J. Silvia

James C. Kaufman

Roni Reiter-Palmon

Jeb S. Puryear

Taking Inventory of the Creative Behavior Inventory:
An Item Response Theory Analysis of the CBI

Rebekah M. Rodriguez, Paul J. Silvia

University of North Carolina at Greensboro

James C. Kaufman

University of Connecticut

Roni Reiter-Palmon

University of Nebraska at Omaha

Jeb S. Puryear

University of Montana

Unpublished preprint: September 9, 2021

Author Note

Rebekah M. Rodriguez and Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro; James C. Kaufman, Department of Educational Psychology, University of Connecticut; Roni Reiter-Palmon, Department of Psychology, University of Nebraska at Omaha; Jeb S. Puryear, Department of Teaching and Learning, University of Montana.

The self-report scales, R code, and raw data are available at Open Science Framework: https://osf.io/h6vj4/.

Please address correspondence to Rebekah M. Rodriguez, Department of Psychology, P.O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC, 27402-6170, e-mail: rmrodriguez@uncg.edu.

Abstract

The original 90-item Creative Behavior Inventory (CBI) was a landmark self-report scale in creativity research, and the 28-item brief form developed nearly 20 years ago continues to be a popular measure of everyday creativity. Relatively little is known, however, about the psychometric properties of this widely used scale. In the current research, we conduct a detailed psychometric investigation into the 28-item CBI by applying methods from item response theory using a sample of 2,082 adults. Our investigation revealed several strengths of the current scale: excellent reliability, suitable dimensionality, appropriate item difficulty, and reasonably good item discrimination. Several areas for improvement were highlighted as well: (1) the four-point response scale should have fewer options; (2) a handful of items showed gender-based differential item functioning, indicating some gender bias; and (3) local dependence statistics revealed clusters of items that are redundant and could be trimmed. These analyses support the continued use of the CBI for assessing engagement in everyday creative behaviors but suggest that the CBI could benefit from thoughtful revision.

Keywords: Creative Behavior Inventory; CBI; everyday creativity; item response theory; psychometrics; assessment

Taking Inventory of the Creative Behavior Inventory:
An Item Response Theory Analysis of the CBI

To study creativity's many interesting relationships, from academic achievement (Gajda et al., 2017) to well-being (Acar et al., 2020) to health (Cohen, 2006), researchers need tools to measure it. In the ever-expanding world of creativity assessment, a popular category of tool seeks to assess creative behavior (Kaufman, 2019; Reiter-Palmon & Schoenbeck, 2020): individual differences in how often people have engaged in activities that are deemed creative. Measures of past engagement in creative activities provide a useful complement to measures of creative thinking, personality traits, and markers of eminence and achievement in a creative domain. In the present research, we take a close look at the Creative Behavior Inventory (CBI), one of the oldest self-report measures in creativity assessment (Hocevar, 1976, 1979) that has been modified over the years and remains popular in modern research. Our aim is to identify major strengths of the scale and highlight promising directions for future revision and refinement.

The Creative Behavior Inventory

The CBI, developed by Hocevar (1976, 1979, 1981), was a milestone in the early era of self-report assessment of creativity. Such scales were uncommon at the time, and the CBI drew inspiration from educational research that used self-reported engagement in activities and accomplishments in high-aptitude adolescents and young adults (Holland, 1961; Holland & Nichols, 1964). Based on a sample of 239 college students, Hocevar proposed a 90-item scale: 75 items belonged to 6 subscales (fine arts, performing arts, math-science, crafts, literature, and music), and 15 items were placed in a "nonscalable items" category. The 90 items were activities, awards, and accomplishments, such as "Wrote a short story," "Gave a music recital," and "Received an award for acting." A noteworthy feature is that many items were qualified to exclude creative activities people did as part of their classes. Examples include "Knitted or crocheted something (excluding school or university course work)" and "Made a craft out of

metal (excluding school or university course work)."

For each item, participants indicate how often they engaged in a particular creative behavior or attained an accomplishment using a 0-3 scale: 0 = *Never did this,* 1 = *Did this once or twice,* 2 = *3–5 times,* 3 = *More than 5 times*. The frame of reference was limited from the adolescent years to the present, so creative activities in childhood were excluded. This makes the CBI an accumulative scale—scores should increase with time as people age. Assessing activities-to-date is common in measures of creative accomplishment (e.g., Carson et al., 2005), and it distinguishes the CBI from scales of creative activities that impose a rolling time window, such as activities and accomplishments during the past 12 months, like the Biographical Inventory of Creative Behaviors (BICB; Batey, 2007) or the past 10 years, like the Inventory of Creative Activities and Achievements (ICAA; Diedrich et al., 2018) and Creative Actions Scale (CAS; Elisondo, 2020).

Like many self-report scales, the 90-item CBI eventually fell into disuse, likely because of its unwieldly length and the increasingly dated quality of many of its items (e.g., "Wrote clever or humorous letters"). The scale got a new lease on life from research by Dollinger (2003), who crafted a focused, 28-item short form of the CBI. Few details are available, however, for how the 90-item scale was chopped to 28 items. In the article in which the 28-item CBI is first presented, Dollinger (2003) noted only that the full 90-item CBI had been used in a prior study (Dollinger et al., 2004) and that "supplementary analyses from that sample were the basis for derivation of a 28-item measure" (p. 104). Nevertheless, one obvious aim of Dollinger's brief CBI was to avoid facets and subscales and instead craft a unidimensional scale that yields a single score. In addition, implicit in the items that were selected was an emphasis on engagement in common, everyday creative activities over public achievements.

Dollinger's (2003) 28-item short form of the CBI—which for convenience we'll refer to simply as the CBI from here onward—has become a popular measure in creativity research, with more than 5000 downloads from the publicly available Creativity and Arts Tasks and Scales

archive on Open Science Framework since 2013 (as of June 2021; Silvia & Benedek, 2021). The scale has been featured in several reviews of creativity assessment (Kaufman, 2019; Puryear et al., 2019; Silvia et al., 2012), and translations of the CBI into German (Form et al., 2017) and Russian (Lebedeva et al., 2019) have recently been developed. The CBI's usage has been prominent in topics as diverse as personality, education, neuroscience, political ideology, and mental health (e.g., Dollinger, 2007; Lee & Kemple, 2014; McAleer et al., 2020; Nusbaum & Silvia, 2011; Silvia et al., 2020; Zedelius et al., 2020; Zhu et al., 2016). Although their topics and hypotheses vary, nearly all studies using the CBI apply cross-sectional research designs to study individual differences, usually with the CBI as one of several markers of between-person variation in creativity.

Locating the CBI Within Creativity Assessment

What the CBI measures shifted from Hocevar's (1976) original version, which mixed activities, awards, and achievements, to Dollinger's (2003) shorter form, which emphasizes activities. The CBI has been described as a measure of everyday creativity (Jauk et al, 2014; Silvia et al., 2012), creative activity (Diedrich et al. 2018; Nusbaum & Silvia, 2011), and creative production (Puryear, 2015). Nearly all the items assess the frequency of engagement in common creative domains, with an emphasis on crafts and the fine and performing arts, so the scale can be viewed as predominantly a behavioral measure of everyday creativity.

To locate the CBI within the broader world of self-report tools in creativity assessment, we see the CBI as part of the family of scales that inquire about common creative behaviors. Other tools in this category are the BICB (Batey, 2007), a *Yes/No* checklist of 34 creative activities that people might have engaged in over the past year, and the Creative Actions Scale (CAS; Elisondo, 2020), a recent scale that assesses engagement in common activities in 7 domains over the past 10 years. Like these scales, the CBI emphasizes engagement in common behaviors and activities. Unlike them, the CBI has a fixed starting point (since the start of adolescence) instead of a rolling window (the past year or past 10 years). The CBI also has a

narrower content focus. The BICB and CAS cast a broad net over everyday creative activity and intend to capture a wide range of domains, including interpersonal domains (e.g., leading, coaching, and managing). The CBI, in contrast, focuses on domains that are stereotypically creative in Western cultures, such as popular arts and crafts and fine arts, so it has a more narrowly defined sense of creative behavior.

The CBI can be contrasted with measures of creative achievement, such as the Creative Achievement Questionnaire (CAQ; Carson et al., 2005) and the ICAA (Diedrich et al., 2018), which seek to capture the public markers of achievement and eminence that one finds in "Pro-c" and "Big-C" creators. People with significant high-level attainments will presumably have significant engagement in common creative activities, but for most of the CBI items, high scores simply reflect frequent engagement in the activity, not public recognition or notable achievement in a domain. Likewise, the CBI can be contrasted with measures of people's self-attributed beliefs about their creativity. It seeks to assess how often people have done different activities, so the CBI does not provide the kind of information afforded by scales that assess people's own beliefs about how creative they are in different areas (Kaufman, 2012), their self-efficacy for creative goals (Karwowski et al., 2018), or their motives for pursuing creative activities (Benedek et al., 2020; Taylor & Kaufman, 2021).

The Present Research

The CBI's popularity in modern research seems out of proportion to the field's knowledge of the scale's psychometric features. The original CBI was developed with a small sample of 239 adults in the 1970s, and Dollinger's (2003) version was presented with essentially no information about how items were selected. For example, Puryear et al. (2019) pointed out that 24 items assess frequency of engagement in creative activities, but 4 items assess creative quality via accomplishments and awards. This imbalance is probably a vestige of the original scale's emphasis on both activities and achievements. Either way, the uneven item content highlights the lack of information about how the original 90 items were whittled down to 28 and

suggests that the revised CBI is an impure measure of "everyday creativity."

In the present research, then, we used psychometric tools from item response theory (IRT) to evaluate the CBI. Using a large sample of over 2,000 adults, we apply IRT to illuminate the behavior of the CBI's items as well as to identify possible targets for refining and improving the CBI's psychometric properties. Our analysis will focus on a few key issues. First, we examined the scale's reliability and dimensionality, especially if a single factor is credible. The CBI is always treated as unidimensional, but the factor structure of the CBI has not been thoroughly evaluated. Second, we examined the items' features, such as how well they fit the model, their level of difficulty (how "easy" they are to endorse), how well they discriminate between different underlying levels of creativity, and whether the 4-point response scale was suitable for the items. Third, we checked for possible measurement bias, particularly whether any items favored women or men in analyses of differential item functioning. These analyses can reveal if an observed gender difference reflects a true underlying group difference or if it reflects the influence of construct-irrelevant factors, so they can inform how researchers interpret possible gender differences in the CBI.

Method

Participants

Our analysis involved a sample of 2,082 adults who completed the 28-item CBI. This sample was formed by combining data from many studies in which the CBI had been included that were conducted at the authors' current and former institutions over the past 15 years. Nearly all the participants (around 96%) were college or university students enrolled at the institution; the rest were adults recruited from the surrounding community who were paid as part of the broader study. This sample represented the final dataset after filtering for inattentive, random, and careless responding using the R package *careless* (Yentes & Wilhelm, 2021). Of the total sample, 1590 were female (76.37%) and 492 were male (23.63%). Age data was available for around 82% of the participants (n = 1707), and this group tended to be young, ranging from

18 to 59 years old (M = 21.72, SD = 5.85, Mdn = 20). All participants provided informed consent.

Data Analysis

We invite readers to download the raw data and R code at Open Science Framework (https://osf.io/h6vj4). The data analysis was carried out in R 4.1 (R Core Team, 2021) using the *psych* (Revelle, 2021) and *TAM* (Robitzsch et al., 2021) packages. We used TAM to conduct the item response theory analysis. Because the CBI has a polytomous, ordinal response scale, we estimated a generalized partial credit model (GPCM; Ostini & Nering, 2006). A GPCM estimates each item's difficulty (b; how hard an item is to endorse), discrimination (a; the strength of an item's association with the latent trait), and category thresholds (the three boundaries between the four response options). The model was estimated using marginal maximum likelihood and identified via case constraint, which gives the underlying latent variable a mean of zero.

Results

Evaluation of Dimensionality, Local Dependence, and Reliability

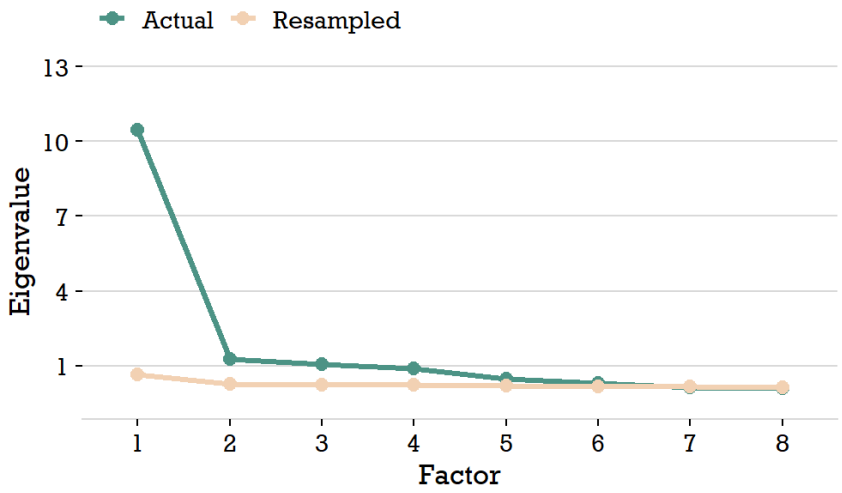
Prior to estimating our GPCM model for the CBI, we assessed the validity of two critical assumptions of IRT—unidimensionality of the latent trait and local independence of items—and explored reliability measures.

Dimensionality. Our exploration of dimensionality focused on an evaluation of *essential unidimensionality*, using a variety of criteria. This view of unidimensionality is less stringent than strict unidimensionality and recognizes that psychological constructs are rarely purely unidimensional, but rather have one dominant dimension with additional dimensions that are so minor that the construct validity is not compromised (Slocum-Gori & Zumbo, 2011). We used *psych* (Revelle, 2021) to conduct factor analyses using maximum-likelihood factor analysis. Because the CBI uses an ordinal response format, the factor analyses used polychoric correlations.

We applied three techniques for our assessment of dimensionality: parallel analysis,

Velicer's (1976) minimum average partial (MAP) criterion, and a greater-than 4:1 ratio of the first-to-second eigenvalues. Parallel analysis suggested 6 factors, but a scree plot showing the actual and resampled data clearly conveyed a dominant first factor with minor additional factors (see Figure 1). For the remaining criteria, MAP suggested 4 factors and the ratio of first and second eigenvalues clearly indicated 1 factor (Slocum-Gori & Zumbo, 2011). Altogether, viewing the CBI as "essentially unidimensional" seems credible, but the pattern of minor factors suggests that there are likely redundant item pairs or clusters that impair unidimensionality.

Figure 1. Scree plot from a parallel analysis of the CBI items.



Note. Only the first 8 eigenvalues are shown.

Local dependence. Local dependence, the residual covariation that remains after contributions from the latent trait are modeled, can mar the unidimensionality of a scale (Chen & Thissen, 1997). The minor factors revealed during our examination of dimensionality may indicate such local dependence. Locally dependent items often have overlapping meaning, and flagging these items provides a good starting point for future scale revisions for a shorter, more unidimensional scale.

We estimated local dependence in the CBI using the adjusted Q_3 (aQ_3) statistic based on

the Q_3 established by Yen (1984). This statistic is in the *r* correlation metric and represents the residual correlation between two items (i.e., the correlation after accounting for the shared influence of the latent trait). As described in Marais (2013), a negative sampling bias in Yen's (1984) original statistic can by corrected by centering the Q_3 values on zero using the mean item residual correlation. Values more extreme than |.20| were flagged for notable local dependence (Christensen et al., 2017). In total, this critical value flagged 17 locally dependent item pairs out of a possible 378 unique combinations within the CBI.

Although local dependence can arise from a variety of factors, in the CBI it commonly took the form of overlapping creative activities (e.g., writing poetry and writing songs, or fashion design and costume making) as shown in Table 1. These local dependencies are important for researchers using this measure to consider because even when the essential unidimensionality of the measure remains intact, local dependence can lead to inflated reliability estimates and a false sense of scale precision (Christensen et al., 2017). Patterns of local dependence can also aid future CBI development. The flagged items seen in the current analysis suggest many pairs of partly redundant items, which are good places to trim the CBI while also improving its unidimensionality.

Reliability. Reliability was explored with several coefficients. In line with previous literature, Cronbach's alpha was very high ($\alpha = .91$). Omega-hierarchical was also good but somewhat lower ($\omega_H = .73$). Finally, the GPCM IRT model provides an estimate of the expected a posteriori (EAP) reliability of the CBI trait scores. EAP reliability was good (.89). Taken together, score reliability for the CBI appears to be very good.

Item and Test Information

Item fit. Item fit was examined using mean-square infit and outfit statistics along with item RMSD (see Table 2). Infit and outfit values indicated good item fit, with most values hovering around 1.00, the ideal value (Bond et al., 2020). Only one item indicated possible outfit (item 3, "Made a craft out of metal": outfit = 1.17) according to significance tests. For

RMSD values, we followed size definitions suggested by Köhler et al. (2020) whereby RMSD values less than .02 were considered "negligible," values between .02 and .05 reflected "small" misfit, and values between .05 and .08 indicated "medium" misfit. All items in the CBI showed negligible to small misfit, with the highest value reaching .038 (item 9, "Wrote poems"). Taken together, the item fit statistics suggest good item fit overall for the current scale.

Item thresholds. Because the CBI uses a polytomous response scale of four ordered categories, a generalized partial credit model estimates three thresholds—the tau parameters in Table 2—that represent the underlying trait level at which someone has a 50:50 chance of selecting one response or the other. For example, the first response threshold for item 4 is -.90, so an underlying trait score of -.90 is the point at which someone has a 50:50 chance of endorsing the first option (*Never did this*) vs the second option (*Did this once or twice*). Because the response options are ordered, ascending from low to high values of creative engagement, the thresholds should be ordered, moving from lower trait values to higher trait values (Linacre, 2002). For item 12, for example, the thresholds ascended from .10 to .26 to 1.02, so as the trait level increased, the higher response options became more probable, as they should.

Most items, however, showed *disordered thresholds* (21 out of 28; see Table 2), which suggests that the 4-point rating scale is inappropriate. As a creative behavior increases, participants should choose increasingly higher rating categories (from *Never did this* to *More than 5 times*). However, with disordered thresholds, the step-like nature of the rating scale is broken. For example, item 20's thresholds were .23, .59, and -.83. The first two are ordered—people are equally likely to respond with 0 or 1 at trait level of .23, and equally likely to respond with 1 or 2 at .59, a higher trait level—but the third threshold isn't. People are equally likely to respond with a 2 or 3 at a trait level of -.83, which is lower than the others.

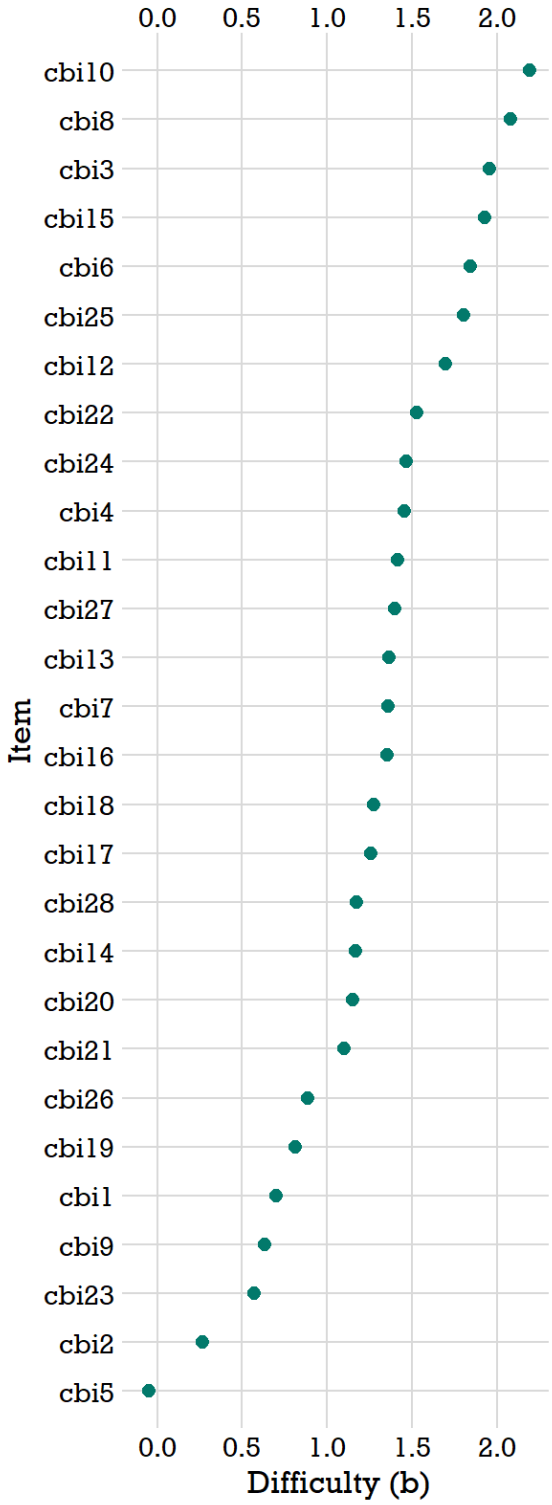
The notion of a disordered threshold is unfamiliar to many researchers who are grounded in the classical test theory model of assessment, but disordered thresholds are well understood in the Rasch and IRT literatures. In large samples like ours, disordered thresholds

reflect unusual distributions of responses across the options (Adams et al., 2012; Bond et al., 2020). They usually mean that the response scale has too many options, so participants underuse some sections of the rating scale (e.g., Silvia & Rodriguez, 2020). Because most items in the CBI have one disordered threshold out of three, it functionally resembles a 3-point rating scale instead of a 4-point scale.

Item difficulty. A generalized partial credit model provides estimates of each item's difficulty. For a self-reported measure of creative activity, it can sound odd to describe the "difficulty" of an item. For polytomous scales such as this one, the item difficulty parameter (*b*) indicates the amount of the latent trait (in this case, creative behavior) needed to endorse an item. The higher the *b* parameter an item has, the more creative behavior a respondent must identify in their everyday life to endorse it (see Table 2).

The current analyses suggest that the CBI is moderately difficult overall (see Figure. 2), with all but one item returning difficulty parameters above 0. Because trait ability scores are centered at zero in this model, only people with levels of creative behavior that are above average are likely to endorse most of these items. Even so, the range of difficulty is reasonable: even the hardest item on the CBI (item 10, "Wrote a play") isn't exceptionally difficult, with a *b* parameter of 2.19.

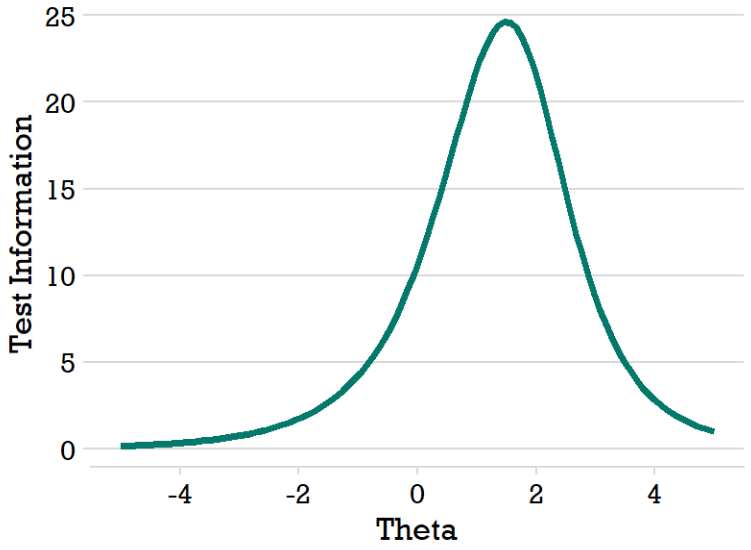
Figure 2. Difficulty (*b*) values for the CBI items, sorted hardest to easiest.



Item discrimination. Much like loadings in a confirmatory factor analysis, an item's discrimination (*a*) value is related to how closely that item can be linked to the underlying trait. For the CBI, discrimination (also known as *slope*) parameters varied considerably, ranging from .39 (*low*) to 1.34 (*high*), but the values largely fell within a moderate range (see Table 2). These results indicate that most items had a moderate ability to differentiate levels of creative behavior among respondents. Nevertheless, the handful of items with fairly low values (around .60 and below) suggests that some CBI items are providing relatively little information about people's relative standing.

Test information. In IRT, scales are more informative at certain ranges of the trait being measured. Test information functions describe the reliability of the measurement at different trait levels. Figure 3 illustrates the test information curve for the CBI. Consistent with the difficulty and discrimination parameters that inform it, the CBI reaches an informational peak at a trait level of about +1.85. This indicates that the scale is most reliable when measuring ability for respondents with moderately high levels of creative behaviors, which fits its intended use in creativity research well.

Figure 3. Test information function for the CBI.



Differential Item Functioning

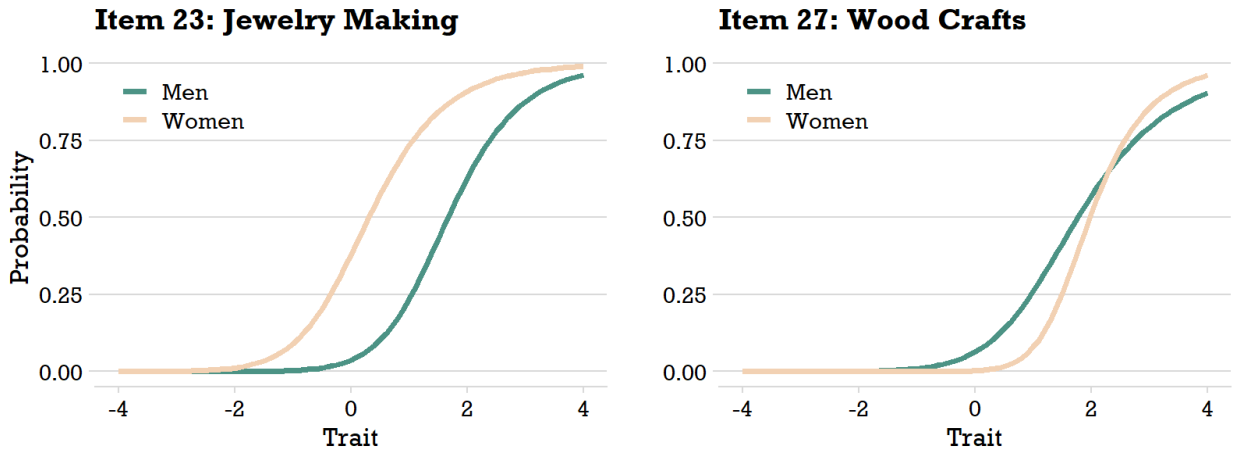
IRT analysis is built on examining the relationship between people's responses to test items and their underlying trait level (Osterlind & Everson, 2009). The expected response for an item should be a function only of the construct of interest—everyday creativity, in this case—instead of secondary and irrelevant constructs, such as age, gender, or cultural background. To take gender as an example, women and men can vary in their creativity trait scores, but women and men with identical traits scores should have identical expected item responses for the CBI items. Occasionally, however, respondents with the same trait level but different group membership are more likely to endorse an item. When this conditional responding happens, it is labeled as *differential item functioning* (DIF) or simply *item bias* (Osterlind & Everson, 2009; Penfield & Camilli, 2006).

To see if any of the CBI items were biased in favor of women or men, we evaluated gender-based DIF in the current sample using *lordif* (Choi et al., 2011), which uses the ordinal logistic regression approach to DIF combined with IRT trait scores and iterative purification (Osterlind & Everson, 2009). Men comprised around 24% of our total sample, but the absolute number of men (n = 492) was large enough for to reliably estimate possible gender-based DIF. We chose McFadden's R^2 for our effect size measure of DIF due to our large sample size (Jodoin & Gierl, 2001; Meade, 2010). Total DIF was flagged with a criterion of R^2 = .02, a common benchmark for a "small" effect size. As a result, all items with at least small gender-based DIF would be flagged.

At this cutoff, 6 items were identified as having gender-based DIF. Two items favored men: item 3 ("Made a craft out of metal") and item 27 ("Designed and constructed a craft out of wood"). The other four items favored women: item 5 ("Made your own holiday decorations"), item 17 ("Designed and made a piece of clothing"), item 18 ("Prepared an original floral arrangement"), and item 23 ("Made jewelry"). Figure 4 illustrates the difference using items 23

and 27 as examples. In the context of the CBI, DIF appears to reflect culturally-informed gender norms. For women and men with identical values on the underlying trait, women were nevertheless more likely to endorse culturally feminine decorative craft activities, and men were nevertheless more likely to endorse culturally masculine activities (woodworking and metalworking).

Figure 4. Item true score functions for CBI items 23 and 27.



Discussion

Our psychometric examination gave a much-needed check-up on the 28-item CBI, which was developed nearly 20 years ago (Dollinger, 2003) but has not received much psychometric scrutiny. An item response theory analysis conducted on responses from over 2,000 adults revealed some key strengths for the CBI as well as some obvious weak spots for future development of the scale.

First, the CBI has high score reliability and appears to be essentially unidimensional, with factor analyses suggesting one dominant factor and several minor factors. As noted earlier, surprisingly little information was given about how the 90-item CBI was pared back to 28-items (Dollinger, 2003), especially because the full scale was intended to measure 6 different domains

(Hocevar, 1976, 1979) yet the short form measures a single factor. The minor, secondary factors are thought to be results of related individual creative behaviors. An examination of local dependence showed overlap in item pairs that were redundant (i.e., drawing and keeping a sketchbook), shared a greater artistic domain (i.e., theater), or shared cultural significance (i.e., making holiday décor and jewelry). These redundant item pairs undermine the scale's unidimensionality and represent a good place to start for future revisions to the CBI.

Second, the CBI's 4-point response scale appears to be inapt. Most of the items showed disordered thresholds. For large samples, this usually indicates that participants are being given too many options for making their judgment (Linacre, 2002), which results in participants underusing some options. Based on the CBI's threshold profiles, we strongly suspect that people don't distinguish well between the intermediate levels. It is easy to know if you have never done something (scored 0) or if you have done something more than 5 times (scored 3), but remembering and distinguishing between the middle categories—doing something once or twice or between 3 to 5 times—is much harder. For a participant in their early 20s, recalling whether they have done a common creative activity twice (scored 1) or three times (scored 2) since the start of their teen years seems like a tall order, and the unreliability of such judgments are probably behind the disordered thresholds. The response scale of the CBI ought to be revised to have fewer categories. It's noteworthy that Qian et al. (2019; Qian & Plucker, 2018) condensed the CBI to a binary scale in their analyses of a 53-item version of Hocevar's (1979) original CBI, and a three-option scale seems promising (e.g., people engage in an activity never, occasionally, or frequently). We suggest that future revisions should consider the merit of reducing the number of response options that are conceptually rooted in the ways people explore and take up creative hobbies and behaviors.

Third, the items were generally well-behaved. Only one item ("Made a craft out of metal") was flagged for likely misfit. Regarding discrimination, most of the items showed acceptable discrimination values, but a handful of items had fairly low scores. These items

contribute less information to the scale and merit a close look in future revisions to the CBI. Regarding difficulty, the scale's items were moderately difficult, with no items being excessively easy or hard for the scale's intended population. For practical purposes, the CBI more reliably differentiates participants at the higher end of the latent creativity trait than the lower end.

Most self-report scales of creativity have "hard" items, in the IRT sense, because they are asking people about activities or accomplishments that are relatively uncommon (Silvia et al., 2021; Wang et al., 2014). For the CBI, however, the item difficulty for some items may be due to the additional constraint placed on the items (i.e., *excluding school or university course work*). This constraint makes it harder to endorse the CBI items and should be given a critical look in future work. Hocevar (1979) intended the qualification to restrict the responses to self-directed activities—things people did voluntarily as leisure activities—but the mix of qualified and normal items is an awkward and clunky feature of the scale, and it seems needlessly restrictive. After all, many people choose certain courses and programs of study because they want to learn about and engage in creative activities like creative writing, theatre, music, and visual art. We think future revisions of the CBI should consider omitting this qualification.

Finally, we examined possible gender-based measurement bias in the CBI with analyses of differential item functioning (DIF). Six items were flagged for DIF, with women favored for four items and men favored for two. In general, some differences in creative activity are expected based on cultural gender norms. Many creative domains are culturally gendered, from scrapbooking to woodworking, so cultural factors can steer women and men of equal creativity toward different domains. As a result, some degree of differential item functioning could be expected and not necessarily be seen as a serious psychometric problem. With flagged DIF items favoring women 2:1, however, there is a slight measurement bias favoring women in the CBI, which is enough to bias the overall scale scores (i.e., for women and men with the same true trait scores, women are expected to have slightly higher observed scores). As a result, if researchers find that women have a slightly higher overall CBI score in their sample, it is hard to know how

much the small difference is a real effect or an expression of item bias.

Overall, gender proved to be an interesting determinant of the types of creative behavior people tend to engage in, but the IRT methods employed here would be well-suited to examining any variable for item response biases. The ability to evaluate DIF is a key advantage of the family of Rasch and IRT methods, and researchers can explore whether item responses are biased by many demographic factors. Age, socioeconomic status, and racial and ethnic identity, for example would be good candidates for future exploration. We didn't evaluate them for the CBI because these variables were unmeasured or lacked variance, but they are clearly worth considering in future research with broader, more diverse samples. For continuous or ordered variables (e.g., age or household income), DIF analyses will call for different methods (e.g., Schauberger & Mair, 2020; Strobl et al., 2015), such as the recursive partitioning Rasch trees used to evaluate age-based DIF in the BICB (Silvia et al., 2021), but the logic of the analysis is largely the same.

In conclusion, our psychometric evaluation of the CBI highlights many good qualities: it has good score reliability, acceptable unidimensionality, appropriate difficulty, and generally informative items. Our evaluation also highlights opportunities for improvement and future possible development. Specifically, items that were flagged for multiple concerns in our analyses may be appropriate to drop from the scale. For example, items like "wrote poems" and "wrote the lyrics to a song" both provide relatively little information and are redundant with other items. Additional items showing redundancy include item 27, which favored men, and items 5, 17, and 23, which favored women, so eliminating these items would help correct gender bias due to DIF and make the scale more efficient. Finally, researchers have pointed out that a handful of items, like 11 ("received an award for an artistic accomplishment") and 12 ("received an award for making a craft"), reflect creative achievement (received public awards) instead of everyday creative activity (Puryear et al., 2017), so dropping these items would sharpen the CBI's focus on the construct of everyday creativity. However, it is important to be cautious when considering

scale alterations. It can be hard to predict unintended consequences from dropping certain items, and some aspects of the CBI (e.g., changing the 4-point response scale) call for a more thorough overhaul than dropping a handful of items. Given the scale's popularity in recent research, it's worth considering some thoughtful revisions so that this long-standing self-report scale could continue to serve creativity researchers for many more years.

References

Acar, S., Tadik, H., Myers, D., Van der Sman, C., & Uysal, R. (2020). Creativity and well-being: A meta-analysis. *Journal of Creative Behavior*.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547-573.

Batey, M. (2007). *A psychometric investigation of everyday creativity*. Unpublished doctoral dissertation. University College, London.

Benedek, M., Bruckdorfer, R., & Jauk, E. (2020). Motives for creativity: Exploring the what and why of everyday creativity. *Journal of Creative Behavior*, *54*(3), 610-625.

Bond, T. G., Yan, Z., & Heine, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, *17*(1), 37-50.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*, 178–194.

Cohen, G. (2006). Research on creativity and aging: The positive impact of the arts on health and illness. *Generations*, *30*(1), 7-15.

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*(3), 304–316.

https://doi.org/10.1037/aca0000137

Dollinger, S. J. (2003). Need for uniqueness, need for cognition, and creativity. *Journal of Creative Behavior, 37,* 99–116.

Dollinger, S. J. (2007). Creativity and conservatism. *Personality and Individual Differences*, *43*(5), 1025–1035. https://doi.org/10.1016/j.paid.2007.02.023

Dollinger, S. J., Urban, K. K., & James, T. A. (2004). Creativity and openness: Further validation of two creative product measures. *Creativity Research Journal*, *16*(1), 35-47.

Elisondo, R. C. (2020). Creative Actions Scale: A Spanish scale of creativity in different domains. *Journal of Creative Behavior, 55,* 215-227.

Form, S., Schlichting, K., & Kaernbach, C. (2017). Mentoring functions: Interpersonal tensions are associated with mentees' creative achievement. *Psychology of Aesthetics, Creativity, and the Arts, 11*(4), 440–450.

Gajda, A., Karwowski, M., & Beghetto, R. A. (2017). Creativity and academic achievement: A meta-analysis. *Journal of Educational Psychology*, *109*(2), 269.

Hocevar, D. (1976). Dimensionality of Creativity. *Psychological Reports*, *39*(3), 869–870.

Hocevar, D. (1979, April). *The development of the Creative Behavior Inventory (CBI).* Paper presented at the annual meeting of the Rocky Mountain Psychological Association (ERIC Document Reproduction Service No. ED 170 350).

Hocevar, D. (1981). Measurement of creativity: Review and critique. *Journal of Personality Assessment*, *45*(5), 450-464.

Holland, J. L. (1961). Creative and academic achievement among talented adolescents. *Journal of Educational Psychology*, 52, 136–147.

Holland, J. L., & Nichols, R. C. (1964). Prediction of academic and extra-curricular achievement in college. *Journal of Educational Psychology, 55*(1), 55–65.

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*(1), 95–105.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Karwowski, M., Lebuda, I., & Wisniewska, E. (2018). Measuring creative self-efficacy and creative personal identity. *International Journal of Creativity and Problem Solving, 28,* 45-57.

Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts, 6*(4), 298–308.

Kaufman, J. C. (2019). Self assessments of creativity: Not ideal, but better than you think. *Psychology of Aesthetics, Creativity, and the Arts, 13,* 187-192.

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*, 251-273.

Lebedeva, N., Schwartz, S. H., Van De Vijver, F. J., Plucker, J., & Bushina, E. (2019). Domains of everyday creativity and personal values. *Frontiers in Psychology*, *9*, 2681.

Lee, I. R., & Kemple, K. (2014). Preservice teachers' personality traits and engagement in creative activities as predictors of their support for children's creativity. *Creativity Research Journal*, *26*(1), 82–94. https://doi.org/10.1080/10400419.2014.873668

McAleer, J. T., Bowler, J. L., Bowler, M. C., Schoemann, A. M. (2020). Implicit and explicit creativity: Further evidence of the integrative model. *Personality and Individual Differences, 154,* 109643.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage.

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. *Handbook of Statistics*, *26*, 125–167.

Qian, M., & Plucker, J. A. (2018). Looking for renaissance people: Examining domain specificity-generality of creativity using item response theory models. *Creativity Research Journal*, *30*(3), 241-248.

Qian, M., Plucker, J. A., & Yang, X. (2019). Is creativity domain specific or domain general? evidence from multilevel explanatory item response theory models. *Thinking Skills and Creativity*, *33*, 100571.

Marais, I. (2013). Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 111-130). John Wiley & Sons.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728–743.

Nusbaum, E. C., & Silvia, P. J. (2011). Are Openness and Intellect distinct aspects of Openness to Experience? A test of the O/I model. *Personality and Individual Differences*, *51*(5), 571–574. https://doi.org/10.1016/j.paid.2011.05.013

Puryear, J. S. (2015). Metacognition as a moderator of creative ideation and creative production. *Creativity Research Journal*, *27*(4), 334-341.

Puryear, J. S., Kettler, T., & Rinn, A. N. (2019). Relating personality and creativity: Considering what and how we measure. *Journal of Creative Behavior*, *53*(2), 232–245.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Downloaded from https://www.R-project.org

Reiter-Palmon, R., & Schoenbeck, M. (2020). Creativity equals creativity–or does it? How creativity is measured influences our understanding of creativity. In V. Dörfler & Marc Stierand (Eds.), *Handbook of research methods on creativity* (pp. 290-300). Edward Elgar Publishing.

Revelle, W. (2021). *psych: Procedures for psychological, psychometric, and personality research*. R package version 2.1.3. https://CRAN.R-project.org/package=psych

Robitzsch, A., Kiefer, T., Margaret, W. (2021). *TAM-package: Test Analysis Modules*. R package

version 3.6.45. https://CRAN.R-project.org/package=TAM

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*, 279-294.

Silvia, P. J., & Benedek, M. (2021). *Creativity & Arts Tasks and Scales: Free for Public Use*. https://doi.org/10.17605/OSF.IO/4S9P6

Silvia, P. J., Eddington, K. M., Harper, K. L., Burgin, C. J., & Kwapil, T. R. (2020). Depressive anhedonia and creative self-concepts, behaviors, and achievements. *Journal of Creative Behavior*.

Silvia, P. J., & Rodriguez, R. M. (2020). Time to renovate the Humor Styles Questionnaire? An item response theory analysis of the HSQ. *Behavioral Sciences*, *10*(11), 173.

Silvia, P. J., Rodriguez, R. M., Beaty, R. E., Frith, E., Kaufman, J. C., Loprinzi, P., & Reiter-Palmon, R. (2021). Measuring everyday creativity: A Rasch model analysis of the Biographical Inventory of Creative Behaviors (BICB) scale. *Thinking Skills and Creativity, 39,* 100797.

Silvia, P. J., Wigert, B., Reiter-Palmon, R., & Kaufman, J. C. (2012). Assessing creativity with self-report scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 19–34. https://doi.org/10.1037/a0024071

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*, 443-461.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316.

Taylor, C. L., & Kaufman, J. C. (2021). The creative trait motivation scales. *Thinking Skills and Creativity*, *39*, 100763.

Velicer, W. (1976). Determining the number of components from the matrix of partial

correlations. *Psychometrika*, *41*, 321–327.

Wang, C. C., Ho, H. C., Cheng, C. L., & Cheng, Y. Y. (2014). Application of the Rasch Model to the measurement of creativity: The Creative Achievement Questionnaire. *Creativity Research Journal*, *26*(1), 62-71.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yentes, R., & Wilhelm, F. (2021). *careless: Procedures for computing indices of careless responding*. R package version 1.2.1. https://cran.r-project.org/package=careless.

Zedelius, C. M., Protzko, J., Broadway, J. M., & Schooler, J. W. (2020). What types of daydreaming predict creativity? Laboratory and experience sampling evidence. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000342

Zhu, W., Chen, Q., Tang, C., Cao, G., Hou, Y., & Qiu, J. (2016). Brain structure links everyday creativity to creative achievement. *Brain and Cognition, 103,* 70-76.

Table 1

Local Dependence Among CBI Item Pairs

First Item		Second Item		aQ_3
Number	Topic	Number	Topic	
9	Poetry	21	Wrote Short Story	0.38
9	Poetry	20	Wrote Song	0.36
11	Artistic Award	12	Craft Award	0.31
20	Wrote Song	21	Wrote Short Story	0.29
11	Artistic Award	24	Art Displayed Publicly	0.27
17	Fashion Design	28	Made Costume	0.24
8	Published Literature	10	Wrote Play	0.24
3	Metalworking	27	Woodworking	0.23
19	Drawing	26	Kept Sketch Book	0.23
2	Made Greeting Cards	15	Made Leather Craft	0.23
10	Wrote Play	25	Set Design	0.22
10	Wrote Play	21	Wrote Short Story	0.22
5	Made Holiday Décor	23	Made Jewelry	0.22
14	Made Cartoons	26	Kept Sketch Book	0.22
1	Painting	26	Kept Sketch Book	0.22
4	Puppet Show	15	Made Leather Craft	0.20
17	Fashion Design	23	Made Jewelry	0.20

Note. Not all scale items are featured in this table. Locally dependent items have aQ_3 correlations greater than |.20|.

Table 2

CBI Psychometric Features

Item	Slope (*a*)	Difficulty (*b*)	Tau 1	Tau 2	Tau 3	Infit	Outfit	RMSD	Disordered?
1. Painted an original picture*	1.059	.702	-.510	.532	-.023	1.014	1.004	.020	Yes
2. Designed and made your own greeting cards	.607	.268	-.815	.826	-.011	1.009	.996	.024	Yes
3. Made a craft out of metal*	.890	1.953	-.142	.118	.024	.987	1.168	.026	Yes
4. Put on a puppet show	.724	1.453	-.901	.868	.033	1.017	1.032	.028	Yes
5. Made your own holiday decorations	.784	-.050	-1.081	.614	.467	1.008	1.003	.027	Yes
6. Built a hanging mobile*	1.019	1.845	-.290	.119	.171	1.015	1.052	.020	
7. Made a sculpture*	1.336	1.361	-.487	.219	.268	1.008	1.006	.016	
8. Had a piece of literature (e.g., poem, short stories, etc.) published in a school or university publication	.674	2.079	-.026	.471	-.446	1.001	1.054	.023	Yes
9. Wrote poems*	.531	.634	.330	.623	-.953	1.011	1.014	.038	Yes
10. Wrote a play*	.956	2.192	.133	-.197	.064	1.050	.942	.023	Yes

11. Received an award for an artistic accomplishment	.863	1.412	-.497	.368	.130	1.001	1.045	.026	Yes
12. Received an award for making a craft	1.326	1.694	-.359	.102	.257	1.021	.982	.016	
13. Made a craft out of plastic, plexiglass, stained glass, or a similar material*	1.029	1.365	-.304	.325	-.021	1.010	1.015	.019	Yes
14. Made cartoons	.780	1.170	-.245	.562	-.317	1.012	1.000	.021	Yes
15. Made a leather craft*	1.254	1.930	-.159	.024	.135	.996	.951	.010	
16. Made a ceramic craft*	1.163	1.356	-.266	.101	.165	1.026	.980	.021	
17. Designed and made a piece of clothing*	.918	1.258	-.258	.176	.082	1.012	1.012	.017	Yes
18. Prepared an original floral arrangement	.720	1.276	-.229	.369	-.139	1.002	1.009	.021	Yes
19. Drew a picture for aesthetic reasons*	.863	.814	.277	.339	-.616	1.012	1.024	.018	Yes
20. Wrote the lyrics to a song*	.502	1.148	.226	.599	-.825	1.007	1.018	.022	Yes
21. Wrote a short story*	.660	1.101	-.276	.466	-.190	1.008	1.026	.023	Yes
22. Planned and presented an original speech*	.620	1.527	.138	.323	-.461	1.004	1.028	.023	Yes
23. Made jewelry*	.801	.571	-.283	.160	.123	1.023	.975	.028	Yes

24. Had artwork or craft work publicly exhibited	1.093	1.466	-.249	.116	.132	.999	1.042	.016	
25. Assisted in the design of a set for a musical or dramatic production*	.869	1.803	.299	.258	-.557	1.036	1.036	.027	Yes
26. Kept a sketch book*	.930	.885	-.024	.334	-.309	1.017	.997	.024	Yes
27. Designed and constructed a craft out of wood*	.931	1.399	-.127	.209	-.082	1.022	1.058	.029	Yes
28. Designed and made a costume	.992	1.175	-.567	.186	.381	1.003	.998	.013	

Note. Item labels with an asterisk were marked with the constraint, "(excluding school or university course work)."