2-2012

Assessing Creativity With Self-Report Scales : A Review and Empirical Evaluation

Paul J. Silvia
University of North Carolina at Greensboro

Benjamin Wigert
University of Nebraska at Omaha, bwigert@unomaha.edu

Roni Reiter-Palmon
University of Nebraska at Omaha, rreiter-palmon@unomaha.edu

James C. Kaufman
California State University - San Bernardino

Assessing Creativity With Self-Report Scales: A Review and Empirical Evaluation

Paul J. Silvia
Department of Psychology, University of North Carolina at Greensboro

Benjamin Wigert
Department of Psychology, University of Nebraska at Omaha

Roni Reiter-Palmon
Department of Psychology, University of Nebraska at Omaha

James C. Kaufman
Learning Research Institute and Department of Psychology, California State University, San Bernardino

---

Paul J. Silvia, Department of Psychology, University of North Carolina at Greensboro; Benjamin Wigert and Roni Reiter-Palmon, Department of Psychology, University of Nebraska at Omaha; James C. Kaufman, Learning Research Institute and Department of Psychology, California State University, San Bernardino.

We thank Mark Batey and Stephen Dollinger for providing copies of the BICB and CBI. The last three authors are listed in reverse alphabetical order and contributed equally.

Correspondence concerning this article should be addressed to Paul J. Silvia, Department of Psychology, P.O. Box 26170, University of North Carolina at Greensboro, Greensboro, NC 27402-6170. E-mail: p_silvia@uncg.edu

Abstract: This article reviews recent developments in the assessment of creativity using self-report scales. We focus on four new and promising scales: the Creative Achievement Questionnaire, the Biographical Inventory of Creative Behaviors, the revised Creative Behavior Inventory, and the Creative Domain Questionnaire. For each scale, we review evidence for reliability, validity, and structure, and we discuss important methodological features for users to consider. We then present new analyses of each scale based on a large, diverse sample. We evaluate each scale's item-level and scale-level psychometric features, using both classical test theory and item response theory, and we examine how the scales converge. All four scales performed well and covaried highly with each other. Based on the latest generation of tools, self-report creativity assessment is probably much better than creativity researchers think it is.

Keywords: creativity, assessment, psychometrics, measurement, reliability, validity

The psychology of creativity, like many scientific fields, is split over some hard problems. Is creativity associated with mental illness? Is there a domain-general creative trait or ability? Is creativity primarily individual or sociocultural? But creativity researchers do agree on some things, and one of those things is that measuring creativity is hard. Assessment has been a vexing problem for creativity researchers over the decades, in part because creativity research aspires to observe and measure things that are atypical, novel, innovative, and unusual, be they products, ideas, or people.

Creativity assessment broadly sorts into a few categories of measurements: creative products, creative cognition, creative traits, and creative behavior and accomplishments (Kaufman, Plucker, & Baer, 2008). In this article, we describe and evaluate several recent self-report methods for assessing creativity, particularly everyday creativity, creative achievements, and creative self-concepts. We intend this article to serve as a primer for researchers interested in measuring creativity with self-report scales. We describe each scale in detail (e.g., what it measures, whether it covers different creative domains), review any statistical quirks or issues (e.g., skewed scores, low base rates, ordinal scales), discuss past work that provides evidence for reliability and validity, and present new data on item-level and scale-level psychometric properties based on a large, diverse sample. At the end, we evaluate the relationships between the different self-report scales, thus illuminating how they converge and diverge.

Because of our emphasis on self-reported creativity, we won't cover several popular kinds of creativity assessment. For example, assessing creative products, such as with the consensual assessment technique (CAT; Amabile, 1982), is outside of the scope of this article. Several recent books and articles review how to assess creative products generally and the CAT specifically (Baer, Kaufman, & Gentile, 2004; Hennessey, Kim, Guomin, & Weiwei, 2008; Kaufman & Baer, in press; Kaufman, Plucker, & Baer, 2008). Likewise, we don't review the assessment of creative abilities using performance tasks; divergent thinking is the best known example of such assessment. Many recent articles and several handbook chapters review this controversial assessment domain (Nusbaum & Silvia, 2011; Plucker & Makel, 2010; Plucker & Renzulli, 1999; Silvia et al., 2008).

About the Sample

To evaluate the self-report measures of creativity, we collected data from a sample of 1304 people. The participants were enrolled at the University of Nebraska at Omaha (48%) or California State University, San Bernardino (52%). People completed the questionnaire online and received credit toward a research participation option in a psychology course. The final sample was diverse. The sample consisted primarily of young adults: the average age was 22.9 years old (*SD* = 6.63, *Mdn* = 21), and age ranged from 17 to 66. Approximately 76% of the sample was female. Based on self-reported race and ethnicity, approximately 55% of the sample was European American, 26% was Hispanic/Latino, 7% was African American, and 6% was Asian American. Consistent with the high proportion of Hispanic and Latino participants, 17% of the sample said that Spanish was the primary language spoken at home. The median past-year family income was $50,000 to $59,999, and the median past-year personal income was less than $15,999.

We excluded respondents with extensive missing data and respondents who indicated at the end of the questionnaire that they didn't pay much attention while filling it out, both perils of web-based data collection. Because of different patterns of missingness, the sample size varies from scale to scale.

Creative Achievement Questionnaire

The Creative Achievement Questionnaire (CAQ; Carson, Peterson, & Higgins, 2005) measures creative accomplishments in 10 domains: Visual Arts, Music, Dance, Architectural Design, Creative Writing, Humor, Inventions, Scientific Discovery, Theater and Film, and Culinary Arts. Unlike most self-report scales, the CAQ aims to capture Pro-c or Big-C creativity (Beghetto & Kaufman, 2007; Kaufman & Beghetto, 2009), so it focuses on significant, observable accomplishments. Only people with significant achievements in at least one domain receive high scores on the CAQ. By design, then, the CAQ yields highly skewed scores that pile up near the floor of the scale. Such scores are awkward to analyze, as we'll see later, but they reflect the true distribution of Big-C accomplishments.

The CAQ uses an innovative and complex scoring approach. The items and scoring instructions are in an appendix to Carson et al. (2005) article. Each domain has eight items, numbered 0 through 7, that represent increasing levels of creative achievement. For all domains, the first item indicates no training, experience, or accomplishment. For the Creative Writing domain, for example, the first item is “I do not have training or recognized talent in this area.” If people endorse the first item, they receive zero points for the domain and skip to the next one. The remaining items ask about increasingly rare levels of accomplishment that are logically connected, so endorsing a high item implies endorsing prior items. People receive more points for the items involving higher accomplishment. Most of the items are binary—people check whether an item applies to them—but some items involve writing a number, such as the number of patents awarded to their work and the number of awards received. For the Creative Writing domain, for example, item 7 asks people to provide a number in response to “My work has been reviewed in national publications.” For the free-response items, the participant's response is multiplied by the item number. Someone with three reviews in national publications, for example, would receive a score of 21 (3 reviews × 7, the item number) for the item. To get a domain score, researchers simply sum the domain's items. Scores range from zero (someone endorsed only the “no training or talent” item) to unrestricted high values.

Evidence for Reliability, Validity, and Structure

Regarding internal consistency, the CAQ does not lend itself to traditional analyses of internal consistency. Each domain includes eight items, but the items are not like Likert items, which are presumably tau-equivalent (i.e., each item has equal weight). For the CAQ, different items deliberately have different weights, and the items would not be viewed as interchangeable elements of a broader item pool. Likewise, the eight items aren't independent: if someone endorses the first item (indicating no accomplishments), then the other items receive zeros; similarly, if someone endorses a high-achievement item, then the prior items ought to have been endorsed. Within a domain, then, estimating Cronbach's alpha is not necessary. Researchers have done so— Carson et al. (2005), for example, report alpha values ranging from .70 to .87—but the CAQ domain scores don't meet the assumptions of a classical reliability analysis. It is best to think of each CAQ domain scale as a means for generating a single score, not as a conventional Likert scale.

At the scale level, estimating the internal consistency of the 10 domain scores is more informative. Such an analysis presumes that the 10 domains are markers of the same underlying factor. The 10 domains don't seem to form one factor (Carson et al. 2005; Silvia, Kaufman, & Pretz, 2009), but researchers commonly compute an overall score, which implies a single underlying factor.

Regarding evidence for validity, the CAQ has been extensively used. Research to date suggests that the scale effectively captures differences in creative achievements. In the original article, Carson et al. (2005) report that people with higher overall CAQ scores created better products in a collage-making task (r = .59), had higher divergent thinking scores (r = .47), and were higher in openness to experience (r = .33). Openness to experience has been a robust predictor of CAQ scores in several later studies (Hirsh & Peterson, 2008; Silvia, Nusbaum, Berg, Martin, & O'Connor, 2009). CAQ overall scores didn't appreciably correlate with measures of anxiety, depression, or social anxiety symptoms, but they did correlate with the Creative Behavior Inventory (r = .59), a measure of everyday creativity (Dollinger, 2003), and divergent thinking (r = .21; Silvia & Kimbrel, 2010). People with higher CAQ scores displayed more flexible cognitive control in a Stroop paradigm (Zabelina & Robinson, 2010) and had higher IQ scores (r = .34) and cumulative grade point averages (r = .18; Mar, DeYoung, Higgins, & Peterson, 2006). In samples with high education and intelligence, higher CAQ scores were associated with a genetic marker of psychosis risk (Kéri, 2009) and with decreased latent inhibition (Carson, Peterson, & Higgins, 2003).

Regarding factor structure, Carson et al. (2005) reported a series of exploratory factor analyses that suggested that the 10 domains sorted into two or three factors. In the two-factor model, they suggested an Arts factor (composed of drama, writing, humor, music, visual arts, and dance) and a Science factor (composed of invention, science, and culinary arts); the architecture domain was excluded. In the three-factor model, they suggested an Expressive (composed of visual arts, writing, and humor), Performance (composed of dance, drama, and music), and Scientific (composed of invention, science, and culinary arts) structure; the architecture domain was excluded. The inconsistent evidence regarding the factor structure probably stems from a few things, such as the huge differences in variances between the 10 domain scores and the extreme skew in each score. Both solutions are unsatisfying: each excludes the architecture domain, and it is odd to see the culinary arts domain lumped in with the science and

invention domains. Subsequent research has generally collapsed across the 10 domains to get an overall score, a decision that presupposes a single factor.

Notable Issues

Data screening

One reasonable concern with self-report measures of creativity is whether people will honestly report low scores. This concern is overstated, we think. Most of the self-report scales reported in this paper have scores that pile up at the low end, and most people receive zeros on most domains of the CAQ. It thus seems like people are perfectly comfortable reporting a lack of accomplishments. Creativity researchers probably overestimate how much the public at large values creativity.

The CAQ should be particularly robust against score inflation because it asks about concrete activities, not general self-perceptions of creative tendencies. It is also possible, however, that its scores underestimate creative accomplishments that are typical of college students. For example, maximal points are given for publications and recognition from critics (i.e., accomplishments more likely for graduate students and professionals), but some scales don't account for gradations of awards (e.g., local, regional, national, international)—a more common reflection of creative activities encountered by younger students and those performing on a team (e.g., dance team, musical group, forensics team).

A more realistic problem, however, concerns implausible values. In theory, the CAQ's items function as stair steps: people shouldn't endorse the high-achievement items without also endorsing the preceding items. For example, no one should endorse "I have been paid to act in theater or film" without having endorsed the earlier item "I have performed in theater or film." But some participants have aberrant response patterns, such as endorsement of only a single high value or endorsement of illogical patterns of accomplishments. It isn't clear what to do with such response patterns, and the literature on the CAQ to date has not discussed it. (Only a small percentage of a sample will have unusual patterns, but this subgroup can have a large influence simply because they are part of the minority that didn't get a score of 0 or 1.)

As with many problems, prevention is probably the best solution. For computer-based administration of the CAQ, the items can be presented conditionally: if people endorse the "no achievement" item, the program can move to the next domain, thus preventing certain illogical patterns. Another option is to include a scale that screens for random or inattentive responding. One example is the infrequency scale (Chapman & Chapman, 1983), which asks items that nearly everyone endorses or rejects (e.g., "There have been a number of occasions when people I know have said hello to me"). People who endorse more than a few items in the aberrant direction are probably responding randomly, not paying attention, or trying to fake an aberrant response profile.

Another realistic problem concerns the response criterion people use when judging each item, particularly the open-ended quantity items. Participants can interpret these items loosely, and one sometimes wonders whether participants are giving themselves too much credit. This is true of all self-report items—people can interpret the items in different ways and subjectively define the scale anchors (Biernat, 2003). For the CAQ, the problem is pressing because of the scale's extreme skew. If people interpret the multiplier items

loosely, they will end up with huge, outlying scores. For example, one multiplier for the Scientific Discovery item asks how many research grants people have been awarded. Scientists can view this loosely (e.g., small internal seed grants or travel grants) or stringently (e.g., grants from major federal funding agencies). Novices and experts may view the item differently. A novice might consider a small internal grant as being worthy, whereas an expert might list only major awards, resulting in the novice and the expert receiving the same score.

Score skew and overdispersion

By design, the CAQ yields highly skewed scores: most scores pile up at the scale's floor, and a few people have extreme scores. Because of the multiplier items, the range of possible scores is vast. Such skew was intended because the true population distribution of creative achievements must be skewed: by definition, Big-C creativity is uncommon.

Psychology has traditionally used two methods for dealing with highly skewed data. The first, denial, involves ignoring the skew and just running the analyses anyway; the second, faith, involves recognizing the skew but asserting that the general linear model is "robust" to violations of normality. We discourage researchers from these venerable methods. The degree of skew in the CAQ shouldn't be ignored, and conventional regression models aren't robust to such severe violations of their assumptions.

Instead, researchers should consider models designed for what Long (1997) calls "limited outcomes," such as ordinal, nominal, truncated, censored, and count data. Count models are especially valuable for studying creative achievements. Count outcomes follow a Poisson distribution, and one curious property of Poisson distributions is that the mean equals the variance. As a result, count data are inherently heteroskedastic. In psychological data, however, the variance is often larger than the mean, a condition known as overdispersion. The CAQ domains are overdispersed because of excessive zeros. Researchers can handle such distributions with a family of extended Poisson models, such as negative binomial models (Hilbe, 2007) and zero-inflated Poisson models (Long, 1997). Some work with the CAQ has used such models. Silvia and Kimbrel (2010) used negative binomial models to predict CAQ domain scores, and Silvia et al. (2009) treated the domains as count variables in their latent class analysis of the CAQ. Popular statistics software has good options for count models, and there are many good sources for interested researchers (Cohen, Cohen, West, & Aiken, 2003; Long, 1997).

Another way to handle the severe skew in the CAQ is to condense the scores into ordinal categories, such as low, medium, and high. For example, people with zero for a domain would be scored as 0, people with scores between 1 and 10 would be scored 1, and people with scores of 10 or greater would be scored 2. These are off-the-cuff suggestions—researchers would need to empirically evaluate such a scoring scheme—but we offer them as food for thought for future research. Ordinal data are much easier to work with than the CAQ's awkward raw scores. Although it might sound like much information is being lost, an ordinal scoring system probably captures most of what researchers want to know about the respondents, particularly for samples of young adults who are early in their creative development.

Summary scores for factors and classes.

The CAQ provides 10 scores, one for each domain. Research to date has not settled on ways to reduce the 10 domains into a smaller number of scores, or even whether collapsing across domains is a good idea. In the factor analysis approach, the domains would be assigned to factors, such as the two-factor and three-factor models suggested by Carson et al. (2005). Our sense is that the factor analytic evidence for the CAQ is still preliminary. Neither of Carson et al.'s factor solutions included all 10 domains, and the analyses did not seem to take steps to accommodate the serious skew in the scores. Additional research (particularly work involving confirmatory models) is needed.

An alternative approach to factor analysis, suggested by Silvia et al. (2009), is to explore whether the participants form discrete groups defined by patterns of creative accomplishments. This approach uses latent class analysis to look for nominal clusters of creative accomplishments. Unlike factor analysis, which looks for groupings of domains, latent class analysis looks for groupings of participants that have similar patterns of domain scores. In their analysis of the CAQ, Silvia et al. found evidence for three kinds of classes. The first class (around 66% of the sample) was a *no creativity* class: people in this group had low scores on all 10 CAQ domains. A second class (around 17% of the sample) was a *visual arts* class: people had elevated scores on the visual arts domain but not on any other domain. The third class (around 17% of the sample) was a *performing arts* class: people had elevated scores on the theater and film, music, dance, and creative writing domains. The latent class approach thus suggests that the creative achievements measured by the CAQ are domain-specific rather than domain-general: the undergraduate students in that sample had no accomplishments, accomplishments in the performing arts and writing, or accomplishments in the visuals arts.

Factor analysis and latent class analysis aren't exclusive approaches to the structure of the CAQ domains. For most research purposes, factors are more practical than latent classes. For studies of whether creativity is domain-general or domain-specific, Silvia et al. (2009) argued that latent class analysis can provide better evidence for domain-specificity than factor analysis can. Either way, examining the structure of the CAQ domains deserve more attention in future work.

The Present Data

Table 1 depicts the descriptive statistics for the 10 CAQ domains. In addition to the usual suspects, Table 1 displays the percentage of the sample that endorsed a 0 or a 1 for the domain. The severe skew is obvious: for each domain, at least half the sample has only a 0 or 1. Cronbach's alpha for the 10 domains was .60, which is consistent with presumed domain differences.

To evaluate the CAQ's structure, we started with confirmatory factor analyses (CFA) of the factor structures suggested by Carson et al. (2005): a two-factor model (Arts and Sciences) and a three-factor model (Expressive, Performance, and Scientific). For illustration purposes, we treated the CAQ domains as continuous scores, as in most past work. Both the two-factor model (CFI = .763, RMSEA = .101, SRMR = .058) and the three-factor model (CFI = .764, RMSEA = .105, SRMR = .058) fit quite poorly on most of the indices, according to conventional cut-offs (Kline, 2010). It is difficult to know whether the poor fit is attributable to poor model specification, to the non-normal indicators, or to both.

As an alternative, we specified the indicators as count variables. These models were estimated in Mplus 6.1, using maximum likelihood estimation with robust standard errors and Monte Carlo integration. We estimated CFA models with one, two, and three factors and compared them using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both the AIC and BIC favored the three-factor model over the other two; it is important to note, however, that this does not "confirm" the three-factor model per se, but it does indicate that it is relatively better than the other two.

We also examined the latent class structure of the CAQ. Silvia et al. (2009) found distinct classes of creative achievement in their analysis of the CAQ, and we were curious to know whether latent classes would replicate in a new sample. [1] We settled on a five-class solution, which is depicted in Figure 1. The classes were defined by different domain-specific patterns of creative achievement. The largest class (61.2% of the sample) was a *no creativity* class: people in this group had low scores on all 10 domains. Three of the classes reflected domain-specific achievement, such as a *visual arts* class (11.9%), a *dance* class (13.9%), and a *music* class (9.3%). Finally, a small *polymath* class appeared. People in this group—only 3.7% of the sample—had high scores in the visual art, dance, and creative writing domains. Like our past study, then, we found that scores on the CAQ define latent nominal groups with distinct patterns of achievement. In particular, most people have low scores on all the domains, and people with high scores tend to have them in only one domain.

Overall Evaluation

The CAQ is a clever scale. The innovative scoring approach accomplishes the developers' goal of creating a scale that yields highly skewed scores. The rarity of high scores is consistent with the notion that Big-C creativity is uncommon, and it reassures researchers who fear that most participants will claim accomplishments that they don't really have. The wide use of the scale shows that it works in the feral world of research with real participants. It's a good choice for creativity researchers who want to measure variation in Big-C creative achievements.

Analyzing the CAQ requires more thought than the usual self-report scale. First, conventional regression will nearly always be inappropriate for analyzing the domain scores. Models for overdispersed count data, such as negative binomial models and zero-inflated models, will work well for the count structure of the CAQ domain scores. Researchers should also explore the value of simplified ordinal scoring systems based on cut points. Second, research has not yet settled on methods for distilling the 10 domains into summary scores. The most common approach—summing or averaging across the 10 domains—is convenient, but it ignores the heterogeneity within the domains. Factor analyses find that the 10 domains don't seem to load on a single factor, and latent class analyses find clusters of people with high scores in different domains, not a group that is high on everything.

---

[1] Regarding the nuts and bolts of the latent class analysis, we explored at least 1000 initial starting values and allowed at least 100 to terminate to a final solution. Final models were then replicated across seed values. To choose a final class solution, we used classification quality, entropy (.930 in this solution), the BIC, and the corrected and bootstrapped versions of the likelihood ratio test (Nylund, Asparouhov, & Muthén, 2007). We also preferred, all else equal, models with fewer classes, which are more likely to replicate in future research. Three cases with extreme CAQ scores were dropped to enable model convergence—this is another example of the CAQ's vexing skew.

Biographical Inventory of Creative Behaviors

The Biographical Inventory of Creative Behaviors (BICB), developed by Batey (2007), is a 34-item scale that assesses everyday creativity across a broad range of domains. The BICB presents people with a range of behaviors and asks whether people have done them. It thus broadly resembles the CAQ in its behavior-anchored approach, but it has some interesting and useful differences. First, the domain coverage of the BICB is quite broad. The items cover the common domains of everyday creativity, such as arts, crafts, and creative writing, but they also cover social creativity, such as leadership, coaching, and mentorship. As a group, the items represent the most common kinds of everyday creative actions.

Second, the BICB uses a forced-choice *yes/no* response format. The scale's instructions ask people to indicate which activities they have been actively involved in during the past 12 months. The 34 activities are listed, and people simply check the ones that they have been actively involved in. The items are thus categorical—the forced-choice format yields binary 0/1 scores for each item. To obtain an overall scale score, researchers simply sum across the items, which yields a possible range of 0–34. The BICB does not have subscales.

Evidence for Reliability, Validity, and Structure

As a newcomer to the creativity assessment scene, the BICB is just starting to gain attention from researchers. Thus far, the scale has performed well in several studies of creativity. Regarding evidence for reliability, past work has reported Cronbach's alphas of .74 (Furnham, Batey, Anand, & Manfield, 2008), .78 Batey & Furnham, 2008; Batey, Furnham, & Safiullina, 2010), and .76 (Silvia & Nusbaum, 2010). Regarding evidence for validity, research has found correlations that one would expect. For example, the BICB correlates positively with divergent thinking fluency (r = .22, Furnham et al., 2008; r = .21, Batey, Furnham, & Safiullina, 2010) and with openness to experience (r = .38, Furnham et al., 2008; r = .33, Batey, Furnham, & Safiullina, 2010), two common markers of creativity.

Regarding the scale's factor structure, it appears that published research has not yet examined whether the BICB's items yield a single “creative behavior” factor. Given the diverse item content and the tendency for measures of creative behavior to yield domain-specific factors, it's important to examine the scale's factor structure.

Notable Issues

The BICB has several issues that researchers using the scale should take into account. First, categorical data pose analytic problems for general linear models. Just as binary outcomes in a regression model require generalized linear models (e.g., logistic or probit regression; Long, 1997), binary outcomes in an exploratory or confirmatory factor analysis require generalized models. Latent variable models for diverse kinds of outcomes are well-developed in the statistics literature (Skrondal & Rabe-Hesketh, 2004), but they have been slow to catch on in the world of applied research. Treating binary data as continuous violates several assumptions of conventional linear models: the residuals won't be normally distributed or homoskedastic, and the model can predict severely out-of-range values.

Second, the BICB has significant skew in its overall score. This was a deliberate feature of the scale, given its roots in Eysenck's (1993) notion that traits follow a normal distribution but creative achievement follows a Poisson distribution. We can illustrate the skew using our dataset; Figure 2 depicts the distribution of BICB scores in our sample. The skew in the BICB is not nearly as severe as the skew in the CAQ domains. Researchers with small sample sizes can apply transformations to the raw scores, such as a log transformation, that will make the distribution more normal.

Researchers with larger sample sizes have more options for handling the skewed scores. The binary items can be specified as binary indicators of a latent BICB variable in a confirmatory factor analysis (CFA). A CFA with binary data is simply an item response theory (IRT) model, and IRT models estimate normally distributed latent trait scores based on the binary responses. Researchers in the Bayesian tradition can use Bayesian Marcov Chain Monte Carlo (MCMC) methods (Lynch, 2007), which don't make the distributional assumptions made by traditional frequentist statistics, although we suspect that it will be a while before Bayesian MCMC models catch on in creativity research.

The Present Data

In our data, the BICB had good internal consistency. Cronbach's alpha was .89 when computed traditionally. An alternate way of estimating alpha that recognizes the binary quality of the data is to compute alpha from CFA loadings (for examples and details, see Hancock & Mueller, 2001; Silvia, 2011). We conducted a CFA that specified the items as tau-equivalent binary indicators and estimated Cronbach's alpha from the standardized loadings. Alpha for this model was .95.

Table 2 reports descriptive statistics for the items. The "% Yes" column lists the percent of the sample that responded with a *yes*, meaning that they did the creative action within the last year. The BICB items clearly capture a wide range of difficulty: some actions are common (e.g., items 6 and 31, which refer to decorating a personal space and to coming up with jokes) and others are quite rare (e.g., items 2 and 23, which refer to writing novels and publishing research).

To explore the scale's structure, we first conducted exploratory factor analyses using Mplus 6.1. We considered models with up to three factors with an oblique geomin rotation (Browne, 2001). The items were modeled as categorical (i.e., as binary outcomes). We used maximum likelihood with robust standard errors and Monte Carlo integration. A single factor model was the best representation of the data. Statistically, models with several factors generally had poor structure, and the secondary factors lacked clear interpretations. Practically, the 34-item BICB covers a lot of ground, and it doesn't have enough items for any particular domain to yield reliable domain-specific factors.

To gain insight into the BICB's item-level features, we examined the items using item response theory (IRT). We estimated a two parameter logistic (2PL) model using Mplus 6.1. Such a model estimates each item's discrimination and difficulty values. In IRT, an item's difficulty is expressed on the same scale as the underlying latent trait, which is in the standard normal *z* metric. Difficulty values thus have an intuitive interpretation: they are the trait level at which someone has a 50% chance of saying *yes* to the item. For example, item 20 ("made someone a present"), an easy item, had a difficulty value of −.694. This means that someone with a below-average level of everyday creativity (around −.69 SDs below the average) has a 50% chance of saying yes to the item. In contrast, item 2 ("wrote a novel"), a hard item,

had a difficulty value of 2.121. This means that only people who are more than 2.121 standard deviations above the mean in everyday creativity have more than a 50% chance of endorsing this item—it takes a lot of creativity to have written a novel.

The IRT discrimination values refer to how quickly item responses change as the underlying latent trait changes. They are conceptually (and statistically) related to factor loadings in a CFA: items with high loadings change more quickly as the latent trait changes, whereas items with low loadings change more slowly as the latent trait changes. Discrimination values range from 0 to around 3. All the discrimination values for the BICB were good, and many were quite high. It's worth noting that it isn't necessarily desirable for all the items to have very high discrimination values. Such items provide a lot of information but only for a narrow range of the trait (Embretson & Reise, 2000).

One way to summarize the items as a whole is to estimate the scale's total information function, which depicts the amount of information the BICB provides at different levels of the trait. Unlike classical test theory, which assumes that measurement error is constant across the trait, IRT allows researchers to estimate which range of the trait is estimated most precisely. For the BICB, the information function peaked at a trait level of around 1.50, which is fairly high (see Figure 3). One way to interpret this finding is that the BICB produces more reliable scores for people with high scores than for people with low scores. Stated differently, the BICB is good at discriminating the very high from the high from the somewhat high, but it is less good at discriminating the low from the very low.

The IRT analyses allow us to revisit the issue of the BICB's significant skew. An IRT model is a kind of generalized latent variable model (Skrondal & Rabe-Hesketh, 2004), in which a continuous, normally distributed latent trait is presumed to cause scores for indicators that aren't continuous (e.g., indicators that are ordinal, censored, truncated, or counts). Figure 4 shows the relation between a simple BICB sum score and the IRT-based trait score. As expected, they're related nonlinearly, and the IRT-based scores are less skewed.

Overall Evaluation

The BICB has not yet been widely used, but our analyses suggest that it deserves more attention from creativity researchers. The scale had a solid factor structure and internal consistency, and the IRT analyses indicated that the scale has good item-level properties (difficulty and discrimination values) and good scale-level properties (test information). The primary practical issue for researchers stems from the binary response format. Researchers with smaller samples should consider transformations of the overall score; researchers with larger samples should consider CFA models with binary indicators or IRT-derived trait scores. The scale is brief and easy to use, so we imagine that it will be a popular choice in future work.

Creative Behavior Inventory

The Creative Behavior Inventory (CBI), initially developed by Hocevar (1979), was one of the first self-report measures of creative behavior and accomplishment to be widely used in research. Hocevar generated an item pool by asking college students to list their most creative achievements and behaviors in several domains. Experts then rated each item in the pool for creativity. Items receiving low expert

ratings of creativity (i.e., items the experts felt didn't reflect creativity) were discarded, and applications of classical test methods eventually yielded a 90-item scale that measured creative behavior in several domains: fine arts, crafts, literature, math-science, performing arts, and music.

The original CBI has been used in several past studies. Hocevar (1980), for example, explored how divergent thinking, intelligence, and creative behavior covaried. He found that ideational fluency scores correlated with total CBI scores as well as several CBI subscales (crafts, performing arts, and math-science), and that crystallized intelligence (measured with synonym-antonym and analogies tasks) correlated with the total CBI score.

Recent research, however, has preferred a shortened form of Hocevar's scale developed by Dollinger (2003). Dollinger's short form didn't simply abbreviate the original inventory: the new scale shifted the underlying construct. Hocevar's scale includes many items that refer to infrequent, high-level accomplishments of the sort that appear on the CAQ, such as founding literary magazines and publishing literary and scientific work. Dollinger's short form weeded out the high-level items and retained items referring to common creative behaviors, such as making a costume, writing poems and song lyrics, and sketching. As a result, the short form should be considered a measure of everyday creativity, whereas the long form covers both everyday creativity and eminent creative achievement.

In addition to shifting the inventory toward common creative behaviors, Dollinger eliminated many of the domains in favor of a single factor consisting of common behaviors, most of which come from the art, crafts, and writing subscales of the original inventory. The revised scale thus lacks domain subscores. Researchers interested in domain differences may lament the collapsing across domains, but Hocevar's (1979) full scale doesn't seem to capture domain differences as well as the CAQ does. In his reanalysis of Hocevar's data, for example, Plucker (1999) found that the full CBI yielded only a single factor.

Evidence for Reliability, Validity, and Structure

Regarding reliability, past work with the revised CBI has reported Cronbach's alphas of .88 (Dollinger, Burke, & Gump, 2007) and .89 (Dollinger, 2003; Dollinger, Clancy Dollinger, & Centeno, 2005). Regarding evidence for validity, research has shown that the CBI correlates with many other markers of creativity, such as a creative drawing task (r = .31; Dollinger et al., 2005), an overall score from the CAQ (r = .59), self-rated creativity (r = .52), and divergent thinking (r = .19; Silvia & Kimbrel, 2010). Moving beyond measures of creativity, the CBI correlates with several aspects of personality, such as openness to experience— r = .37 in a correlation analysis (Dollinger, 2007) and β = .62 in a structural equation model (Silvia et al., 2009)—political conservatism (r = −.27; Dollinger, 2007), Gough's (1979) creative personality scale (r = .43; Dollinger et al., 2005), need for cognition (r = .35; Dollinger, 2003), and need for uniqueness (r = .33; Dollinger, 2003). Regarding evidence for factor structure, it doesn't seem that the factor structure of the CBI has received much attention.

Notable Issues

Researchers should note that the CBI uses a four-point ordinal response scale. People are asked to complete each item using the following scale: A = *Never did this*, B = *Did this once or twice*, C = *3–5 times*, and D = *More than 5 times*. Most studies using the scale have treated the items as continuous rather than ordinal, but the overall scale score is sufficiently skewed for researchers to be concerned. Figure 5 shows the overall scores from our dataset. The practical options are the same as for the BICB. Researchers who are using the observed overall score could simply transform it; a log transformation, for example, would make the overall distribution more normal. Researchers using latent variable methods with large samples, in contrast, could specify the items as ordinal indicators for a latent CBI variable, which would yield a polytomous IRT model (Embretson & Reise, 2000). The IRT option is impractical without a substantial sample size, given that the CBI has 28 items.

The Present Data

In our dataset, internal consistency was high: Cronbach's alpha was .92. Table 3 presents item-level descriptive statistics. The four response options (A, B, C, and D) were scored as 1, 2, 3, and 4. The lowest response option (A = *Never did this*) was the most frequent response for 26 of the 28 items, and 1 was the median for most of the items, indicating significant skew. Not surprisingly, then, the overall average of the 28 items was skewed as well, as shown in Figure 5.

To explore the CBI's factor structure, we conducted an exploratory factor analysis in Mplus 6.1 using maximum likelihood estimation with robust standard errors and an oblique geomin rotation. The results suggested that a one-factor model reasonably described the data. Scree plots strongly suggested only one factor. Furthermore, the factors after the first correlated highly with the first factor, were defined by few items, and lacked an obvious meaning. Overall, then, the single-factor model implied by past work seems suitable.

So far, we've presented analyses that treated the CBI items as continuous, but they're clearly ordinal. The response options should yield ordinal scores, and the pile-up of scores at the floor further indicates that the items are categorical instead of continuous. We don't want to condone pretending that categorical data are continuous, but we recognize that few researchers will have the sample size needed to model the items appropriately. A good option is to model the items as categorical (ordinal indicators with four levels) in a CFA, which yields a polytomous IRT model. Such a model requires a large sample because many parameters are estimated. Each item, for example, has a discrimination value and three difficulty values (the trait levels at which people cross a threshold from one scale value to the next), and the CBI has 28 items. Table 3 displays the factor loadings for the CBI based on a CFA with ordinal indicators. Overall, the items had good loadings.

To demonstrate the value of an IRT approach, Figure 6 depicts the relationship between IRT-based trait scores and the simple average of the CBI items. The relation between the two variables is nonlinear, which reflects that the two scoring methods diverge the most at the low and high ends of the trait. The distribution of IRT-based trait scores was substantially more normal, as would be expected.

Overall Evaluation

The 28-item CBI is a good choice for research. It captures a range of creativity, with an emphasis on everyday creative actions, and it has performed well in past work and in our own data. Both the item-level and scale-level properties were good, and past research has found expected relationships with many other measures of creativity and several dimensions of personality. The main issue concerns the significant skew in the overall scores. Researchers with large samples should consider IRT-based trait scores; researchers with smaller samples should consider transformations of the overall score.

Creativity Domain Questionnaire

The Creativity Domain Questionnaire (CDQ) measures people's beliefs about their level of creativity in different domains. Whereas the prior scales focus on observable behaviors and accomplishments, the CDQ focuses on people's self-concepts. People are asked "How creative would you rate yourself in…" and then given items from different domains, such as acting, computers, dancing, leadership, and writing. People's self-concepts are interesting in their own right, and they have been extensively studied in social psychology (Epstein, 1973; McConnell & Strain, 2007). Self-beliefs about creativity are also interesting for practical reasons. People use their beliefs about their traits, preferences, and abilities when choosing hobbies, careers, friends, and relationship partners, so self-beliefs play a role in many high-stakes decisions.

The CDQ has its roots in early work by Kaufman and Baer (2004), which explored the structure and correlates of self-rated creativity. They developed a brief 11-item scale, known as the Creativity Scale for Different Domains (CSDD). Respondents were asked to rate their creativity in different areas—for example, "How creative are you in bodily/physical movement (for example, dance, sports, etc.)?"—using a five-point scale (1 = *Not at all*, 5 = *Extremely*). Ten items referred to different domains of creativity; the final item asked people to give a global rating of their creativity. Kaufman and Baer found evidence for an empathy/communication factor (composed of communication, interpersonal relationships, solving personal problems, and writing), a hands-on creativity factor (composed of crafts, art, and bodily/physical), and a math/science factor (composed of math and science). The CSDD has been used in several studies of creativity as a complement to measures of creative abilities and achievements (e.g., Rawlings & Locarnini, 2007; Silvia & Kimbrel, 2010; Silvia et al., 2009).

For studying creative self-perceptions across domains, however, the CSDD was too short. The item coverage was too sparse to capture the many domains of human creativity, so a longer version was developed. The first CDQ (Kaufman, 2006) contained 56 items, each answered on a six-point scale. Exploratory and confirmatory factor analyses (Kaufman, Cole, & Baer, 2009) suggested seven domains: performance, math/science, problem solving, artistic-visual, artistic-verbal, entrepreneurial, and interpersonal. Except for the math/science domain, all the domains loaded strongly on a higher-order creativity factor. Kaufman (2006) used the long CDQ to appraise ethnic and gender differences in self-reported creativity, and Silvia et al. (2009, Study 2) found that the CDQ domains were better represented as continuous factors than as nominal latent classes.

The most recent incarnation of the CDQ, the Revised Creativity Domain Questionnaire (CDQ-R), balances the breadth of the long CDQ with the brevity of the original scale (Kaufman et al., 2009). The

CDQ-R has 21 items that form four factors: drama (e.g., acting, singing, writing), math/science (e.g., chemistry, logic, computers), arts (e.g., crafts, painting, design), and interaction (e.g., teaching, leadership, selling). Each item is completed on a six-point scale ranging from *Not at all creative* to *Extremely creative*. The four domain scores can be averaged to obtain a global creativity score. The items and instructions are published in Kaufman et al. (2009).

Evidence for Reliability, Validity, and Structure

Regarding evidence for reliability, Kaufman et al. (2009) reported Cronbach's alphas of .71 for the math/science, arts, and interaction subscales and .76 for the drama subscale. Alpha was .82 for the full 21-item scale. Regarding evidence for validity, Kaufman et al. (2009) found that creativity ratings across the domains correlated with personality in coherent ways. Openness to experience strongly predicted all four domains and the global score; the remaining factors of the Big Five had domain-specific patterns (e.g., extraversion predicted higher drama ratings but not higher math/science ratings), consistent with past work that has found both domain-general and domain-specific aspects of personality and creativity (e.g., Feist, 1998; Silvia et al., 2009).

Regarding the structure of the CDQ-R, Kaufman et al. (2009) settled on the four factor structure based on both exploratory and confirmatory factor analysis. Exploratory analyses suggested a good solution for the preferred four-factor model, but they also found a good three-factor solution that mapped onto the artistic, intellectual, and everyday creativity factors proposed by Ivcevic and Mayer (2009). A confirmatory model found that the four-factor solution showed acceptable fit (CFI = .92, RMSEA = .05) after several items' residuals were allowed to covary. This suggests that the evidence to date for the four factor solution is somewhat preliminary.

Notable Issues

After grappling with the awkward distributions and response scales of the prior scales, the ordinariness of the CDQ-R is refreshing. People complete each item using a six-point scale—there's no midpoint—and the item-level distributions usually aren't horribly skewed. The four subscales and the global creativity index have essentially normal distributions. The distributional differences reflect a difference between measures of creative achievements (which are always skewed) and creative aspects of personality (which are normally distributed).

The Present Data

The CDQ-R's internal consistency was higher in the present data than in Kaufman et al.'s (2009) sample. Cronbach's alpha was .78 for the drama subscale, .84 for the math/science subscale, .78 for the arts subscale, .79 for the interaction subscale, and .89 for the full 21-item scale. Table 4 displays the descriptive statistics for the items.

To examine the CDQ-R's structure, we estimated a CFA of the four subscales using Mplus 6.1. Each item served as an indicator for its respective subscale, no cross-loadings or correlated residuals were specified, and each subscale factor covaried freely with the others. The fit of this model was mixed: some fit

statistics were poor (CFI = .847), and others were acceptable but could be better (RMSEA = .071, 90% CI = .068 to .075; SRMR = .062). Table 4 displays the loadings from the CFA.

To gain insight into the sources of strain within the model, we examined modification indices based on a threshold of 50 (Brown, 2006). (We didn't make these changes to the model, but doing so would of course improve model fit.) The modification indices suggested two cross-loadings (loading the dance item on both the drama and arts factors, and loading the logic/puzzles item on both the math/science and interaction factors) and several correlated residuals (correlating algebra/geometry with chemistry, chemistry with logic/puzzles, and teaching/education with playing with children).

Overall Evaluation

The most recent version of the CDQ is a good choice for researchers interested in subjective beliefs about creativity. It provides a good balance between domain coverage and scale length, and the body of work on the CDQ-R and its predecessors (the CSDD and the 56-item CDQ) shows that the scale performs well. The main issue for future research concerns the structure of the CDQ-R. Shortening the longer CDQ required collapsing some domains: the long version had seven domains, but the short version has only four. The marginal fit of the CFA suggests that the items and subscales should receive more psychometric attention in the future, either by adding items (e.g., creating a distinct teaching/mentoring domain) or by sharpening the distinction between the arts and drama domain scales.

Evaluating the Coherence of the Creativity Scales

Thus far, we have examined each of the four self-report scales individually, using past research and our new data. A natural question, of course, is how the self-report scales relate to each other. Table 5 displays the descriptive statistics for the scales and the correlation matrix. For completeness, and to illustrate the influence of distribution shape, Table 5 displays descriptive statistics for raw scores, transformed scores, and IRT-based trait scores.

To evaluate whether the creativity scales reflect the same underlying factor, we estimated a confirmatory factor analysis of the four scales. For the indicators, we selected the log-transformed CAQ, the IRT-based trait scores for the BICB and CBI, and the CDQ-R global score (the average of the four domain scores). The variance of the latent variable was fixed to 1, and the model was estimated with Mplus 6.1 using maximum likelihood with robust standard errors. Model fit was good. The chi-square test was significant, $\chi^2(2\ df) = 15.056$, $p = .0005$, which isn't surprising given our large sample size, but the remaining fit indices suggested good fit based on conventional cut-offs: CFI = .982, SRMR = .020, RMSEA = .071 (90% CI = .040 to .106).

The standardized factor loadings were high and significant for all four creativity scales: CAQ ($\beta = .643$), BICB ($\beta = .698$), CBI ($\beta = .744$), and CDQ-R ($\beta = .605$). The reliability of the latent creativity variable can be expressed via *H*, known as maximal reliability (Drewes, 2000; Hancock & Mueller, 2001). *H* indicates the proportion of variance in the latent variable that is explained by the indicators, and it's interpreted much like Cronbach's alpha. For the latent creativity variable, *H* was .78. [2]

[2] We don't recommend using the highly skewed raw scores, but for the morbidly curious, a CFA of the four indicators' raw scores showed similar results. Model fit was good, $\chi^2(2\ df) = 7.727$, $p = .021$, CFI = .978, SRMR = .018, RMSEA = .047 (90% CI = .016 to .084). Maximal reliability for this model was $H = .76$. The indicators had good loadings on the latent variable, although several were notably lower: CAQ ($\beta = .450$), BICB ($\beta = .715$), CBI ($\beta = .766$), and CDQ ($\beta = .571$). Furthermore, the model with raw scores had more outlying and influential cases, appraised using Cook's *D* and the individual log-likelihoods, than the other model. Such findings aren't surprising, and they reinforce the need for creativity researchers to pay attention to distributional assumptions with scales that yield skewed scores.

In short, the four scales covaried as one would hope, which should reassure researchers interested in using them. Despite their differences in scaling and specific constructs, the scales overlap substantially. It isn't surprising that the three measures of creative achievement (CAQ, BICB, and CBI) covaried with each other, but it is interesting that the CDQ-R, a measure of self-perceived creativity, covaried just as highly. Our study isn't the first to find such an effect, but it does reinforce the notion that people's beliefs about their creativity are usually grounded in real differences in abilities and accomplishments, albeit self-reported ones.

A natural next step would be to incorporate other assessment methods, such as peer ratings of creativity, performance tasks, consensual assessments, and archival data. The self-report scales in our study covaried well, but they share both common constructs and a common method (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). We doubt that method variance is a major factor in the present data—if it were, evidence for single-factor solutions at the scale-level would be better than it is—but a large-scale study that examined a range of constructs and methods would illuminate the joint contributions of constructs and methods.

Conclusions and Recommendations

Creativity research has had many reviews of creativity assessment over the years, and they often end on a grim note (e.g., Hocevar, 1981; Michael & Wright, 1989). Many creativity researchers, in our experience, share this glum sense of the field's assessment tools. We think, however, that the state of self-report creativity assessment is much better than creativity researchers think it is. In this article, we focused on a group of relatively new self-report measures of creativity. Based on past research, our scale-specific analyses, and our analyses of how the scales converge, we think that this group of scales offers good choices for researchers interested in simple self-report measures of creative behavior, achievement, and self-perception.

Throughout, we have pointed to notable methodological issues, most of which concern item scaling and skew. Many of the scales yield categorical or count scores, and the resulting overall scores have significant positive skew. We have suggested a range of methods for dealing with the non-normal outcomes, including transforming the raw scores or estimating latent trait scores using IRT models. Regardless of the method, we do encourage researchers to take skew and scaling more seriously than they have been taken in past work. In our data, the raw scores did not perform as well based on CFA loadings, patterns of correlations, and number of outlying and influential cases, as one would expect.

A second issue concerns the structural features of the scales. The two scales that focus on domains—the CAQ and the CDQ-R—are the most complex structurally. For the CAQ, there's no ideal way to distill the 10 domains. Most research has simply averaged or summed the domain scores. In our data, factor analyses suggested more than one factor, and latent class analyses found that people with high scores sort into different nominal groups, not a single high-scoring group. People with equally high overall scores thus have high achievements in different domains. Given the many successful applications of the CAQ, however, we suspect that using a transformed sum score is probably okay for the low-stakes basic research that the scale is used for. For the CDQ-R, the four factor domain structure could probably use some additional refinement, given the mixed fit of the confirmatory factor analyses.

The self-report scales discussed here have fared well in past research and in the present research, but we should emphasize that they were all developed for low-stakes assessment purposes. To date, there's no evidence to support their valid use for high-stakes purposes, such as screening people for employment or making educational decisions. “Faking good” is easy to do on these scales for respondents who want to appear to be more creative than they really are, so scores from high-stakes contexts will be suspect. We doubt that many participants in a research project do this—the achievement and behavior scales (CAQ, BICB, and CBI) have skewed scores that pile up on the floor, so at least the majority of the sample feels comfortable giving low scores. The scales reviewed here also have consistent, replicated effects with other variables, which provide evidence for the scores' validity. Nevertheless, researchers should keep in mind that even extensive evidence for validity in low-stakes domains doesn't mean that an assessment tool is suitable for other purposes (Messick, 1995).

We imagine that many researchers would ask “Which scale should I use?” These scales vary on several dimensions that can aid in assessment decisions. First, the scales vary in how they treat domains. Some scales provide scores for distinct domains (the CAQ and CDQ-R), and others provide only domain-general scores (the BICB and CBI). Second, the scales vary in whether they evaluate eminent or everyday creativity. Some scales enable researchers to assess uncommon Big-C levels of creativity (the CAQ), and others focus on common little-c creativity (the BICB, CBI, and CDQ-R). Third, the scales vary in whether they assess public behaviors or inner self-beliefs. Some scales focus on creative behaviors and observable accomplishments (the CAQ, BICB, and CBI), and others focus on subjective self-beliefs regarding creativity (the CDQ-R). Table 6 classifies each scale on these dimensions.

That said, the best answer to “Which scale should I use?” is “all of them.” We encourage researchers to use as many measurement tools as is feasible, given the limits on time and resources, including several self-report measures of creative actions and thinking styles that weren't included or reviewed here (Durmysheva & Kozbelt, 2010; Ivcevic & Mayer, 2009; Runco, Plucker, & Lim, 2001). Including multiple measures is better for many obvious reasons, but one less obvious reason is that it will accelerate the growth of evidence in creativity assessment. We agree with the many researchers who have pointed out that creativity assessment has not advanced as quickly as other fields have (see Plucker & Makel, 2010). The field's knowledge about how its tools perform will develop more quickly when more of the scales are used more often.

References

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

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, *16*, 113– 117.

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

Batey, M., & Furnham, A. (2008). The relationship between measures of creativity and schizotypy. *Personality and Individual Differences*, *45*, 816– 821.

Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge, and personality as predictors of creativity. *Learning and Individual Differences*, *20*, 532– 535.

Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *1*, 73– 79.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist*, *58*, 1019– 1027.

Brown, T. D. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111– 150.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2003). Decreased latent inhibition is associated with increased creative achievement in high-functioning individuals. *Journal of Personality and Social Psychology*, *85*, 499– 506.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, *17*, 37– 50.

Chapman, L. J., & Chapman, J. P. (1983). *Infrequency Scale*. Unpublished test (copies available from T. R. Kwapil, Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, 27402–6170).

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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*, 1025– 1035.

Dollinger, S. J., Burke, P. A., & Gump, N. W. (2007). Creativity and values. *Creativity Research Journal*, *19*, 91– 103.

Dollinger, S. J., Clancy Dollinger, S. M., & Centeno, L. (2005). Identity and creativity. *Identity*, *5*, 315– 339.

Drewes, D. W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods*, *5*, 214– 227.

Durmysheva, Y., & Kozbelt, A. (2010). The creative approach questionnaire: Operationalizing Galenson's finder-seeker typology in an non-expert sample. *International Journal of Creativity and Problem Solving*, *20*, 35– 55.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, *28*, 404– 416.

Eysenck, H. J. (1993). Creativity and personality: Suggestions for a theory. *Psychological Inquiry*, *4*, 147– 178.

Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review*, *2*, 290– 309.

Furnham, A., Batey, M., Anand, K., & Manfield, J. (2008). Personality, hypomania, intelligence and creativity. *Personality and Individual Differences*, *44*, 1060– 1069.

Gough, H. G. (1979). A creative personality scale for the Adjective Check List. *Journal of Personality and Social Psychology*, *37*, 1398– 1405.
Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R.Cudeck, S.du Toit, & D.Sörbom (Eds.) , *Structural equation modeling: Present and future* (pp. 195– 216). Lincolnwood, IL: Scientific Software International.
Hennessey, B. A., Kim, G., Guomin, Z., & Weiwei, S. (2008). A multi-cultural application of the Consensual Assessment Technique. *International Journal of Creativity and Problem Solving*, *18*, 87– 100.
Hilbe, J. M. (2007). *Negative binomial regression*. New York: Cambridge University Press.
Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Research in Personality*, *42*, 1323– 1333.
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. (1980). Intelligence, divergent thinking, and creativity. *Intelligence*, *4*, 25– 40.
Hocevar, D. (1981). Measurement of creativity: Review and critique. *Journal of Personality Assessment*, *45*, 450– 464.
Ivcevic, Z., & Mayer, J. D. (2009). Mapping dimensions of creativity in the life-space. *Creativity Research Journal*, *21*, 152– 165.
Kaufman, J. C. (2006). Self-reported differences in creativity by ethnicity and gender. *Applied Cognitive Psychology*, *20*, 1065– 1082.
Kaufman, J. C., & Baer, J. (2004). Sure, I'm creative—But not in mathematics! Self-reported creativity in diverse domains. *Empirical Studies of the Arts*, *22*, 143– 155.
Kaufman, J. C., & Baer, J. (in press). Beyond new and appropriate: Who decides what is creative?*Creativity Research Journal*.
Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The Four C Model of Creativity. *Review of General Psychology*, *13*, 1– 12.
Kaufman, J. C., Cole, J. C., & Baer, J. (2009). The construct of creativity: A structural model for self-reported creativity ratings. *Journal of Creative Behavior*, *43*, 119– 134.
Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. Hoboken, NJ: Wiley.
Kaufman, J. C., Waterstreet, M. A., Ailabouni, H. S., Whitcomb, H. J., Roe, A. K., & Riggs, M. (2009). Personality and self-perceptions of creativity across domains. *Imagination, Cognition and Personality*, *29*, 193– 209.
Kéri, S. (2009). Genes for psychosis and creativity: A promoter polymorphism of the *Neuregulin* 1 gene is related to creativity in people with high intellectual achievement . *Psychological Science*, *20*, 1070– 1073.
Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.
Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.
Mar, R. A., DeYoung, C. G., Higgins, D. M., & Peterson, J. B. (2006). Self-liking and self-competence separate self-evaluation from self-deception: Associations with personality, ability, and achievement. *Journal of Personality*, *74*, 1047– 1078.
McConnell, A. R., & Strain, L. M. (2007). Content and structure of the self-concept. In C.Sedikides & S. J.Spencer (Eds.) , *The self* (pp. 51– 73). New York: Psychology Press.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741– 749.

Michael, W. B., & Wright, C. R. (1989). Psychometric issues in the assessment of creativity. In J. A.Glover, R. R.Ronning, & C. R.Reynolds (Eds.) , *Handbook of creativity* (pp. 33– 52). New York: Plenum Press.

Nusbaum, E. C., & Silvia, P. J. (2011). Are creativity and intelligence really so different? Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence, 39*, 36– 45.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535– 569.

Plucker, J. A. (1999). Reanalyses of student responses to creativity checklists: Evidence of content generality. *Journal of Creative Behavior*, *33*, 126– 137.

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C.Kaufman & R. J.Sternberg (Eds.) , *Cambridge handbook of creativity* (pp. 48– 73). New York: Cambridge University Press.

Plucker, J. A., & Renzulli, J. S. (1999). Psychometric approaches to the study of human creativity. In R. J.Sternberg (Ed.) , *Handbook of creativity* (pp. 35– 61). New York: Cambridge University Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879– 903.

Rawlings, D., & Locarnini, A. (2007). Validating the creativity scale for diverse domains using groups of artists and scientists. *Empirical Studies of the Arts*, *25*, 163– 172.

Runco, M. A., Plucker, J. A., & Lim, W. (2001). Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal*, *13*, 393– 400.

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, *6*, 24– 30.

Silvia, P. J., Kaufman, J. C., & Pretz, J. E. (2009). Is creativity domain-specific? Latent class models of creative accomplishments and creative self-descriptions. *Psychology of Aesthetics, Creativity, and the Arts*, *3*, 139– 148.

Silvia, P. J., & Kimbrel, N. A. (2010). A dimensional analysis of creativity and mental illness: Do anxiety and depression symptoms predict creative cognition, creative accomplishments, and creative self-concepts?*Psychology of Aesthetics, Creativity, and the Arts*, *4*, 2– 10.

Silvia, P. J., & Nusbaum, E. C. (2010). *What's your major? College majors as markers of creativity*. Manuscript under review.

Silvia, P. J., Nusbaum, E. C., Berg, C., Martin, C., & O'Connor, A. (2009). Openness to experience, plasticity, and creativity: Exploring lower-order, higher-order, and interactive effects. *Journal of Research in Personality*, *43*, 1087– 1090.

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*, 68– 85.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Zabelina, D. L., & Robinson, M. D. (2010). Creativity as flexible cognitive control. *Psychology of Aesthetics, Creativity, and the Arts*, *4*, 136– 143.

Table 1
Creative Achievement Questionnaire (CAQ): Descriptive Statistics

Domain	*M*	*SD*	*Mdn*	*Min, Max*	% 0	% 1
Visual Arts	1.92	3.54	1	0, 35	46.8	23.6
Music	1.63	4.95	0	0, 91	58.0	19.8
Dance	1.52	4.72	0	0, 90	69.6	9.6
Architectural Design	.22	1.19	0	0, 22	91.3	4.8
Creative Writing	1.53	3.01	1	0, 43	37.0	40.7
Humor	1.35	2.04	1	0, 37	29.0	46.8
Inventions	.49	1.36	0	0, 14	73.3	19.5
Scientific Discovery	.41	1.15	0	0, 16	75.6	17.9
Theatre and Film	.72	1.83	0	0, 15	67.6	23.2
Culinary Arts	.78	1.34	1	0, 22	40.4	55.7

Note. Sample size is $n = 848$ for this analysis.

Table 2
Biographical Inventory of Creative Behaviors (BICB): Item Statistics

Item	% Yes	CFA loading	IRT discrimination	IRT difficulty
1. Wrote short story	30.6	.605	.811	.807
2. Wrote novel	5.7	.727	1.131	2.121
3. Organized event	37.6	.607	.815	.506
4. Produced script	5.8	.823	1.547	1.836
5. Designed textile	21.3	.647	.904	1.183
6. Decorated room	54.8	.529	.666	−.201
7. Invented product	10.0	.699	1.043	1.775
8. Drew cartoon	26.6	.633	.872	.950
9. Started club	12.6	.720	1.106	1.535
10. Made picture	29.5	.625	.854	.829
11. Published article	6.6	.738	1.167	1.982
12. Made sculpture	12.9	.745	1.193	1.459
13. Criticized scientific theory	18.1	.654	.922	1.332
14. Made recipes	36.5	.564	.729	.587
15. Produced short film	9.5	.804	1.444	1.567
16. Made webpage	12.8	.662	.942	1.653
17. Created a theory	14.7	.687	1.010	1.473
18. Invented game	25.9	.723	1.115	.874
19. Chosen to lead	42.9	.625	.855	.290
20. Made a present	65.7	.543	.689	−.694
21. Wrote poem	43.5	.528	.663	.304
22. Adapted object	32.1	.653	.920	.691
23. Published research	7.2	.717	1.096	1.982
24. Choreographed dance	15.2	.641	.890	1.545
25. Designed garden	19.2	.588	.776	1.416
26. Made photography portfolio	20.2	.655	.924	1.224
27. Acted	11.4	.706	1.065	1.647
28. Gave speech	48.9	.577	.754	.060
29. Mentored others	48.3	.679	.986	.080
30. Designed experiment	24.4	.695	1.030	.963
31. Wrote jokes	50.9	.466	.562	−.034
32. Served as leader	33.9	.553	.708	.718
33. Composed music	11.6	.709	1.072	1.630
34. Made collage	46.1	.413	.484	.230

Note. Sample size for this analysis is $n = 1256$. The CFA loading is standardized. The item labels are abbreviated stems, not the actual items. The instructions and items are available from Mark Batey.

Table 3
Creative Behavior Inventory (CBI): Item Statistics

Item	*M*	*SD*	*Mdn*	CFA loading
1. Painted picture	1.91	1.03	2	.651
2. Made cards	2.30	1.09	2	.559
3. Made metal crafts	1.43	.75	1	.665
4. Held puppet show	1.63	.80	1	.591
5. Made decorations	2.37	1.05	2	.605
6. Built mobile	1.44	.74	1	.708
7. Made sculpture	1.53	.82	1	.750
8. Published literature	1.43	.76	1	.578
9. Wrote poetry	2.12	1.17	2	.522
10. Wrote play	1.26	.63	1	.654
11. Received art award	1.59	.86	1	.656
12. Received craft award	1.36	.69	1	.748
13. Made plastic craft	1.53	.83	1	.710
14. Made cartoons	1.70	.95	1	.621
15. Made leather craft	1.31	.67	1	.744
16. Made ceramic craft	1.56	.87	1	.741
17. Designed clothing	1.71	.95	1	.679
18. Made floral arrangement	1.79	.99	1	.605
19. Drew picture	1.78	1.07	1	.717
20. Wrote lyrics	1.78	1.04	1	.535
21. Wrote story	1.77	.97	1	.596
22. Planned speech	1.61	.94	1	.618
23. Made jewelry	2.08	1.11	1	.634
24. Had art exhibit	1.49	.85	1	.734
25. Did set design	1.30	.68	1	.707
26. Kept sketch book	1.71	.99	1	.688
27. Made wood craft	1.52	.86	1	.705
28. Designed costume	1.75	.92	1	.634

Note. Sample size for this analysis is $n = 1294$. The CFA loading is standardized. The CBI items were scaled 1 through 4. The item labels are abbreviations, not the actual items. The instructions and items are available from Stephen Dollinger.

Table 4
Revised Creativity Domain Questionnaire (CDQ-R): Item Statistics

Item	*M*	*SD*	*Mdn*	Subscale CFA loading	Subscale
1. Acting	2.97	1.43	3	.642	Drama
2. Algebra/geometry	3.03	1.52	3	.651	Math/science
3. Chemistry	2.53	1.42	2	.763	Math/science
4. Computers/computer science	3.07	1.44	3	.622	Math/science
5. Crafts	4.03	1.40	4	.782	Arts
6. Dancing	3.45	1.50	4	.518	Drama
7. English literature/criticism	3.52	1.42	4	.629	Drama
8. Interior design/decorating	3.74	1.43	4	.726	Arts
9. Keeping a journal/blog	3.33	1.53	3	.613	Drama
10. Leadership	4.01	1.43	4	.663	Interaction
11. Life sciences/biology	2.96	1.47	3	.746	Math/science
12. Logic/puzzles	3.68	1.43	4	.687	Math/science
13. Mechanical abilities	2.80	1.47	3	.655	Math/science
14. Money management	3.66	1.45	4	.545	Interaction
15. Painting/drawing	3.47	1.54	4	.688	Arts
16. Playing with children	4.58	1.35	5	.577	Interaction
17. Selling people things	3.36	1.48	3	.552	Interaction
18. Solving personal problems	4.34	1.22	5	.702	Interaction
19. Teaching/education	4.11	1.37	4	.652	Interaction
20. Vocal performance/singing	2.91	1.59	3	.565	Drama
21. Writing poetry/prose	3.07	1.58	3	.686	Drama

Note. Sample size is $n = 1054$ for this analysis. The CFA loading is standardized. See the text for a description of the subscales and CFA model. The actual items are presented. The instructions and items are published in Kaufman et al. (2009–2010) and are available from James Kaufman.

Table 5
Summary Scale-Level Statistics and Correlations

Scale	*M*	*Mdn*	*SD*	*Variance*	*Min, Max*	1	2	3	4	5	6	7	8	9
1. CAQ sum (raw)	10.605	7	13.422	180.137	0, 176	1								
2. CAQ sum (log)	1.900	2.01	1.069	1.143	−.69, 5.17	.752	1							
3. BICB sum (raw)	8.905	8	6.559	43.016	0, 34	.370	.456	1						
4. BICB (log)	1.911	2.14	.949	.900	−.693, 3.541	.312	.468	.863	1					
5. BICB (IRT)	−.002	0	.916	.838	−1.793, 3.148	.357	.475	.969	.950	1				
6. CBI average (raw)	.671	.571	.514	.264	0, 3	.311	.415	.533	.469	.520	1			
7. CBI (log)	.066	.069	.429	.184	−.693, 1.253	.311	.448	.530	.499	.532	.969	1		
8. CBI (IRT)	−.002	.053	.954	.911	−2.009, 3.899	.293	.435	.525	.495	.528	.946	.981	1	
9. CDQ-R global score	3.489	3.5	.824	.680	1, 6	.240	.421	.385	.363	.382	.456	.473	.467	1

Note. Raw refers to simple sums or averages of raw scores; Log refers to log transformations of the raw scores; IRT refers to trait scores estimated via 2PL IRT models. For this model, the CBI items were scaled 0 through 3 (cf. Table 3). The full matrix that includes subscale scores (the 10 CAQ domain scores and the four CDQ-R domain scores) is available from Paul Silvia.

Table 6
A Classification of the Self-Report Scales

	Does it provide domain scores?	Does it primarily measure eminent achievement or everyday creativity?	Does it measure public behaviors or subjective beliefs?
Creative Achievement Questionnaire (CAQ)	Yes	Eminent achievement	Public behaviors
Biographical Inventory of Creative Behaviors (BICB)	No	Everyday creativity	Public behaviors
Creative Behavior Inventory (CBI)	No	Everyday creativity	Public behaviors
Revised Creative Domain Questionnaire (CDQ-R)	Yes	Everyday creativity	Subjective beliefs

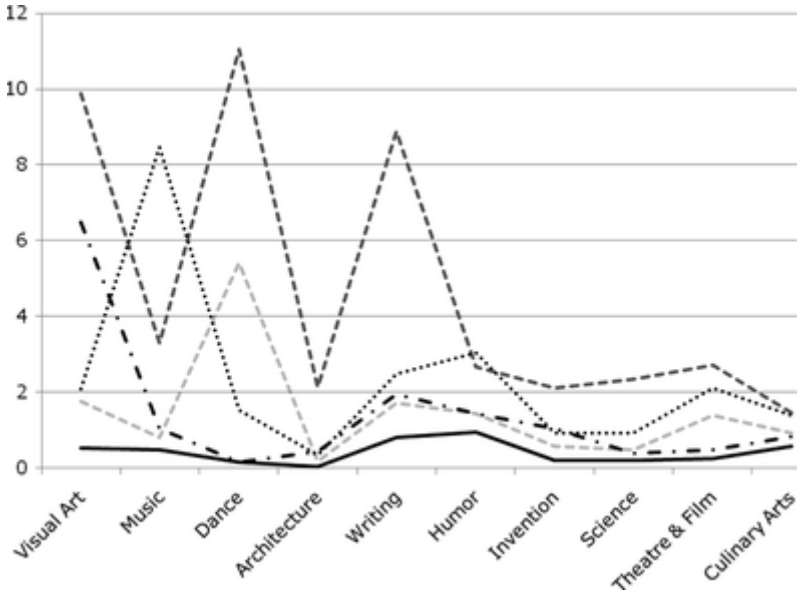

Figure 1. Latent classes of creative achievement in the CAQ.

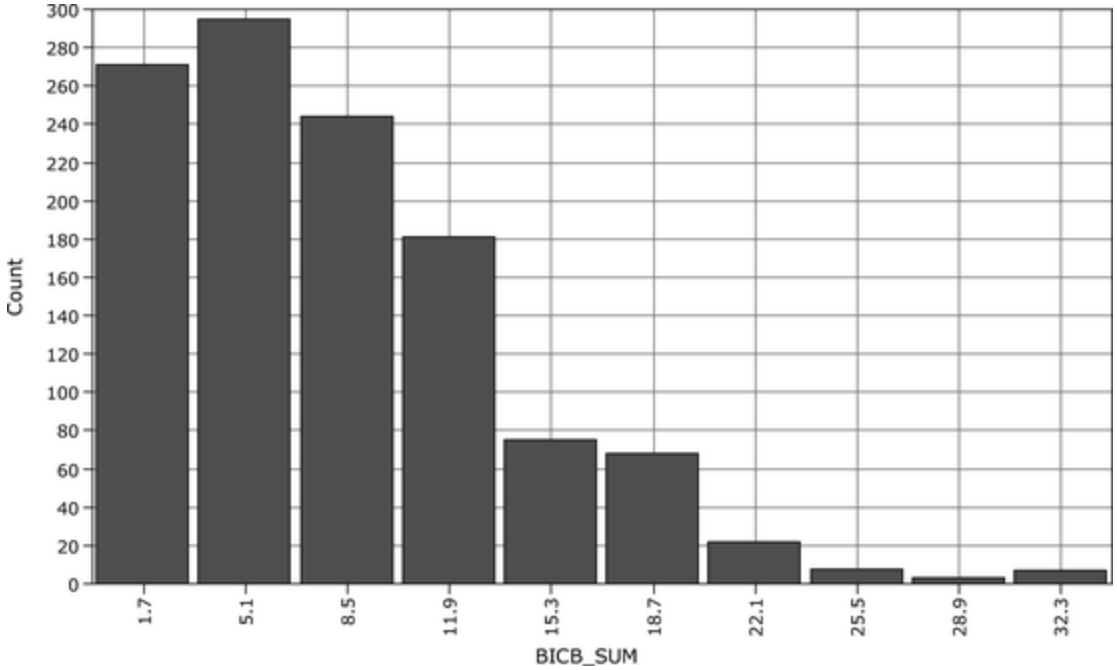

Figure 2. Distribution of BICB scores.

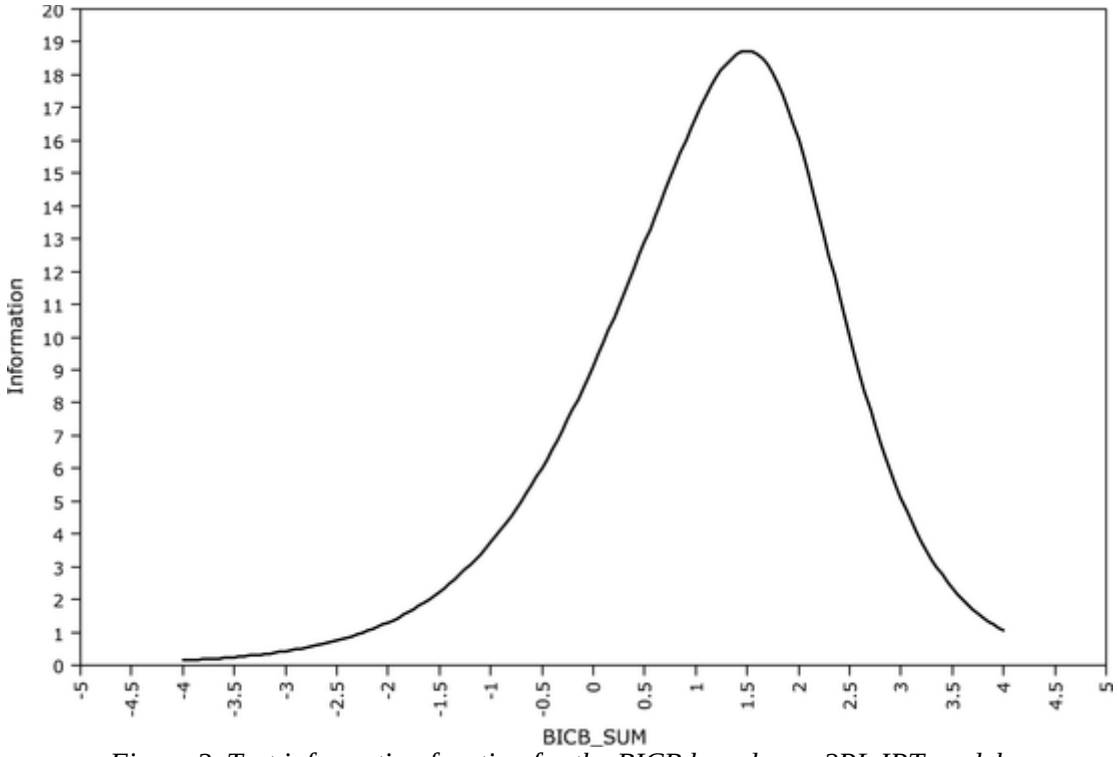

Figure 3. Test information function for the BICB based on a 2PL IRT model.

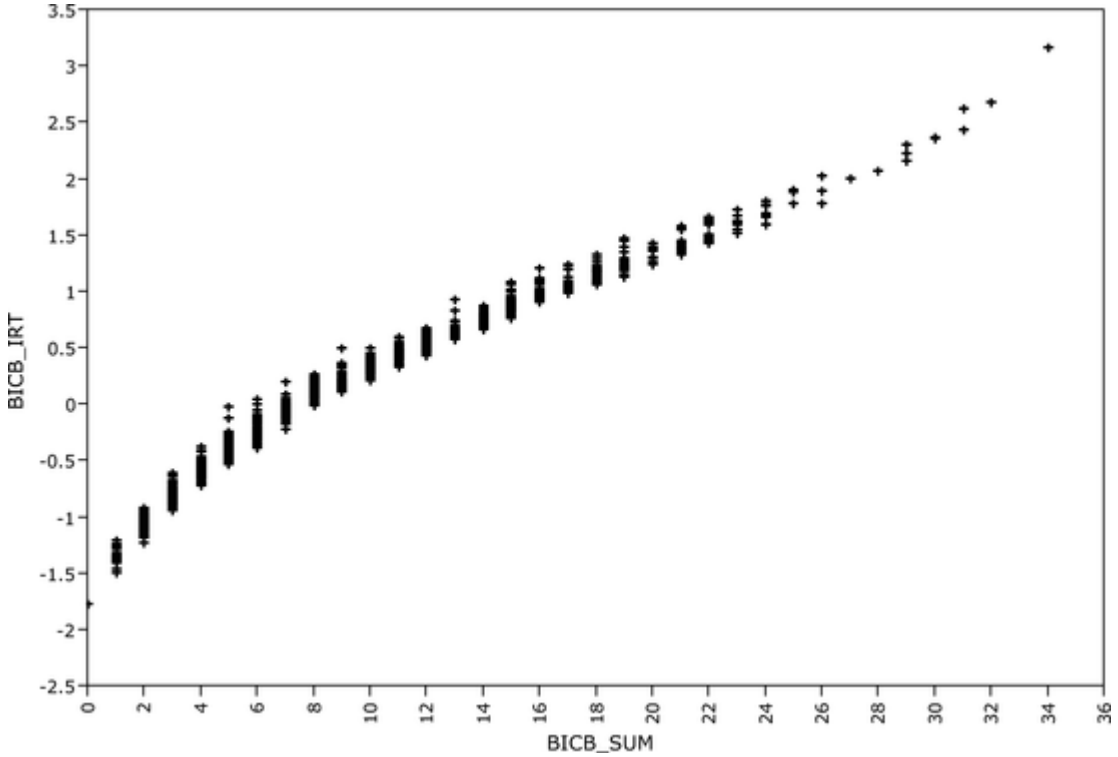

Figure 4. Relationship between simple sum scores and IRT-based trait scores for the BICB.

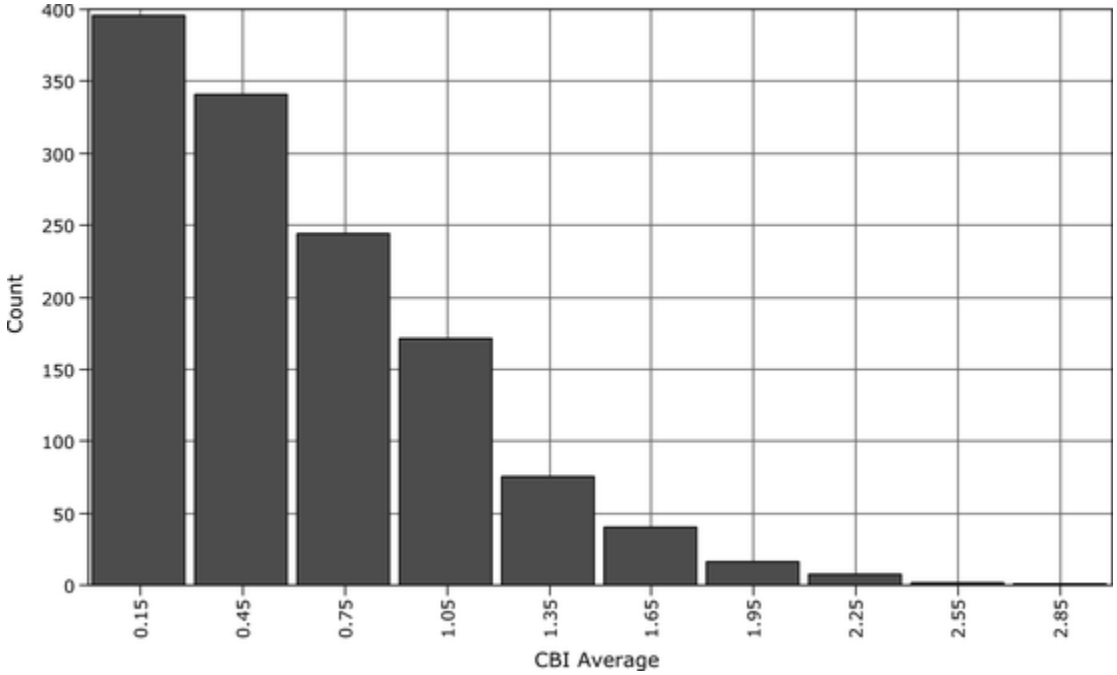

Figure 5. Distribution of CBI scores.

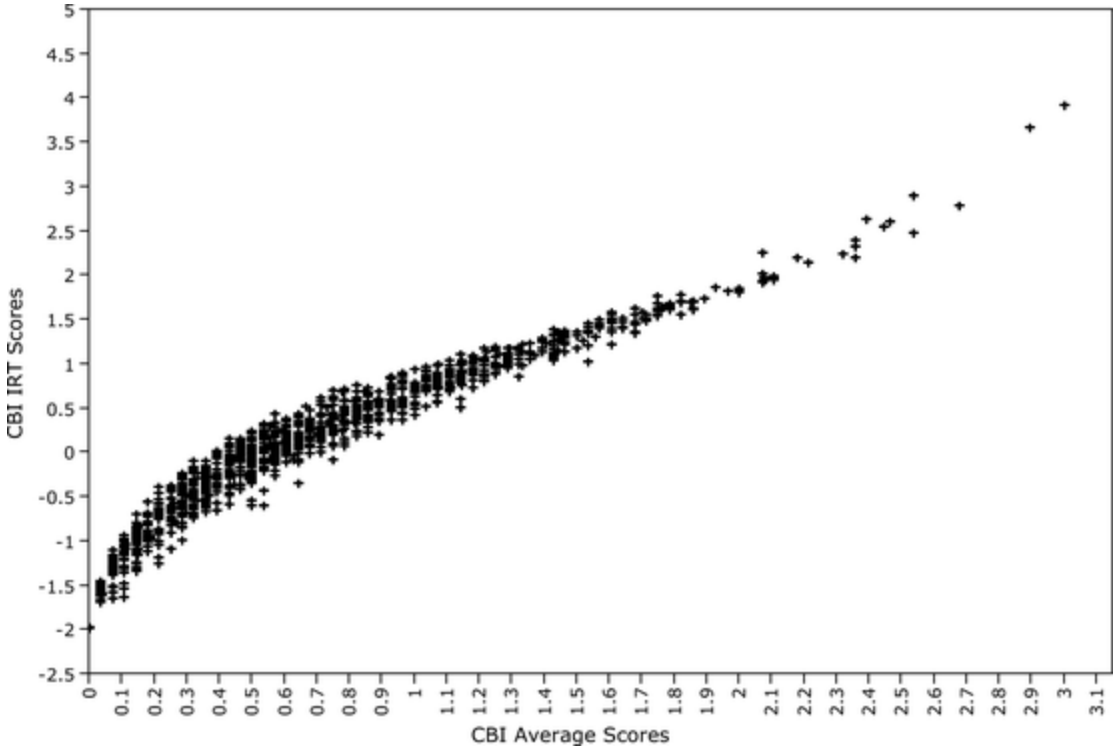

Figure 6. Relationship between simple item averages and IRT-based trait scores for the CBI.