

Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts

By: William Bradley McKibben and [Paul J. Silvia](#)

This is the peer reviewed version of the following article:

McKibben, W.B. and Silvia, P.J. (2017), Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts. *Journal of Creative Behavior*, 51: 57-69. <https://doi.org/10.1002/jocb.86>

which has been published in final form at <https://doi.org/10.1002/jocb.86>. This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Use of Self-Archived Versions](#).

*****© 2015 by the Creative Education Foundation, Inc.. Reprinted with permission. No further reproduction is authorized without written permission from Wiley. This version of the document is not the version of record. *****

Abstract:

Inattentiveness and social desirability might be particularly problematic for self-report scales in creativity and arts research. Respondents who are inattentive or who present themselves favorably will score highly on scales that yield positively skewed distributions and that assess socially valued constructs, such as scales measuring creativity and arts knowledge. A total of 204 undergraduates completed an online survey with several self-report measures (the Creative Achievement Questionnaire, the Biographical Inventory of Creative Behavior, and the Aesthetic Fluency Scale). Many metrics of inattentiveness were included, such as directed response items, self-reported attentiveness, and scales to catch inconsistent and patterned responses. The Balanced Inventory of Desirable Responding measured facets of social desirability. We found high rates of inattentive responding. A latent class analysis indicated that around 25% of the sample was potentially inattentive, and inattentive respondents received higher scores on the creativity and arts scales. Excluding problematic participants increased the effect sizes of interest, consistent with the coarsening effect of careless participants. Social desirability, in contrast, had essentially no relationship with the creativity and arts scales. These findings suggest that social desirability is probably less of a problem than researchers fear, but that inattentiveness is an underappreciated threat. Some practical guidelines are suggested.

Keywords: creativity | carelessness | inattentiveness | validity | social desirability | survey methods

Article:

(In)attention must be paid.

(from the Prolegomenon to the Oakland Poetics Co-Op's Manifesto)

Psychology is in an introspective period about its methods, particularly issues related to replication (Makel, 2014), false positives (Murayama, Pekrun, & Fiedler, 2014), and statistical reasoning (Cumming, 2012). One part of this trend is a renewed interest in participant behavior as sources of bias and threats to validity (Maniaci & Rogge, 2014; Meade & Craig, 2012). Most research participants show great goodwill and diligence, and our appreciation for their contributions to scientific knowledge cannot be overstated. But some participants, deliberately or not, respond in ways that coarsen the data, reduce power, and distort the conclusions we can draw.

In the present study, we build upon and apply recent work on two distorting influences: inattentive responding and socially desirable responding. These issues have received relatively little formal attention in research on aesthetics, creativity, and the arts, but they have been discussed informally and should be studied in light of the current emphasis on effect sizes, power, and reproducibility. After giving an overview of inattentiveness and social desirability and why they might be particularly troublesome for the constructs studied in our field, we report a study that evaluates the distorting effect of problematic participants for measures of creative behavior and arts knowledge. On the basis of these and other findings, we conclude with some practical recommendations.

Inattentiveness and Social Desirability: Distinct Sources of Bias

Inattentive responding, as its name suggests, refers to participants who respond to survey items without regard to the content. People who are responding inattentively may frequently skip items, misread items, or respond without reading items (Johnson, 2005). Sometimes the responses appeared patterned, such as clicking the midpoint response for a string of items or alternating between options. Researchers have shown that inattentiveness is probably quite common. Research by Berry et al. (1992), Johnson (2005), and Oppenheimer, Meyvis, and Davidenko (2009) suggests that as few as 3% or as many as 46% of participants respond inattentively in survey research. In recent work, Meade and Craig (2012) found this number to be between 10% and 12%, and Maniaci and Rogge (2014) found rates of 3–9%.

Inattentive responding is probably more frequent in online surveys than in-person surveys (Johnson, 2005). The lab environment has stronger social norms, fewer distractions, and a researcher to observe the session. In contrast, the online environment largely consists of the researcher's survey, an unpredictable physical environment, and several thousand websites that venerate cats and demonize gluten. Participants who respond inattentively may add noise to data, reduce statistical power and effect sizes, or render an instrument invalid.

Inattentiveness can be assessed with a range of methods. Some methods use questions to flag potentially careless participants. For example, *infrequency scales* ask items for which every attentive participant will give the same answer (e.g., the true/false item “I eat cement occasionally”; Huang, Bowling, Liu, & Li, 2015). *Inconsistency scales* give essentially synonymous items twice, such as early and late in the survey; inattentive participants will show more deviations in their responses. *Directed response items* instruct respondents to give a specific response, such as skipping an item or clicking a specific value (e.g., “Please click *strongly agree*”); someone not reading an item or giving patterned responses will usually

fail to give the directed response. Other methods use features of the data, such as how long it took someone to complete the survey (e.g., someone who blew through a long survey in a few minutes), the number of times a response was repeated (e.g., someone clicking three for a dozen items in a row), and measures of multivariate distance (e.g., Mahalanobis's *D*).

In contrast to inattentive responding, socially desirable responding refers to a participant's tendency to self-present in an overly positive way on survey items (Crowne & Marlowe, 1964). As participants are reacting to the content of the survey, participants are responding based on their reactions and motives rather than to the item content itself. Although an inattentive participant will usually add noise, a participant motivated by social desirability can inject systematic error by giving overly favorable responses. Social desirability yields attentive yet biased responses, thus compromising validity via a different route.

Social desirability has historically been measured with self-report scales. Researchers can then evaluate correlations between self-reported desirability and their constructs of interest. In their pioneering work, Crowne and Marlowe (1960, 1964) developed items that capture a reluctance to admit to unflattering behaviors and thoughts. In recent years, newer scales, such as the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984, 1991), have supplanted earlier ones. The BIDR consists of two subscales whose names and interpretation have varied over the years (Gignac, 2013; Musch, Ostapczuk, & Klaiher, 2012; Paulhus & John, 1998). One facet—known as *self-deceptive enhancement* or *egoistic bias*—reflects a motive to view oneself positively that people might not be aware of. The other—known as *impression management* or *moralistic bias*—reflects a deliberate attempt to appear favorably to others.

Inattentiveness and Social Desirability as Biasing Factors in Creativity and Arts Scales

Are inattentiveness and social desirability a problem in the kinds of self-report instruments common in creativity and arts research? These issues have not received much formal examination, but they are often raised informally in discussions and reviews. Social desirability seems potentially problematic because much of what we measure is culturally valued. Consider, for example, scales that ask people to rate how creative they are in different domains (Kaufman, 2012), how easily they come up with creative ideas (Plucker, Runco, & Lim, 2006), how many creative activities they have done in the past year (Batey, 2007; Hocevar, 1979), how confident they feel during creative tasks (Karwowski, 2011), how much they know about the arts (Smith & Smith, 2006), or how many creative achievements they have in different areas (Carson, Peterson, & Higgins, 2005). If most people would prefer to be seen as creative, knowledgeable sorts who have good ideas, then self-report measures of creativity could be biased by social desirability.

In some of our past writing, we have dismissed, perhaps too lightly, the role of social desirability. In a review of self-report scales (Silvia, Wigert, Reiter-Palmon, & Kaufman, 2012), we pointed out that most scales that measure creative achievements, behaviors, and knowledge have highly skewed distributions, in which most scores pile up at the scale's floor. If most people have very low scores, we suggested, then widespread exaggeration seems unlikely. In hindsight, the issue is not whether exaggerated favorable responses are widespread, but rather whether the findings would be cleaner and more replicable if a small subgroup of problematic responses

could be objectively identified and excluded. One need not have many problematic cases for power and effect sizes to suffer appreciably (Maniaci & Rogge, 2014).

Inattentiveness could be a serious problem for many popular scales. Inattentive participants commonly mark the midpoints of a scale or give random or patterned responses, such as alternating around the midpoint values. For a scale with normally distributed scores (e.g., scales for personality traits), such people add noise but otherwise end up among the middle of the distribution, like most people. But for scales with positively skewed distributions, an inattentive participant will end up far above the mean. If someone responds inattentively to a scale like the Creative Achievement Questionnaire, a scale that yields very low average scores, then he or she will be pegged as a highly creative person. In short, for the skewed scales common in our field, inattentive participants will distort the estimate of the sample mean and act as outlying influential cases (Stevens, 1984).

Not much research has evaluated creativity assessment in light of inattentive and socially desirable responding, but the results so far are discouraging. One recent study (Runco et al., 2014) evaluated the RIBS, a self-report scale measuring ideational behavior (Runco, Plucker, & Lim, 2000–2002). The long form and short form of the RIBS were given along with (1) a group of reverse-scored “contraindicative” items; (2) a “distractor scale” consisting of construct-irrelevant items that should not correlate with the RIBS (e.g., “I am reflective”); and (3) a “lie scale” (also known as an infrequency scale) that asks items that nearly everyone would agree or disagree with, thus allowing researchers to catch inattentive or random responding. Both the short and long form of the RIBS correlated over .70 with the distractor scale (i.e., a set of irrelevant items) and over .50 with the lie scale (i.e., a set of items that everyone should reject or endorse). These correlations suggest a common “response set” that is consistent with inattentive responding. If a subset of participants gives every item similar responses, then correlations between scales sharing a method will be inflated.

The Present Research

Our study had two aims. First, we wanted to investigate the prevalence of inattentive responding and socially desirable responding in an online survey of creativity and the arts and to evaluate their potentially distorting effects. Do inattentive responders end up with higher scores on self-report measures of creativity and art knowledge? Do they appreciably influence effect sizes and the conclusions one would draw? We wanted to evaluate a “worst case scenario” so we could estimate the effects of these distorting factors at their grimmest. We thus collected data using a method and a sample that should maximize carelessness. For the method, we collected data online and used a lengthy survey that took roughly 30 minutes, both factors known to reduce attentiveness (Johnson, 2005; Meade & Craig, 2012). For the sample, we collected data from college student participants in an undergraduate research pool during the end of a semester, a time known to attract relatively less conscientious participants (Stevens & Ash, 2001). With a worst-case scenario as a benchmark, we can put the normal range of bias into perspective.

Second, we wanted to apply and illustrate some practical methods for identifying these distorting influences and suggest some good practices for future research. If inattentiveness and social desirability have non-trivial influences, then researchers should include methods for identifying

problematic cases as part of standard research practice. Our study describes and applies some straightforward methods that are easy to work into the kinds of designs and analyses that are common in creativity and arts research.

Methods

Participants

A total of 204 undergraduate students (133 women, 68 men, and three who declined to state) from UNCG volunteered to participate in this study and received credit toward a research option in a psychology course. Age ranged from 18 to 28 years ($M = 19.02$ years, $SD = 1.29$). The sample was diverse, according to self-reported racial identification: 31% African American, 7% Asian-American, 55% Caucasian, and 3% Native Alaskan/American Indian. (People could choose more than one category or decline to choose any.) Regarding ethnicity, 9% identified their ethnicity as Hispanic or Latino; 4% of the sample indicated that English was not their native language. Eleven people (5%) began but did not complete the study.

Procedure

Participation was completely online. People who wished to participate signed up for the study on Experimentrix, a research participation webpage, and were directed to the survey, which was delivered via Qualtrics. Participants first saw the informed consent form. People who provided consent could then complete the survey. Upon completion of the study, participants were redirected to a separate survey in which they could enter their name and e-mail address to receive credit for participation. This second survey was not connected to the first, so the survey responses remained anonymous.

Given our goal of evaluating the worst-case scenario, data collection started mid-semester and ended on the semester's last day. Informal researcher lore suggests—and empirical research shows (Stevens & Ash, 2001; Witt, Donnellan, & Orlando, 2011)—that the start of the semester yields relatively more conscientious participants, whereas the end of the semester yields a higher proportion of participants who are flakier, disinterested, and noncompliant. The findings should thus represent the upper end of inattentiveness that researchers using similar procedures should expect to find.

Substantive Measures

Aesthetic fluency scale (AFS)

The AFS (Smith & Smith, 2006) was designed to measure knowledge about the arts. It does so by presenting people with a list of 10 important figures, periods, and ideas in art history (e.g., Mary Cassatt, Fauvism, Abstract Expressionism, Impressionism), and people then rate how much they know about each one on a 5-point scale, in which 0 = *I have never heard of this artist or term*, 1 = *I have heard of this but don't really know anything about it*, 2 = *I have a vague idea of what this is*, 3 = *I understand this artist or idea when it is discussed*, and 4 = *I can talk intelligently about this artist or idea*. The scale has been used in many studies (e.g., DeWall,

Silvia, Schurtz, & McKenzie, 2011; Silvia, 2007; Silvia & Barona, 2009; Silvia & Nusbaum, 2011), and variants for specific domains, such as literature, decorative arts, and film (e.g., Silvia & Berg, 2011), have been used as well. In this study, the scale demonstrated good internal consistency reliability for the full sample (Cronbach's $\alpha = .80$). We included four additional items related to literature and creative writing (Toni Morrison, Italo Calvino, Salman Rushdie, *Infinite Jest*)—along with a bogus item (The Oakland Poetics Co-Op) described later—but this brief scale showed poor internal consistency, $\alpha = .50$.

Authenticity inventory—3 (AI-3)

The AI-3 (Goldman & Kernis, 2004; Kernis & Goldman, 2006) is a 45-item questionnaire scored on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The AI-3 yields an overall composite score of dispositional authentic functioning as well as scores along four subscales (awareness, unbiased processing, behavior, and relational orientation) that comprise authenticity. Higher scores on the AI-3 reflect higher authentic functioning. In this study, the AI-3 showed good internal consistency (Cronbach's $\alpha = .86$).

Biographical inventory of creative behaviors (BICB)

The BICB (Batey, 2007) was designed to measure everyday creativity (Richards, 2007, 2010), the common “little c” creative hobbies and activities that people do in daily life. The scale asks people if they have done 34 different creative activities in the past year (e.g., “Written a short story,” “Choreographed a dance,” “Produced a short film,” “Made someone a present”). People marked the activities they had done with an X. The number of marked responses was summed to yield an overall score of creativity ($\alpha = .90$); higher scores reflect more frequent engagement in creative behaviors. The scale has been used in many recent studies (e.g., Silvia et al., 2014) and has good psychometric properties (Silvia et al., 2012).

Creative achievement questionnaire (CAQ)

The CAQ (Carson et al., 2005) is a popular self-report scale for measuring creative achievement in a range of domains (visual arts, music, dance, architectural design, creative writing, humor, inventions, scientific discovery, theater and film, and culinary arts). Unlike many other scales, it is intended to capture high-level achievements, so few people receive high scores. The scoring for the CAQ is intricate and described elsewhere (Carson et al., 2005; Silvia et al., 2012). In this study, we created scores for each of the 10 domains and then summed them for an overall CAQ score.

HEXACO-60 personality inventory

The HEXACO-60 (Ashton & Lee, 2009) measures six major dimensions of personality: Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to Experience. Most of the dimensions broadly resemble the traditional big five factors (Ashton, Lee, & de Vries, 2014). The instrument contains 60 items scored on a Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Our primary interest was in Openness to Experience (Cronbach's $\alpha = .73$). Because openness strongly predicts creativity across a range of creative

outcomes (e.g., Karwowski, Lebuda, Wisniewska, & Gralewski, 2013; Kaufman, 2013), it can be used as a benchmark variable for understanding how including or excluding problematic participants affects substantive findings.

Perceptual aberration scale and social anhedonia scale

We included two of the four subscales from the recent short forms of the Wisconsin Schizotypy Scales (Gross, Silvia, Barrantes-Vidal, & Kwapil, 2012; Winterstein et al., 2011). The Perceptual Aberration Scale has 15 items dealing with unusual bodily experiences (e.g., “Sometimes I have had a passing thought that some part of my body was rotting away”). The Social Anhedonia Scale has 15 items dealing with social disinterest (e.g., “Having close friends is not as important as many people say”). People completed each item using a True/False scale, scored 1 or 0. The items were scored so that higher scores represent more deviant levels of the traits. A 7-item infrequency scale, described later, was mixed within the 30 items.

Validity Checks

We evaluated a family of methods that might potentially catch inattentive and socially desirable responding. Many of these methods are long-standing tools in the survey-research toolkit; others are newly developed as part of recent research on detecting inattentiveness (Maniaci & Rogge, 2014).

Response inconsistency

The “attentive responding scale” was developed by Maniaci and Rogge (2014) to catch inconsistent patterns of responding. We used the brief form, which consists of two matched 6-item scales. The two sets have nearly synonymous items completed on a 5-point (1 = *not at all true*, 5 = *very true*) scale. The first set, for example, presents the item “I am a very energetic person,” and the second set presents the item “I have a lot of energy.” People complete one set at the start of the survey and the other at the end. As the item content is nearly identical, variance in responses is likely due to inattentive or random responding. For each item pair, we calculated the absolute difference. These differences were then summed to create an overall response inconsistency score that can quantify how inattentive someone likely was.

Infrequency scale

Infrequency scales present items for which essentially all participants should give the same answer. We used a revised version of the Chapman and Chapman (1983) infrequency scale, which was developed as part of the Wisconsin Schizotypy Scales (Kwapil, Barrantes-Vidal, & Silvia, 2008). The scale has seven items that people complete using a True/False format. Given the nature of the items—for example, “At times when I was tired, I yawned or stretched” and “On some nights, I haven't fallen asleep immediately upon getting into bed”—aberrant responses are presumed to reflect careless responding or a “fake-bad” response set. The items were embedded within the Physical Anhedonia and Social Anhedonia scales, which have the same True/False response format. Past research has suggested a cutoff of two for this scale (Kwapil et al., 2008).

Directed response items

Participants who do not read or who “click through” items can be flagged with items that instruct them to give a specific response. We included two such items, known as directed-response items, in two of the scales. In the AI-3, we added an item that read, “This is a system check item. Please mark Strongly Disagree.” In the HEXACO-60, we added an item that read, “This is a system check item. Please mark Strongly Agree.” We chose extreme values because they are more likely to catch inattentive participants who simply click the scale's midpoint. These items were placed roughly at the midway point of each survey, which is when inattentiveness tends to increase (e.g., Berry et al., 1992). An accurate response was scored as a 0, which indicated attentive responding. People who gave any response other than the directed response received a 1. Scores for each item were summed for a composite inattentiveness score, with higher scores reflecting higher inattentive responding. We used only two items based on recommendations from Meade and Craig (2012), who speculated that more than three directed response items might irritate participants (p. 452).

Overclaiming

On the aesthetic fluency scale, we included a bogus item (e.g., The Oakland Poetics Co-Op) as a validity check item. Higher responses to this item probably reflect either socially desirable or inattentive responding, given that the Co-Op does not exist. We reasoned that participants who indicated that they have heard of or can speak about the Co-Op either were not paying attention or were trying to present themselves in an overly positive way. Prior to including the item, we scoured the Internet to ensure that there is no such organization that participants might find while completing the survey online. Thus, if attention was being paid by participants in the study, honest responses to the item should be low. Similar items and scales have been used in past work to evaluate if people are overclaiming knowledge (e.g., Musch et al., 2012).

Survey time

The survey's duration—the time from when participants started the survey to when they submitted their data—was measured by the survey software. As in past work (Maniaci & Rogge, 2014), we truncated extreme values. Nine scores over 100 minutes—the 95% point of the scores and a natural joint in the distribution—were recoded to 100. Time was expressed in decimal time (i.e., a value of 20.75 is 20 minutes and 45 seconds).

End-of-survey self-reported response quality

We included three questions at the end of the survey to assess self-reported attentiveness. These items were developed by Meade and Craig (2012) for detecting inattentive responding. The questions stated: “I put forth ____ effort towards this study” (scored from 1—*Almost No* to 5—*A Lot Of*), “I gave this study ____ attention” (scored from 1—*Almost No* to 5—*My Full*), and “In your honest opinion, should we use your data in our analyses in this study?” (scored as 1—*Yes* or 0—*No*). Instruction sets for the first two questions detailed by Meade and Craig also were

included. In their study, Meade and Craig found that the “keep your data” question performed well at identifying inattentive respondents, and the other two demonstrated mixed utility.

Balanced inventory of desirable responding (BIDR)

The BIDR (Paulhus, 1984; Paulhus & Reid, 1991) was designed to measure socially desirable responding. In this study, we used a 20-item BIDR short form developed by Steenkamp, De Jong, and Baumgartner (2010) as part of their multination study. On this short form, 10 items measure moralistic response tendencies (MRT; e.g., “I have said something bad about a friend behind his or her back”) and 10 items measure egoistic response tendencies (ERT; e.g., “My first impressions of people usually turn out to be right”), which mirror an earlier distinction between impression management and self-deceptive enhancement. The items were completed using a 1 (*not true*) to 7 (*very true*) response scale. Steenkamp et al. (2010) used the short form in a study of over 12,000 participants in 26 countries. Across the 26 countries, average reliability coefficients for ERT and MRT subscales were .67 and .73, respectively. In this study, Cronbach's α for the ERT and MRT subscales were .53 and .62, respectively.

Data Analysis

Given the project's aims, we adopted an exploratory data analysis approach (Tukey, 1977) that focused on descriptive statistics and effect sizes (Cumming, 2012). Confidence intervals are in square brackets. Our analyses first explored the prevalence of inattentiveness and social desirability. We then examined how the many metrics of inattentiveness and social desirability predicted the self-reported measures of creativity and art expertise. Finally, we applied latent class analysis to identify “keep” and “drop” classes and evaluated how substantive effects (e.g., how openness to experience predicts self-reported creativity and art expertise) varied when problematic participants were removed. Readers interested in the raw data can download it (in SPSS and Mplus formats) from Open Science Framework: <https://osf.io/suqdz/>.

Results

Partial Responders

Eleven people, around 5% of the sample, did not complete the survey. Some people completed some of it, but most of their responses were too sparse, so we omitted them from subsequent analyses. Nevertheless, the partial responders offer some useful lessons. Of the three partial responders who completed the CAQ, for example, two were greatly above the sample average, including someone who endorsed every item for every subscale, yielding an incredibly high score.

Measures of Carelessness and Social Desirability

Attentive responding scale

Following Maniaci and Rogge's (2014) approach, we computed attentive responding scale scores by taking the absolute deviation for each of the six item pairs and then summing the deviations.

Maniaci and Rogge's research suggested a cutoff of 6.5. Twenty people (10.4%) exceeded the cutoff, suggesting that they were likely inattentive, and an additional 15 (7.8%) had a score of 6.

Infrequency scale

How many people endorsed the bizarre items on the 7-item infrequency scale? Research with this scale has conventionally used a cutoff score of 2 (Kwapil et al., 2008). Twenty-one people (10.9%) exceeded the cutoff, and another 9% had a score of 2.

Directed response items

Our two directed response items asked people to click *Strongly Disagree* or *Strongly Agree*, respectively. A large proportion of participants missed at least one. For the item embedded in the HEXACO, 48 people (24.9%) gave the wrong response; for the item embedded in the AI-3, 53 people (27.5%) gave the wrong response. Overall, 63.5% of the sample got both right, 20.3% of the sample missed one, and 16.1% missed both.

Knowledge claiming

The Oakland Poetics Co-Op was suitably obscure to the participants. Virtually everyone gave it a 0, indicating that they had not heard of it, a few people (3.1%) gave it a 1, and fewer still (2.1%) gave it higher than a 1.

Self-reported effort, attention, and data use

How did people rate their own effort and attention? Participants rarely said they gave “almost no” or “very little” effort (5.7% gave a 1 or 2 on the 5-point scale) or attention (6.3% gave a 1 or 2). When asked whether the researchers should use their data, 14.1% of the sample said that their data should be omitted.

Survey duration

On average, people took roughly half an hour to complete the survey ($M = 29.43$, $SD = 10.18$, $Mdn = 24.57$). The range, however, was large, from 5.63 minutes to 100 minutes, our truncated value. Around 16% of the sample completed the study in less than 15 minutes; around 12% required more than 45 minutes.

Social desirability

Scores for the two short-form BIDR subscales—egoistic response tendencies and moralistic response tendencies—fit past research using the BIDR. The distributions were fairly normal, and the means were centered near the middle of the response scale (egoistic $M = 4.12$, $SD = .65$, range = 2.20–6.00; moralistic $M = 4.11$, $SD = .80$, range = 1.40–6.20). The sample was thus not unusually high or low in social desirability.

Table 1. Descriptive Statistics and Correlations

Measure	<i>M</i>	<i>SD</i>	<i>Mdn</i>	Min, Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Attentive Responding Scale	3.42	2.62	3.00	0, 13	1													
2. Infrequency Scale	.80	1.32	0	0, 5	.24	1												
3. Directed Response Items	.53	.76	0	0, 2	.29	.61	1											
4. Oakland Poetics Co-Op	.08	.37	0	0, 3	.12	.21	.12	1										
5. Reported Effort	3.89	.83	4.00	1, 5	-.24	-.37	-.50	-.17	1									
6. Reported Attention	4.02	.97	4.00	1, 5	-.28	-.41	-.52	-.06	.65	1								
7. Exclude Own Data	1.14	.35	1	1, 2	.16	.38	.39	.12	-.45	-.46	1							
8. BIDR: Egoistic Bias	4.12	.65	4.10	2.20, 6.00	-.04	.01	-.06	.04	.13	.01	-.08	1						
9. BIDR: Moralistic Bias	4.11	.80	4.00	1.40, 6.20	-.05	.00	-.03	.12	.11	.04	-.01	.40	1					
10. Survey Time	29.41	20.19	24.57	5.63, 100.00	-.01	-.27	-.24	-.11	.26	.17	-.19	.07	-.02	1				
11. CAQ Total	13.21	25.78	6.00	0, 280	.23	.19	.12	.47	-.17	-.04	.05	.01	.05	-.07	1			
12. BICB	8.53	6.36	7.00	1, 34	.28	.04	.15	.22	-.02	-.15	-.06	.01	.00	.08	.36	1		
13. Aesthetic Fluency Scale	.65	.57	.50	.00, 3.30	.07	.04	.07	.35	.02	.06	-.10	.05	.08	.03	.33	.29	1	
14. Openness to Experience	3.12	.58	3.10	1.60, 4.50	-.07	-.01	-.14	.17	.10	.15	-.07	.14	.11	-.03	.23	.20	.36	1

Notes. $n = 193$. Higher scores for variables 1–4 represent higher inattentiveness. Higher scores for 5 and 6 represent higher self-reported effort and attentiveness, and a higher score for 7 indicates that one's data should be excluded. Survey time is in decimal time minutes (e.g., 2.50 equals two and a half minutes, or 2:30); values higher than 100 minutes were truncated to 100.

Did Inattentiveness and Social Desirability Predict Creativity and Art Expertise?

Clearly, there was variability in attentiveness and social desirability. Most of the metrics flagged at least 10% of the sample as potentially problematic, and there was a wide range of social desirability scores. Did it matter for the substantive constructs? Table 1 displays the correlations of self-reported creativity, art expertise, and the metrics of carelessness and social desirability. As a comparison, we included the HEXACO openness to experience scale. Because it has strongly predicted these art and creativity scales in many past studies, openness can serve as a criterion construct for benchmarking the effects

For the creativity scales, inattentiveness did cause an upward bias in the scores: people who responded more carelessly had higher CAQ and BICB scores (see Table 1). For the CAQ, for example, scores increased as people scored higher on the attentive responding scale ($r = .23$ [.09, .36]), made more aberrant responses to the infrequency scale ($r = .19$ [.05, .32]), missed more directed response items ($r = .12$ [-.02, .26]), and claimed to know about the Oakland Poetics Co-Op ($r = .47$ [.35, .57]). For the BICB, a similar pattern appeared. BICB scores increased as people scored higher on the attentive responding scale ($r = .28$ [.14, .41]), missed more directed response items ($r = .15$ [.01, .29]), and claimed to know about the Oakland Poetics Co-Op ($r = .22$ [.08, .35]). To put these effect sizes into perspective, many of these coefficients are larger than the effects of schizotypy and extraversion on creativity (Acar & Runco, 2012; Batey & Furnham, 2006), and they are about as large as the effects of the HEXACO openness to experience scale on the CAQ ($r = .23$ [.09, .36]) and BICB ($r = .20$ [.06, .33]) in the full sample.

Social desirability, on the other hand, seemed immaterial. For both BIDR subscales, the correlations with the CAQ, BICB, and aesthetic fluency scales were essentially zero. Likewise, survey time correlated with several other markers of inattentiveness but had essentially no relationships with the substantive constructs (see Table 1).

Latent Carelessness Classes

One issue with the many validity metrics is how to synthesize them into a holistic judgment about whether to retain or drop a participant. As in past work (Maniaci & Rogge, 2014; Meade & Craig, 2012), we applied latent class analysis to the many markers of inattentiveness and social desirability to classify our participants into groups. Past studies have found a large “keep” class and one or two smaller “drop” classes representing kinds of inattentive responding. We *a priori* specified a model with two latent classes, which we expected to represent “keep” and “drop” classes. We did not evaluate models with three or more classes because our sample size is too small to identify and discriminate between small classes (Bollen, Harden, Ray, & Zavisca, 2014; Tein, Coxe, & Cham, 2013). The latent class indicators were the attentive responding scale, the infrequency scale, the sum of the directed response items, the Oakland Poetics Co-Op item, self-reported effort, self-reported attentiveness, people's “use my data” decision, the BIDR social desirability subscales, and survey time.

A latent class model, estimated in Mplus 7.3, found two sharply separated classes: a large class made up of 75.5% of the sample, and a smaller class made up of 24.5%. Measures of classification quality indicated good class separation (entropy = .955). Figure 1 depicts the

values of the two classes on the measures. (The variables were standardized to put them on the same scale.) Some measures clearly differentiated the two groups: the larger class had lower scores on the attentive responding scale and infrequency scale, missed fewer of the directed-response items, spent more time on the survey, and gave higher scores on the self-reported effort and attention items. Interestingly enough, the two groups were essentially the same on the self-reported “use my data” item, the Oakland Poetics Co-Op item, and on both BIDR subscales.

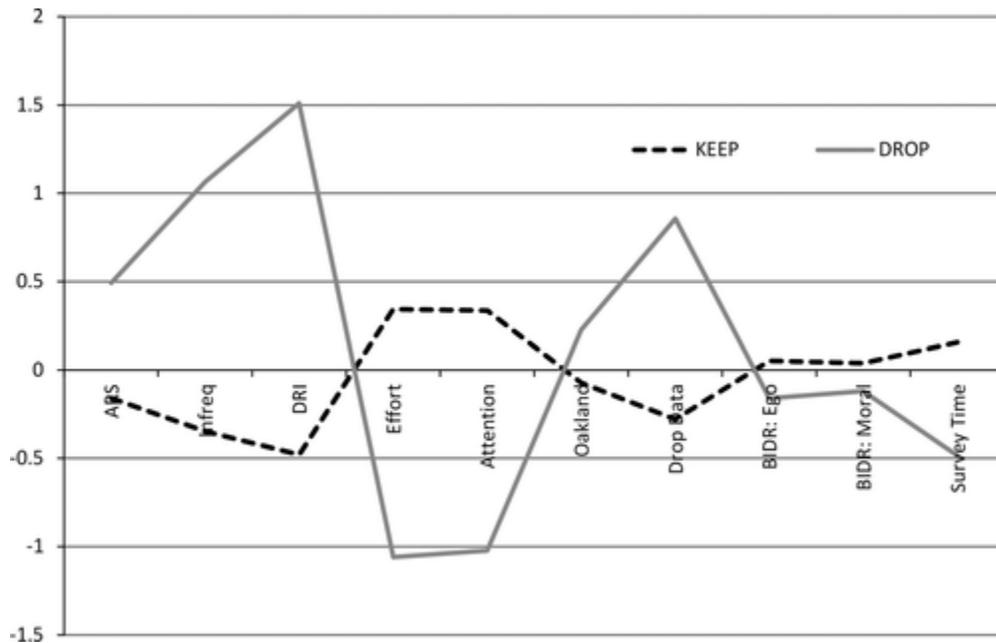


Figure 1. A two-class latent class model based on indicators of attentiveness and social desirability. *Notes.* All indicators have been standardized. The indicators, in order, are scores on the attentive responding scale (ARS), the number of infrequency scale items missed (Infreq), the number of directed response items missed (DRI), self-reported Effort and Attention, scores for the bogus “Oakland Poetics Co-Op” item, whether people said we should drop their data, scores on the BIDR egoistic (BIDR Ego) and moralistic (BIDR Moral) scales, and survey time.

Not all researchers have the software or background needed for latent class analysis, so we evaluated some simple cutoff score methods that would yield similar groupings. The latent classes can be closely approximated by selecting participants who meet three criteria: missing no more than one of the two directed response items, scoring less than 3 on the infrequency scale, and scoring less than 7 on the attentive responding scale.

The latent class analysis revealed several things. First, we indeed got a worst-case scenario. The model estimated that 25% of the participants were potentially problematic, which is much higher than the estimates in recent work (Maniaci & Rogge, 2014; Meade & Craig, 2012). Second, consistent with past work, the attentive responding scale, infrequency scale, survey time, and directed-response items effectively identified problematic participants (Huang et al., 2015; Maniaci & Rogge, 2014), but simply asking people if we should use their data did not. And third, the lack of difference in BIDR subscales supports the view that inattentiveness and social desirability are different kinds of threats to score validity.

Evaluating the Influence of Problematic Cases

So far, we have shown that many of the markers of inattentiveness correlate with our substantive constructs and that a notable number of participants were flagged statistically as probably problematic. The reason for assessing inattentiveness is to have an objective basis for screening out problematic cases, which should improve data quality, power, and reproducibility.

In our sample, omitting the problematic cases appreciably improved the results. We used the latent class analysis as an objective basis for exclusion: participants in the “drop” class were omitted. We then compared results with the “keep” class to results for the full sample. For illustration, we focused on relationships between openness to experience and the substantive constructs, as such hypotheses are actively studied in modern work. As Table 2 shows, openness to experience correlated more strongly with the CAQ, BICB, and aesthetic fluency scale when the problematic cases were dropped; the CAQ, BICB, and aesthetic fluency scales also correlated more strongly with each other. As expected, then, dropping inattentive participants increased the effect sizes of interest, indicating that people in the “drop” class were adding substantial error.

Table 2. Correlations between Openness to Experience and Art and Creativity Variables for the Full Sample (Above the Diagonal) and the “Keep” Subsample (Below the Diagonal)

Construct	1.	2.	3.	4.
1. Openness to Experience	1	.23 [.10, .37]	.20 [.07, .34]	.36 [.24, .48]
2. CAQ	.34 [.23, .45]	1	.36 [.23, .48]	.33 [.20, .45]
3. BICB	.25 [.09, .41]	.55 [.46, .65]	1	.29 [.16, .42]
4. Aesthetic Fluency	.45 [.33, .58]	.46 [.30, .58]	.37 [.19, .55]	1

Note. 95% confidence intervals are in brackets.

Discussion

As our field becomes more concerned with issues of replication, good research practices, and meta-analysis (Makel, 2014), understanding participant factors that coarsen the data becomes particularly important. In this study, we evaluated a wide range of metrics of inattentiveness and social desirability, two factors that have not received much attention in creativity research. For inattentiveness, we found substantial variation in all of the metrics, and a latent class analysis suggested that about a quarter of the participants were probably inattentive. This percentage is higher than in previous studies of attentiveness (Maniaci & Rogge, 2014; Meade & Craig, 2012), and it supports our original concern that inattentive responding is a common and serious threat to validity in online survey research.

As we suspected, inattentive participants ended up among the higher scorers on the creativity and arts scales. Inattentive people who give patterned responses—such as alternating between options or clicking the midpoint—or random responses will end up as high scorers for scales that yield low average scores. For the CAQ, for example, the typical undergraduate in our past studies had a total score of less than 2 (e.g., Silvia, Kaufman, & Pretz, 2009), whereas some of our participants flagged as inattentive in this sample had scores in the hundreds. For the BICB, the typical participant in our work endorsed around only 20% of the items, but someone responding carelessly (such as alternating) will tend to click *yes* to more than that. As a result, inattentiveness will bias the scores upward, yielding a small cluster of high but invalid scores.

Consistent with past work, dropping the problematic cases increased most of the effect sizes, sometimes substantially, thus demonstrating the coarsening effect of careless participants.

Social desirability, on the other hand, was much less important than we expected. Many creativity researchers would suspect that participants would overclaim creative knowledge, skills, and abilities, given creativity's culturally valued status. In our study, however, the two subscales of the BIDR had essentially no relationship with the measures of creativity and arts knowledge. The means, variability, internal consistency, and distribution shapes resembled much past research, so the subscales' near zero effect sizes are probably trustworthy. The role of social desirability nevertheless deserves more attention in future research, but this study suggests that it is probably less problematic than inattentiveness.

We should point out that we sought to evaluate a worst-case scenario. We collected data from the middle of the semester up through the final hour, so we captured the peak period for procrastinators. And the online survey format, for many reasons (Johnson, 2005), is more prone to careless and distracted responding. The typical study should thus be less afflicted than ours. Nevertheless, online methods are increasingly common in creativity research, and people can be flakey and inattentive anywhere, so we would encourage researchers to put into place some simple methods for evaluating inattentiveness.

In our project, the items that discriminated between attentive and careless participants were the 12 paired items from Maniaci and Rogge's (2014) attentive responding scale, the 7-item infrequency scale, the two directed response items, survey time, and the two self-reported effort and attentiveness items. Overall, not many items are needed to discriminate between attentive and inattentive participants. Maniaci and Rogge (2014) suggested that the attentive responding scale and survey time could suffice, and that even a single directed response item would be fruitful. Brief infrequency scales are easy to work into most surveys and effectively catch low-effort participants (Huang et al., 2015). For users of the aesthetic fluency scale, we would recommend including at least one overclaiming item like the Oakland Poetics Co-Op. Only a few people indicated knowing a lot about it, but researchers will want to be able to identify and evaluate that small group in their data.

In future research, researchers should investigate moderators of inattentiveness and social desirability effects. For example, in this study, we found that social desirability was not particularly prevalent; however, this could be because our participant sample represented a broad array of college majors and few were in an art or creativity-related field. Thus, these students may not have felt the need to present themselves as overly artistically astute, in contrast to a sample of students for whom creativity is pivotal to their life goals and personal identities (Karwowski, 2011; Karwowski et al., 2013; Silvia & Nusbaum, 2012). Researchers also should examine ways to reduce inattentiveness effects. For example, offering incentives can increase participant compliance and engagement in research (e.g., Lindsay, Minard, Hudson, Green, & Schmitz, 2014; Patrick, Singer, Boyd, Cranford, & McCabe, 2013), and such methods may reduce careless responding.

References

- Acar, S., & Runco, M.A. (2012). Psychoticism and creativity: A meta-analytic review. *Psychology of Aesthetics, Creativity, and the Arts*, 6, 341– 350.
- Ashton, M.C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91, 340– 345.
- Ashton, M.C., Lee, K., & de Vries, R.E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A review of research and theory. *Personality and Social Psychology Review*, 18, 139– 152.
- Batey, M. (2007). A psychometric investigation of everyday creativity. Unpublished doctoral dissertation. University College, London.
- Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, 132, 355– 429.
- Berry, D.T.R., Wetter, M.W., Baer, R.A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, 4, 340– 345.
- Bollen, K.A., Harden, J.J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling*, 21, 1– 19.
- 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.
- Crowne, D.P., & Marlowe, D.A. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349– 354.
- Crowne, D.P., & Marlowe, D. (1964). *The approval motive*. New York: John Wiley & Sons.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- DeWall, C.N., Silvia, P.J., Schurtz, D.R., & McKenzie, J. (2011). Taste sensitivity and aesthetic preferences: Is taste only a metaphor? *Empirical Studies of the Arts*, 29, 171– 189.
- Gignac, G.E. (2013). Modeling the balanced inventory of desirable responding: Evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment*, 95, 645– 656.
- Goldman, B.M., & Kernis, M.H. (2004). The development of the authenticity inventory, version 3. Unpublished raw data. Athens, GA: University of Georgia.
- Gross, G.M., Silvia, P.J., Barrantes-Vidal, N., & Kwapil, T.R. (2012). Psychometric properties and validity of short forms of the Wisconsin Schizotypy Scales in two large samples. *Schizophrenia Research*, 134, 267– 272.

- Hocevar, D. (1979). 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).
- Huang, J.L., Bowling, N.A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*. DOI 10.1007/s10869-014-9357-6.
- Johnson, J.A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103– 129.
- Karwowski, M. (2011). It doesn't hurt to ask...But sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts*, 5, 154– 164.
- Karwowski, M., Lebuda, I., Wisniewska, E., & Gralewski, J. (2013). Big Five personality traits as the predictors of creative self-efficacy and creative personal identity: Does gender matter? *Journal of Creative Behavior*, 47, 215– 232.
- 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, 298– 308.
- Kaufman, S.B. (2013). Opening up openness to experience: A four-factor model and relations to creative achievement in the arts and sciences. *Journal of Creative Behavior*, 47, 233– 255.
- Kernis, M.H., & Goldman, B.M. (2006). A multicomponent conceptualization of authenticity: Theory and research. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 283– 357). San Diego, CA: Elsevier Academic Press.
- Kwapil, T.R., Barrantes-Vidal, N., & Silvia, P.J. (2008). The dimensional structure of the Wisconsin schizotypy scales: Factor identification and construct validity. *Schizophrenia Bulletin*, 34, 444– 457.
- Lindsay, J.A., Minard, C.G., Hudson, S., Green, C.E., & Schmitz, J.M. (2014). Using prize-based incentives to enhance daily interactive voice response (IVR) compliance: A feasibility study. *Journal of Substance Abuse Treatment*, 46, 74– 77.
- Makel, M.C. (2014). The empirical march: Making science better at self-correction. *Psychology of Aesthetics, Creativity, and the Arts*, 8, 2– 7.
- Maniaci, M.R., & Rogge, R.D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61– 83.
- Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437– 455.
- Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false positive rates. *Personality and Social Psychology Review*, 18, 107– 118.

- Musch, J., Ostapczuk, M., & Klaiber, Y. (2012). Validating an inventory for the assessment of egoistic bias and moralistic bias as two separable components of social desirability. *Journal of Personality Assessment*, 94, 620– 629.
- Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867– 872.
- Patrick, M.E., Singer, E., Boyd, C.J., Cranford, J.A., & McCabe, S. (2013). Incentives for college student participation in web-based substance use surveys. *Addictive Behaviors*, 38, 1710– 1714.
- Paulhus, D.L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598– 609.
- Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17– 59). San Diego, CA: Academic Press.
- Paulhus, D.L., & John, O.P. (1998). Egoistic and moralistic bias in self perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, 66, 1025– 1060.
- Paulhus, D.L., & Reid, D.B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60, 307– 317.
- Plucker, J.A., Runco, M.A., & Lim, W. (2006). Predicting ideational behavior from divergent thinking and discretionary time on task. *Creativity Research Journal*, 18, 55– 63.
- Richards, R. (2007). Everyday creativity: Our hidden potential. In R. Richards (Ed.), *Everyday creativity and new views of human nature: Psychological, social, and spiritual perspectives* (pp. 25– 53). Washington, DC: American Psychological Association.
- Richards, R. (2010). Everyday creativity: Process and way of life—four key issues. In J.C. Kaufman & R.J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 189– 215). New York: Cambridge University Press.
- Runco, M.A., Plucker, J.A., & Lim, W. (2000–2001). Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal*, 13, 393– 400.
- Runco, M.A., Walczyk, J.J., Acar, S., Cowger, E.L., Simundson, M., & Tripp, S. (2014). The incremental validity of a short form of the Ideational Behavior Scale and usefulness of distractor, contraindicative, and lie scales. *Journal of Creative Behavior*, 48, 185– 197.
- Silvia, P.J. (2007). Knowledge-based assessment of expertise in the arts: Exploring aesthetic fluency. *Psychology of Aesthetics, Creativity, and the Arts*, 1, 247– 249.
- Silvia, P.J., & Barona, C.M. (2009). Do people prefer curved objects? Angularity, expertise, and aesthetic preference. *Empirical Studies of the Arts*, 27, 25– 42.

- Silvia, P.J., Beaty, R.E., Nusbaum, E.C., Eddington, K.M., Levin-Aspenson, H., & Kwapil, T.R. (2014). Everyday creativity in daily life: An experience-sampling study of “little c” creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 8, 183– 188.
- Silvia, P.J., & Berg, C. (2011). Finding movies interesting: How expertise and appraisals influence the aesthetic experience of film. *Empirical Studies of the Arts*, 29, 73– 88.
- 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., & Nusbaum, E.C. (2011). On personality and piloerection: Individual differences in aesthetic chills and other unusual aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 5, 208– 214.
- Silvia, P.J., & Nusbaum, E.C. (2012). What's your major? College majors as markers of creativity. *International Journal of Creativity and Problem-Solving*, 22, 31– 43.
- 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– 16.
- Smith, L.F., & Smith, J.K. (2006). The nature and growth of aesthetic fluency. In P. Locher, C. Martindale & L. Dorfman (Eds.), *New directions in aesthetics, creativity, and the arts* (pp. 47– 58). Amityville, NY: Baywood.
- Steenkamp, J.B.E., De Jong, M.G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research*, 47, 199– 214.
- Stevens, J.P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, 95, 334– 344.
- Stevens, C.D., & Ash, R.A. (2001). The conscientiousness of students in subject pools: Implications for “laboratory” research. *Journal of Research in Personality*, 35, 91– 97.
- Tein, J.Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, 20, 640– 657.
- Tukey, J.W. (1977). *Exploratory data analysis*. New York: Pearson.
- Winterstein, B.P., Silvia, P.J., Kwapil, T.R., Kaufman, J.C., Reiter-Palmon, R., & Wigert, B. (2011). Brief assessment of schizotypy: Developing short forms of the Wisconsin Schizotypy Scales. *Personality and Individual Differences*, 51, 920– 924.
- Witt, E.A., Donnellan, M.B., & Orlando, M.J. (2011). Timing and selection effects within a psychology subject pool: Personality and sex matter. *Personality and Individual Differences*, 50, 355– 359.