Article

Social Psychological and Personality Science
1-8
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550618766409
journals.sagepub.com/home/spp
SAGE

# Deeply Confusing: Conflating Difficulty With Deep Revelation on Personality Assessment

Randy Stein[1] and Alexander B. Swan[2]

## Abstract

The factors that contribute to lay expectations of personality assessments are not well understood. Five studies demonstrate that people conflate difficulty of personality assessment items with revelations of deep insights. As a result, popular yet invalid assessments of personality can be seen as "deeper" than assessments from social and personality psychology. In Study 1, participants evaluated items from a popular personality "type" assessment as more difficult and better at revealing deep insights into personality than Big-Five personality inventory items. Studies 2 and 3 replicate this effect experimentally using a manipulation of assessment items' difficulty. Studies 4 and 5 show that the same effect also holds for a less direct method of supposed personality assessment (e.g., assessments that ask about which colors are associated with trivial concepts). Moderating factors and the popularity of shoddy personality assessments are discussed.

## Keywords

lay psychology, personality assessment, true self

Research in the past decade has spurred leaps in the understanding of how lay theories undermine scientific thinking within physical sciences (Gervais, 2015; Lewandowsky, Gignac, & Oberauer, 2013; Shtulman, 2015). People apparently also view their own psychology in unscientific terms. People often believe that social categories have underlying "essential" truths that cannot be directly observed (Bastian & Haslam, 2006; Gelman, 2003; Haslam, Bastian, & Bissett, 2004) and that they have "souls," separate from their bodies and hidden from physical measurements (Bloom, 2009). Despite being ill-defined, people also think their "true self" has an important causal role in judgments and decisions (Strohminger, Knobe, & Newman, 2017). Thus, pseudoscientific ideas about psychology might be intuitively appealing.

Here, we focus on that premise by exploring whether people think pseudoscientific personality assessments actually do a better job at generating insights into the "deep" self than more rigorous ones. Indeed, many people turn to popular personality tests, not academic journals, when they wish to learn about their own psychology (Paul, 2010). Perhaps unsurprisingly, these assessments often make bold claims about making revelations of the hidden self (e.g., Myers & Briggs Foundation, 2016).

However, if people are convinced that a hidden self defines them, why are they also convinced that these pseudoscientific assessments measure it? Our hypothesis is that, when taking personality tests, people conflate judgments of the questions' difficulty with judgments of the questions' ability to tap their deep inner selves. As a result, questions like those from popular "type" assessment tools (and, as we explore in Studies 4 and 5, Buzzfeed-like personality quizzes) seem especially difficult and, correspondingly, revealing.

To explicitly define our terms, *difficulty* (our independent variable) refers to the subjective feeling that an item involves a hard-to-resolve judgment about the self such as a choice between nonopposites or a question that is conceptually ambiguous. Difficulty, as used here, does not refer to "item difficulty" in the item response theory sense (an item's ability to separate low scorers from high scorers).

*Depth* (our dependent variable) refers to the subjective feeling that an item gets at deep down, subconscious elements of the self. Since personality assessments are often marketed as paths to revelations of a hidden yet critically important part of the self, perceptions of an item's depth, while likely a form of naive essentialism (Gelman, 2003), can also be taken as perceptions that the item has value.

Consider, for example, the Keirsey Temperament Sorter (KTS; Keirsey, 1998; Keirsey & Bates, 1984). The KTS is an assessment easily found when searching online for the Myers–Briggs Type Indicator (MBTI), which is both the most popular

[1] College of Business Administration, California State Polytechnic University, Pomona, Pomona, CA, USA
[2] Social Science and Business Division, Eureka College, Eureka, IL, USA

**Corresponding Author:**
Randy Stein, College of Business Administration, California State Polytechnic University, Pomona, 3801 W Temple Ave., Pomona, CA 91768, USA.
Email: rbstein@cpp.edu

personality assessment and highly scientifically suspect (Stromberg & Caswell, 2015). Like the official MBTI assessment (Briggs Myers, McCaulley, Quenk, & Hammer, 1998), KTS questions are sometimes vague and sometimes require choices between items that often are not actually opposites (e.g., a question like "Are you more likely to: a. see how others are useful; b. see how others see" exhibits both characteristics).

Presumably, these features make the questions difficult, compared to more straightforward items like those from the Big-Five Personality Inventory (e.g., Goldberg et al., 2006). We hypothesize here that assessment takers might perceive such difficult questions as engaging deeper, less easily observed aspects of their personality. As the average person is likely undersensitive to validity issues, this perception of depth thus might hold even for less valid assessments.

In exploring this hypothesis, we contribute to personality psychology by presenting a novel view into what features can make personality assessments intuitively appealing. We also add to the true self literature (e.g., Strohminger et al., 2017) by exploring when the "deeper" self is functionally invoked (i.e., to resolve difficult self-assessments).

### The Current Studies

Each of the studies involves having participants evaluate personality assessment items on our constructs of interest. Study 1 emphasizes external validity, showing that KTS items are evaluated (on average) as more difficult and deeper than items from the Big Five. Studies 2–5 show that experimental manipulations of difficulty led to a corresponding increase in perceived depth.

In addition to our main dependent measure, we had participants rate the assessment items' *relevance* to personality, by which we mean the perception that the item has a clear, detectable relation to some aspect of personality that is discriminating and meaningful. Depth and relevance are likely linked. However, to illustrate the difference, consider the Big-Five item that asks whether one "tends to follow a schedule." Such an item has a clear relation to an important, discriminating personality trait (conscientiousness), but answering the item would likely not result in the feeling of reaching deep into the subconscious. Conversely, the Rorschach test supposedly generates insights into the subconscious without explicitly targeting personality traits at all, so it may or may not be seen as relevant. We thus measured relevance to obtain a richer view of how participants viewed the personality items and to explore whether the effects of difficulty are unique to depth.

## Study 1

## Method

### Participants

One hundred two participants (54 males, 46 females, and 2 unknown; $M_{\text{age}} = 33.68$, $SD = 9.84$) completed this study through Amazon's Mechanical Turk (mTurk). There were no a priori expectations about effect size, so we chose 100–110 participants, a relatively well-powered sample for a within-participants design (a sample of 100 has 80% power, one-tailed, for effect size $d = .25$). We found that the mTurk data collection rate can slow considerably as it nears the end. So, we set target ranges that would represent a well-powered sample, and, if it was convenient to do so, stopped data collection once we were in that range. Data collection was stopped without first examining data.

### Procedure

Participants were told we were interested in their opinions on personality assessments. All participants were shown and then evaluated 5 items randomly chosen from the 70-item KTS (Keirsey, 1998; Keirsey & Bates, 1984) and 5 items randomly chosen from the Goldberg et al.'s (2006) 50-item International Personality Item Pool (IPIP) Big-Five assessment (answered on 5-point agreement scales). A different set of 5 items was randomly chosen from each assessment for each participant. The order of the assessments was counterbalanced (for a within-participants design), so half the participants first evaluated the Big-Five items and then evaluated the KTS items, and the other half of the participants did the reverse.

Participants were asked to answer the personality assessment items carefully and honestly as if they were taking the entire personality assessment and told that afterward we would gather opinions of those assessments. Participants were also told we were showing them only part of each assessment. The full set of items from both assessments is in Online Appendix A.

After seeing each set of items, they were shown each item again on a new screen and evaluated each individual item on measures of difficulty, relevance, and depth. Participants rated their disagreement or agreement (on a scale from 1 to 5) with eight statements.

The "difficulty" questions are as follows:

1. I had to think hard about how to answer this question.
2. This was a confusing question.
3. I contemplated giving a different to this question than the answer I actually gave.

The "relevance" questions are as follows:

4. My answer to this question is part of what makes my personality unique.
5. This question targets a meaningful part of my personality.

The "depth" questions are as follows:

6. This question is deep.
7. My subconscious mind was involved in answering this question.
8. I felt like this question was getting at a hidden aspect of my personality.

**Table 1.** Summary of Results for Studies 1 Through 5.

| | Study 1 (Within-Participants Personality Rating) | | | | Study 2 (Between-Participants Personality Rating) | | | | Study 3 (Between-Participants Personality Rating) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Condition | | Difference | | Condition | | Difference | | Condition | | Difference | |
| Average Rating | KTS | Big Five | Hedges' *g* | *p* | Difficult Big Five | Big Five | Hedges' *g* | *p* | KTS | Easy KTS | Hedges' *g* | *p* |
| Difficulty | 1.97 (.64) | 1.50 (.52) | .90 | <.001 | 2.20 (.82) | 1.67 (.68) | .64 | <.001 | 2.43 (.88) | 1.99 (.73) | .54 | <.001 |
| Depth | 2.36 (.78) | 2.15 (.80) | .26 | <.001 | 2.42 (.84) | 2.17 (.82) | .30 | .005 | 2.73 (.81) | 2.51 (.80) | .27 | .02 |
| Relevance | 2.91 (.89) | 2.84 (.91) | .08 | .32 | 2.80 (.89) | 2.71 (.82) | .10 | .37 | 3.01 (.85) | 3.18 (.78) | .21 | .12 |
| | Study 4 (Between-Participants Color Rating) | | | | Study 5 (Between-Participants Color Rating) | | | | | | | |
| | Condition | | Condition | | Condition | | Condition | | | | | |
| | Difficult | Easy | Hedges' *g* | *p* | Difficult | Easy | Hedges' *g* | *p* | | | | |
| Difficulty | 3.09 (1.18) | 2.16 (.92) | .88 | <.001 | 2.82 (1.03) | 1.55 (.64) | .88 | <.001 | | | | |
| Depth | 2.78 (1.08) | 1.97 (.90) | .82 | <.001 | 2.44 (1.00) | 1.88 (.86) | .60 | <.001 | | | | |
| Relevance | 2.54 (1.16) | 1.78 (.90) | .72 | <.001 | 2.14 (1.07) | 2.02 (.92) | .12 | .45 | | | | |

*Note.* Assessments of difficulty, depth, and relevance to personality were made on 5-point scales. Means are represented in condition cells with *SD* in parentheses. In line with our preregistrations, *p* values for the difficulty and depth scales for Studies 2–5 are reported as one-tailed; all other *p* values are two-tailed. KTS = Keirsey Temperament Sorter.

Respondents rated each assessment item individually. Respondents rated only 5 items per test to keep the length of the survey reasonable.

After this section, participants rated 5 summary judgment items (on the 1–5 agreement scale):

1. I'm interested in taking the rest of this assessment.
2. This assessment seems like it would be valid.
3. This assessment seems like it would be useful in helping me know myself.
4. This assessment can measure how people "really" are.

The full set of questions was asked for the first assessment before moving on to the next assessment. For exploratory purposes, in all studies in this article, respondents were next asked to complete the Bullshit Receptivity Scale (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015) and the 10-item Rational-Experiential Inventory (REI-10; Epstein, Pacini, Denes-Raj, & Heier, 1996). Results related to these scales for all studies are in Online Appendix E. Following a series of demographics questions, participants were thanked and compensated.

## Results

Average difficultly, depth, and relevance for each personality assessment item were first computed using the individual eight evaluation questions. These resulting item scores were averaged to create composites for both the KTS and the Big Five (all three $\alpha > .76$). The full set of *t* tests and correlations for the eight individual questions for all studies are in the Online Supplemental Spreadsheet, tables 4 and 5, respectively. The patterns of the individual questions are largely identical to the analyses of the three indices. Scores on these three metrics for each individual personality assessment item are displayed in the Online Supplemental Spreadsheet, table 1. Following Lakens (2013), we report effect sizes as Hedges' $g_{\text{av}}$ for within-participants comparisons and Hedges' $g_{\text{s}}$ for between-participants comparisons. Both are bias-corrected versions of Cohen's *d*.

### Main Hypotheses

Results (summarized in Table 1) supported the hypothesis of the study that people would conflate difficulty of their decisions on the items with how effective the items read the deeper self. For both personality assessments, perceived difficulty was positively correlated to perceived depth (see Table 2 for correlation matrices for each study).

Also, comparing evaluations of the two assessments, the KTS items were rated as both more difficult, $t(101) = 8.69$, $p < .001$; Hedges' $g_{\text{av}} = .80$; mean difference 95% confidence interval (CI) = [0.36, 0.58], and deeper, $t(101) = 4.66$, $p < .001$; Hedges' $g_{\text{av}} = .26$; mean difference 95% CI [0.12, 0.29], than the Big-Five items.

### Relevance

As shown in Table 2, relevance and depth were highly correlated for both assessments. However, relevance and difficulty were uncorrelated for both assessments. Additionally, the two assessments were not seen as significantly different on the relevance measure, $t(101) = 1.00$, $p = .32$. Thus, relevance and depth differ when it comes to their relationship to difficulty,

**Table 2.** Correlations for Studies 1 Through 5.

| | Study 1—KTS | | | Study 1—Big Five | | | Study 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| Average Rating | Difficulty | Relevance | Depth | Difficulty | Relevance | Depth | Difficulty | Relevance | Depth |
| Difficulty | 1 | 0.04 | 0.34* | 1 | 0.08 | 0.33* | 1 | 0.23* | 0.47* |
| Direct relevance | 0.04 | 1 | 0.56* | 0.08 | 1 | 0.65* | 0.23* | 1 | 0.69* |
| Depth | 0.34* | 0.56* | 1 | 0.33* | 0.65* | 1 | 0.47* | 0.69* | 1 |
| | Study 3 | | | Study 4 | | | Study 5 | | |
| | Difficulty | Relevance | Depth | Difficulty | Relevance | Depth | Difficulty | Relevance | Depth |
| Difficulty | 1 | 0.06 | 0.41* | 1 | 0.35* | 0.39* | 1 | 0.27* | 0.42* |
| Direct relevance | 0.06 | 1 | 0.59* | 0.35* | 1 | 0.82* | 0.27* | 1 | 0.82* |
| Depth | 0.41* | 0.59* | 1 | 0.39* | 0.82* | 1 | 0.42* | 0.82* | 1 |

*Note.* For Studies 2–5, correlations within each condition are in the Online Supplemental Spreadsheet, table 6. Correlations for Study 1 are presented here by condition since each condition has a full representation of its source test (as opposed to a smaller set of items intended to be relatively easy or difficult). KTS = Keirsey Temperament Sorter.

*$p < .01$. Other correlations are not significant at the two-tailed $p < .05$ level.

as perceived shifts in difficulty were associated with corresponding shifts in depth, but not relevance.

### *Additional Results*

Additional analyses were conducted on the summary questions, and the KTS items were rated as slightly more valid than the Big-Five items, KTS: $M = 3.75$, $SD = 1.08$; Big-Five: $M = 3.55$, $SD = 1.11$; $t(101) = 2.27$, $p = .03$; Hedges' $g_{\text{av}} = .19$; mean difference 95% CI [0.03, 0.39]. KTS items were also rated as slightly more helpful for a person to better know themselves than Big-Five items, KTS: $M = 3.48$, $SD = 1.24$; Big Five: $M = 3.28$, $SD = 1.29$; $t(101) = 2.17$, $p = .03$; Hedges' $g_{\text{av}} = .16$; mean difference 95% CI [0.02, 0.37]. Respondents were equally interested in taking the two assessments, KTS: $M = 3.48$, $SD = 1.33$; Big Five: $M = 3.32$, $SD = 1.39$, $p = .13$, and saw the two as equally good at revealing how people "really" are (KTS: $M = 3.24$, $SD = 1.18$; Big Five: $M = 3.12$, $SD = 1.19$, $p = .15$). Taken together with the individual question ratings, these results suggest that both tests were perceived positively, though the KTS was perceived as specifically better at reading the deep self, more than the Big Five. Since the focus of our research was the depth ratings, and these summary questions are likely impacted by many factors besides perceived depth, we did not use these summary questions in future studies.

Studies 2 through 5 focus on demonstrating that experimental manipulations of a personality assessment's difficulty lead to increased perceptions of depth. Note that experimentally manipulating difficulty necessitates a change in some potentially confounding feature of the assessment, namely question content or format. Therefore, while we cannot entirely eliminate concerns about confounds in any individual study, to examine the robustness of the effect, we control for content (Study 2, Study 4) and question type (Study 3, Study 5) and use both explicit personality assessment (Study 2, Study 3) and indirect assessment (Study 4, Study 5).

# Study 2

The observations in Study 1 revealed that difficulty of items on a personality assessment was conflated with depth. We should, then, be able to obtain a similar effect by manipulating "easy," less deep Big-Five items and experimentally making them "harder" while keeping content fairly consistent. In Study 2, we used a two-cell (easy vs. difficult) between-participants design.

For the easy condition, respondents saw five Big-Five items (which might have been altered, as explained below). For the difficult condition, we created versions of several Big-Five items that had difficult-to-resolve false dichotomies, akin to many KTS items. These difficult items we hypothesized would be rated as deeper.

## Method

### *Participants*

Two hundred and eighty-seven participants (131 males, 156 females; $M_{\text{age}} = 35.38$) completed this preregistered study (https://osf.io/gfaje/?view_only=c40ee46a41cd4e748c9505bdd7afbe00) through mTurk.

Sample sizes for Studies 2–5 were determined by expectations of the effect size of our most critical comparison, namely, the difference between the two conditions on the depth measure. Previously collected data indicated that we should expect an effect size of approximately Hedges' $g_s = .30$. A corresponding power analysis indicated we would need 139 participants per condition to achieve 80% power, so we aimed for a total of 275–300 participants. For Studies 2 and 3, we examined data after collecting 50 responses to ensure that the difficulty manipulation was working as intended but otherwise did not examine data prior to stopping and did not add data after reaching our target.

### Procedure

Study 2 was similar to Study 1. However, the design was between-participants rather than within. Participants answered and then evaluated 5 items from only one personality inventory, either the regular Big Five (easy condition) or the difficult version we created. For the difficult version, we created 9 items that required respondents to choose between options items that are conceptually similar, but not logical opposites. Thus, each of the nine difficult questions was based on 2 Big-Five items. The easy condition contained only the 18 items used to construct the difficult items. Thus, question content is controlled for between conditions.

Guided by our intuitions and the individual item ratings from Study 1, we selected IPIP Big-Five items that (1) did not already seem to be asking about a deep concept (e.g., an item like "I have a soft heart" already gets at what people are like "deep on the inside" so trying to increase its depth seemed fruitless) and (2) would make linguistic sense when paired with another item from its own dimension in a forced-choice question, while still resulting in a choice that was conceptually vague and/or a false dichotomy.

We drew on the full 100-item IPIP Big-Five set (Goldberg et al., 2006) to have suitable options. Also, since KTS questions sometimes ask about actual behavior tendencies (as most Big-Five items do) and sometimes ask about *preferences* for certain values or personality traits, we allowed our difficult questions to also ask about tendency or preference, based on what best fit the question. When we asked about preference in the difficult condition, we changed the wording in the corresponding original Big-Five item to also refer to preference. This does alter the meaning of the original item but keeps wording consistent between conditions. The full set of items is in Online Appendix B.

After initially seeing each item, as in Study 1, participants were shown the items again and asked to rate each on perceived difficulty, relevance, and depth.

## Results

Results are summarized in Table 1. Reliabilities for the depth, relevance, and difficultly scales were all high ($\alpha > .85$). Scores on these three metrics for each individual personality assessment item are displayed in the Online Supplemental Spreadsheet, table 2. Also, full correlation matrices for these three indices split by between-participants condition for Studies 2–5 are in the Online Supplemental Spreadsheet, table 6.

For Studies 2–5, reported $p$ values for the manipulation check and main hypothesis are one-tailed, as specified in our preregistrations. All effects in these analyses would be significant at the $p < .05$ level even using two-tailed $p$ values. We use two-tailed $p$ values when reporting results for the relevance measures.

### Manipulation Check

As expected, items from the difficult condition were rated as more difficult than those from the easy condition, $t(285) = 6.01$, $p < .001$; Hedges' $g_s = .64$; mean difference 95% CI [0.34, 0.71].

### Main Hypotheses

As predicted, items from the difficult condition were seen as deeper than the easy condition, $t(285) = 2.57$, $p = .005$; Hedges' $g_s = .30$; mean difference 95% CI:[0.06, 0.45]. We also observed a significant correlation between perceived difficulty and depth.

### Relevance

As shown in Table 2, relevance and depth were significantly correlated, as were relevance and difficulty. However, there was *not* a difference in relevance ratings between conditions, $t(284) = .91$, $p > .30$. Shifts in difficulty were associated with shifts in depth, but not shifts in relevance.

In Study 3, we replicate the difficulty/depth effect when holding item type constant, comparing easy forced choice items (KTS items modified to be easier) with more difficult ones (original KTS items), thus demonstrating our effects are not unique to comparing forced choices to single items.

# Study 3

Study 3 was similar to Study 2. However, here, rather than modifying Big-Five items, we compared original KTS questions to "easier" versions we created.

## Method

### Participants

Two hundred and twenty-six participants (97 males, 129 females; $M_{\text{age}} = 38.76$) completed this preregistered study (https://osf.io/etx8w/?view_only=c565d9722bea437d8a7a8be86b325cf0) through mTurk. Based on previous studies we ran, we expected an effect size of approximately Hedges' $g_s = .35$. To detect an effect that size at 80% power, we needed 102 respondents per cell. More conservatively, we set a goal of running about 200–220 respondents (though due to an error in posting we allowed for a maximum of 234).

### Procedure

The procedure of Study 3 was the same as Study 2. However, in the difficult condition, those 5 items were randomly chosen from a set of 8 original KTS items. For the easy condition, we created a set of 8 modified items that were intended to be easier" than their original counterparts. Using the item ratings from Study 1 as a guide, we identified 8 KTS items that were rated as relatively deep and that we thought we could make easier by reducing ambiguity of language or from changing a false dichotomy to a more real dichotomy, while still asking about a similar concept. The full set of items in each condition can be found in Online Appendix C. After initially seeing each item, participants were shown the items again and asked to rate each on difficulty, relevance, and depth as in Studies 1 and 2.

## Results

Results are summarized in Table 1. Reliabilities for the depth, relevance, and difficultly scales were all high ($\alpha >$ .84). Scores on these three metrics for each individual personality assessment item are in the Online Supplemental Spreadsheet, table 3.

### Manipulation Check

Items from the difficult condition were in fact perceived as more difficult as those from the easy condition, $t(224) = 4.09$, $p < .001$; Hedges' $g_s = .54$; mean difference 95% CI [0.23, 0.65].

### Main Hypotheses

As hypothesized, items from the difficult condition were also seen as deeper than the easy condition, $t(224) = 2.01$, $p = .02$; Hedges' $g_s = .27$; mean difference 95% CI [0.01, 0.42]. Perceived difficulty and depth were also significantly correlated (see Table 2).

### Relevance

Relevance and depth were significantly correlated (see Table 2). However, relevance and difficulty were uncorrelated and there was no difference between conditions on relevance, $t(224) = -1.54, p = .12$. Thus, despite a high correlation, shifts in difficulty are associated with shifts in depth, but not relevance.

Note that while the difficulty and depth link observed in Studies 2 and 3 may play some role in the appeal of pseudoscientific assessments that tend to have especially confusing questions, these studies do not necessarily reflect entirely unsound judgments per se. The items in the difficult conditions are not necessarily invalid in an absolute sense. These items are likely relatively invalid in terms of measuring their target traits. However, some could assess relative valuation of traits, and as such they do introduce a greater level of conceptual complexity (and, perhaps, depth) than the easier questions.

In Studies 4 and 5, we thus broaden the scope of the effects, showing that the same pattern of results replicates when using a more plainly invalid method of personality assessment.

## Study 4

Online "personality quizzes" often require people to make abstract judgments about associations, such as one that first asks people about disparate product preferences (e.g., which of nine different ways do you take your coffee?) and then outputs insights about the self ("which city should you actually live in?"; Perez, 2014). We used a methodology inspired by an actual Buzzfeed quiz (Hickey, 2017), in which quiz takers note which of several colors they associate with a set of concepts (such as days of the week, letters, and numbers) and are then given "the true age of [their] soul."

We showed participants purported personality assessment items that required them to judge the extent to which they associated colors with certain concepts (like "Tuesday" and the letter "M"). Respondents decided which of two colors (blue and green) they associated with each concept or they rated colors individually. After initially seeing the items, participants rated them on similar scales to those in Studies 1 through 3.

## Method

### Participants

One hundred and forty-six participants (72 males, 74 females, $M_{\text{age}} = 35.04$) completed this preregistered study (https://osf.io/fnqpx/?view_only=6f1a781d6fd242978a73615dae2a7785) through mTurk. Based on previously collected data, we expected a critical effect size of Hedges' $g_s = .47$, and a power analysis indicated needing at least 56 respondents per cell for 80% power. We aimed for a more conservative 140–150 total respondents. For Studies 4 and 5, data collection was stopped without prior examination of data.

### Procedure

The procedure was similar to Studies 2 and 3. For personality assessment items, respondents judged associations with colors. Stimuli for this task are shown in Online Appendix D. In the difficult condition, respondents were asked five questions of the form "which color do you associate with the letter 'M'?" and given a choice between blue and green. The other four concepts included "Tuesday," "11," "S," and "8." The choice was always between blue and green.

In the easy condition, respondents were asked to note on a 1–5 scale the extent to which they associate one of the two colors (chosen at random) with the concept, for example, "To what extent do you associate the color green with the letter 'M'?"

Participants were then shown each of the assessment items again and rated the items using questions similar to those of Studies 1 and 2. We slightly modified three of the questions to reduce ambiguity in this new context that did not explicitly address personality. We changed "this question targets a meaningful part of my personality" to "this question targets a meaningful personality trait," "this question is 'deep'" to "this question gets at something deep," and "I felt like this question was getting at a hidden part of my personality" to "I felt like this question was getting at a hidden part of me."

## Results

Results are summarized in Table 1. The perceived difficulty, relevance to color, and perceived depth subscales were highly reliable ($\alpha > .82$).

### Manipulation Check

Items from the difficult condition were rated as more difficult as those from the easy condition, $t(144) = 5.38$, $p < .001$; Hedges' $g_s = .88$; mean difference 95% CI [0.59, 1.28].

### Main Hypotheses

Most critically, items from the difficult condition were seen as deeper than the easy condition, $t(144) = 4.94$, $p < .001$; Hedges' $g_s = .82$; mean difference 95% CI [0.48, 1.13]. Additionally, in support of the main hypothesis, a significant correlation was observed between perceived difficulty and depth.

### Relevance

As with Studies 1–3, perceived depth was correlated with relevance, though here the correlation was notably larger (see Table 2). Additionally, relevance was correlated with perceived difficulty, and the difficult condition questions were seen as more relevant than the easy condition questions, $t(144) = 4.41$, $p < .001$; Hedges' $g_s = .73$; mean difference 95% CI [0.421, 1.10]. In this study, since all the assessment items were quite similar and have no explicit link to personality, it is reasonable that the major input into relevance judgments was perceived depth (e.g., there was no other reason for items to be perceived as relevant), hence the stronger relationship between the two (compared to Studies 1–3).

# Study 5

Study 5 provides a conceptual replication of Study 4 wherein both the easy and difficult conditions contain a forced choice.

## Method

### Participants

One hundred and fifty-three participants (91 males, 62 females, $M_{\text{age}} = 36.20$) completed this preregistered study (https://osf.io/x4ngy/?view_only=2f20ba8219fb4a0aa6eb232d69b98f4d) through mTurk. We anticipated a similar effect size as Study 4, so we obtained a similar sample size.

### Procedure

The difficult condition was identical to the difficult condition in Study 4. In the easy condition, rather than being asked "association" questions about five vague concepts, respondents were asked five presumably easier "preference" questions asking them which color (between blue and green) they preferred for a number of products (cars, walls, clothing, M&M's, and phone cases). All dependent measures were the same as in Study 4.

## Results

Results are summarized in Table 1. The perceived difficulty, relevance to color, and perceived depth subscales were all highly reliable ($\alpha > .82$).

### Manipulation Check

The association items from the difficult condition were seen as more difficult than the preference items in the easy condition, $t(151) = 9.07$, $p < .001$; Hedges' $g_s = .88$; mean difference 95% CI [0.99, 1.54].

### Main Hypotheses

As hypothesized, items from the difficult condition were seen as deeper than the easy condition, $t(151) = 3.75$, $p < .001$; Hedges' $g_s = .60$; mean difference 95% CI [0.27, 0.86]. Additionally, a significant correlation was observed between perceived difficulty and depth.

### Relevance

Perceived depth and relevance were again highly correlated (see Table 2). Correspondingly, relevance was also correlated with perceived difficulty. However, looking across conditions, the difficult condition questions were not seen as any more relevant than the easy condition questions, $t(151) = .783$, $p = .45$; Hedges' $g_s = .12$. Thus, our general pattern of findings (with the exception of Study 4) suggests that though depth and relevance are quite related, perceived difficulty has bearing on the former but not the latter.

## General Discussion

Five studies consistently showed that people conflate perceptions of a personality assessment's difficulty with its perceived depth. The breadth of studies highlight that the effect is not unique to any particular assessment or question type (also underscored by Online Supplemental Spreadsheet, table 6; we generally observe correlations between difficulty and depth within each between-participants condition). These studies contribute to knowledge of how people reason about when their "true selves" are invoked (Chen, Urminsky, & Bartels, 2016; Shlegel, Hicks, Arndt, & King, 2009; Shlegel, Hicks, King, & Arndt, 2011). Since popular personality tests often purport to "uncover" insights about "true" personality, these studies also suggest that features of these assessments that could make the tests less valid trait measures ironically might make them seem more able to generate such "hidden" insights.

A natural follow-up question is what might moderate our effects. Appendix E of the Supplemental Online Material is illuminating on this point, as we found some exploratory evidence across studies that the link between difficulty and depth is pronounced among people who are primarily trusting of intuitions (as measured by the REI-10 scale). Thus, at least part of the effects we observed might be due to a failure to correct for intuitions that difficult questions are especially deep.

A limitation of the current studies is that we did not examine the downstream consequences of the in-the-moment perceived link between difficulty and depth (e.g., the effects on motivation to taking and sharing the results of personality assessments). However, thinking of what these consequences might be, the (ironic) feeling of the deep self being invoked while taking personality assessments might make pseudoscientific tests impervious to criticism in the minds of some users, since the tests seem so "useful." Indeed, the specter of pseudoscientific personality assessments that supposedly get at the hidden self loomed large over even the academic literature for decades (Lilienfeld, Wood, & Garb, 2000). Thus, personality assessment might belong on the list of topics for which pseudoscience is more convincing to many people than science. How to increase the mind share of more rigorous ideas about psychology remains a nontrivial question.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Supplemental Material

The supplemental material is available in the online version of the article.

## References

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*, 228–235.

Bloom, P. (2009). *Descartes' baby: How the science of child development explains what makes us human*. New York, NY: Basic Books.

Briggs Myers, I., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *Manual: A guide to the development and use of the Myers-Briggs type indicator*. Palo Alto, CA: Consulting Psychologist Press.

Chen, S. Y., Urminsky, O., & Bartels, D. M. (2016). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science*, *27*, 1398–1406.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford, England: Oxford University Press.

Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition*, *142*, 312–321.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84–96.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, *30*, 1661–1673.

Hickey, A. (2017). This synesthesia test will reveal the true age of your soul. *Buzzfeed.com*. Retrieved from https://www.buzzfeed.com/agh/this-color-association-test-will-reveal-the-age-you-are-at-h?utm_term=.bxq7jW5mV#.tq6bPKYp2

Keirsey, D. (1998). *Please understand me II: Temperament, character, intelligence*. Del Mar, CA: Prometheus Nemesis Book.

Keirsey, D., & Bates, M. (1984). *Please understand me: Character and temperament types*. Del Mar, CA: Prometheus Nemesis Book.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. doi:10.3389/fpsyg.2013.00863

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The role of conspiracist ideation and worldviews in predicting rejection of science. *PLoS One*, *8*, e75637.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

Myers & Briggs Foundation. (2016). *MBTI basics*. Retrieved from http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/

Paul, A. M. (2010). *The cult of personality testing: How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstand ourselves*. New York, NY: Simon & Schuster.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*, 549.

Perez, A. (2014). What city should you actually live in? *Buzzfeed.com*. Retrieved from https://www.buzzfeed.com/ashleyperez/what-city-should-you-actually-live-in?utm_term=.fxXm68LN0#.ggbXmyoVN

Schlegel, R. J., Hicks, J. A., Arndt, J., & King, L. A. (2009). Thine own self: True self-concept accessibility and meaning in life. *Journal of Personality and Social Psychology*, *96*, 473.

Schlegel, R. J., Hicks, J. A., King, L. A., & Arndt, J. (2011). Feeling like you know who you are: Perceived true self-knowledge and meaning in life. *Personality & Social Psychology Bulletin*, *37*, 745–756.

Shtulman, A. (2015). How lay cognition constrains scientific cognition. *Philosophy Compass*, *10*, 785–798.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*, 551–560.

Stromberg, J., & Caswell, E. (2015). *Why the Myers-Briggs test is totally meaningless*. Retrieved from http://www.vox.com/2014/7/15/5881947/myers-briggs-personality-test-meaningless

## Author Biographies

**Randy Stein** is an assistant professor at the College of Business Administration, California State Polytechnic University, Pomona.

**Alexander B. Swan** is an assistant professor of psychology at Eureka College.

Handling Editor: Simine Vazire