Methodology and Research Practice

Malleability of Statistical Perception: Impact of Validity Presentation on College Admission Test Policy Preferences

Don C. Zhang1,2, Jonathan Wai2

1 Department of Psychology, Louisiana State University, Baton Rouge, LA, US, 2 Department of Education Reform, University of Arkansas, Fayetteville, AR, US

Keywords: standardized tests, effect size, statistical communication, cognitive ability tests, educational policy

https://doi.org/10.1525/collabra.24087

Collabra: Psychology
Vol. 7, Issue 1, 2021

Research evidence in the social sciences often relies on effect size statistics, which are hard to understand for the public and do not always provide clear information for decision-makers. One area where interpretation of research evidence has profound effects on policy is college admission testing. In this paper, we conducted two experiments testing how different effect size presentations affect validity perception and policy preferences toward standardized admission tests (e.g., ACT, SAT). We found that compared to traditional effect size statistics (e.g., correlation coefficient), participants perceived admission tests to be more predictively valid when the same evidence was presented using an alternative effect size presentation. The perceived validity of the admission test was also positively associated with admission test policies (e.g., test-optional policy) preferences. Our findings show that policy preferences toward admission tests depend on the perception of statistical evidence, which is malleable and depends on how evidence is presented.

Social policies are most effective when they are enacted on the basis of scientific evidence. Research evidence in the social sciences is commonly presented as the strength of an association between two variables (e.g., wealth and education attainment). This relationship is typically denoted as an effect size (J. Cohen, 1992). Unfortunately, traditional effect size statistics reported in scientific writing (e.g., the Pearson’s correlation coefficient) are hard to understand and difficult to interpret for policymakers, practitioners, and the public (Brooks et al., 2014; Hanel & Mehler, 2019; May, 2004), which may undermine the utility of research evidence in policy decisions. Furthermore, interpretation of effect size evidence is malleable, as illustrated by numerous revisions to what constitutes as a “small” versus “large” effect size – a discourse still not yet resolved amongst psychologists today (Bosco et al., 2015; J. Cohen, 1992; Funder & Ozer, 2019; Gignac & Szodorai, 2016).

Semantics aside, interpretation of effect size evidence can have profound effects on policy and practice. This is particularly evident in the context of policies related to whether standardized test scores should be used for college admissions decisions. In the past decade, a growing number of U.S. universities have opted to make admission tests optional or have removed them altogether (Hubler, 2020). The removal of standardized admission tests has accelerated since the COVID19 pandemic (Hoover, 2021). A common criticism leveraged against standardized tests is their inability to meaningfully predict future performance (Kuncel & Hezlett, 2010). When the new SAT was released in 2008, for example, Robert Schaeffer of Fairtest commented that the SAT “does not predict college success as well as high school grades, so why do we need the SAT, old or new?” (Lewin, 2008). These sentiments, however, contrast sharply with a large body of empirical evidence demonstrating the validity of testing in occupational and educational selection (e.g., Kuncel & Hezlett, 2007; Schmidt & Hunter, 1998). But what if the validity evidence of admission tests was presented using metrics that are easier to understand? Would it change how scores are interpreted and used for policy preferences?

In this paper, we conducted two randomized experiments that examined the impact of effect size presentation on the perceived validity of standardized admission tests and subsequent policy preferences toward their use for college admission decisions. In the first crowdsourced online experiment, we presented participants with the validity evidence of the ACT using either a traditional effect size metric (e.g., correlation coefficient) or an alternative effect size metric (e.g., expectancy chart). Drawing from social psychological research on motivated reasoning, we also examined the moderating role of SES on the association between perceived test validity and test-optional policy preferences. In the second pre-registered field study, we replicated the first experiment with a community sample on a university...
campus. We also obtained a behavioral measure of policy preferences by asking participants if they were willing to physically sign a petition that would render admission tests optional on campus. Together, this research illustrates how different presentations of the same evidence can produce disparate interpretations that could affect attitudes and preferences toward the use of tests in admission policies.

Public debate about standardized admission tests

Over a century of psychological research has established the remarkable robustness of general mental ability (GMA) as an individual differences disposition (Deary, 2000; Hunt, 2010; Lubinski, 2004). Stemming from this work is the development of cognitive ability tests, which are used to accurately assess, identify, and select individuals in work and educational settings (Schmidt & Hunter, 1998; Terman & Merrill, 1937). Since its inception, cognitive ability testing has been the center of considerable public and academic controversy (Cronbach, 1975; Hutt & Schneider, 2018; Murphy et al., 2005). Past president of the American Psychological Association, Barbara Lerner, commented during the 1978 annual convention, "tests are under attack today because they tell us truths about ourselves and our society." History is repeating itself in higher education today, where many students and faculty of U.S. universities have become increasingly wary of using standardized tests for undergraduate and graduate admissions: a sentiment that has been amplified by the COVID19 pandemic (Hoover, 2020).

Against this backdrop, the University of California (UC) system launched a two-year investigation into the efficacy of standardized tests but ultimately found that they were valid predictors of academic performance and recommended against test removal (Report of the UC Academic Council Standardized Testing Task Force, STTF, 2020). Despite the weight of the evidence, the UC system decided to abandon existing standardized tests (i.e., ACT and SAT) for higher education admissions, with the purported intent to create an in-house, UC developed test (Huhler, 2020). According to a report by FairTest, a national education reform organization that is typically anti-testing (Phelps, 2005), over one thousand American colleges and universities have now either removed standardized testing or made it optional for undergraduate admissions.

Criticisms notwithstanding, standardized selection tests such as college admission exams largely reflect individual differences in developed general mental ability (ACT: Koenig et al., 2008; SAT: Frey & Detterman, 2004) and are one of the best predictors of achievements at school, work, and life (Brown et al., 2021; Kuncel et al., 2004; Rohde & Thompson, 2007). In undergraduate admissions, the predictive validity of standardized tests (i.e., ACT, SAT), which denotes the association between test scores and academic achievement, ranges from 0.51 to 0.67 for predicting cumulative GPA and individual course grades. Similarly, in graduate admissions, the predictive validity of the GRE ranges from 0.34 to 0.41 for cumulative graduate GPA and 0.26 to 0.51 for comprehensive exam performance (Kuncel & Hezlett, 2007). Besides grades, other career- and life-achievements such as publications, patents, university tenure, and income are also predicted by standardized test scores (Robertson et al., 2010; Wai et al., 2005). Indeed, predictive validity remains one of the chief features of standardized tests of cognitive ability in work and educational settings.

Despite the abundant support for the use of standardized cognitive ability tests, critics of the tests remain unperturbed by the weight of the evidence. Kidder & Rosner (2002) noted that "the SAT only adds 5.4 percent of variance explained by HSGPA [high school grade point average] alone." (p.193) In response to a large meta-analysis on the validity of graduate admission tests (Kuncel & Hezlett, 2007), Bernard Brown, one of the commentators, noted that "They show a 0.40 correlation... but 0.40 is a limited predictor of an effect." (p.1695) Another commentator, James Shery, stated that "correlations are themselves overstated by the authors. More than half of the variance in 'later performance' cannot be attributed to observed variance in standardized test scores" (pp. 1695-1696) (Lerdau & Avery, 2007). In their view, a validity of 0.40 was not large enough to be useful for real-world applications. Despite the general pessimism from test critics, the same evidence was interpreted as providing "useful information for predicting subsequent student performance across many disciplines" by the meta-analysis's authors (Kuncel & Hezlett, 2007). Indeed, Funder & Ozer (2019) concluded in their review of effect sizes represented in social sciences that "r of 0.30 indicates a large effect that is potentially powerful in both the short and the long run" (p.156).

These divergent conclusions regarding test validity illustrate that the same evidence may result in very different interpretations. Furthermore, attitudes toward admission tests rest not only on the presence but also the interpretation of validity evidence (Kane, 2013; Maclver et al., 2014; Mattern et al., 2009; O'Leary et al., 2017). This is further exemplified in the UC's administrative decision to abandon standardized testing, contradicting the recommendations of their faculty task force. Arguably, failure to appreciate the predictive efficacy of testing can be partly attributed to how validity evidence is traditionally presented to and interpreted by the public. Furthermore, although standardized effect size metrics such as the correlation coefficient are ideal for the accumulation of knowledge across research contexts, they fail to present context-specific information necessary to make informed decisions (e.g., Phelps, 2005).

Alternative effect sizes presentations

Comprehension and interpretation of validity evidence may be improved using alternative presentations of statistical effect size (Brooks et al., 2014). Instead of a point estimate of the linear relation between predictor (e.g., ACT score) and outcome (e.g., GPA), statistical associations can be expressed in terms of probabilities or likelihoods of the desired outcome (e.g., academic achievement) based on a candidate's score on a predictor. Taking this approach, Rosenthal & Rubin (1982) developed the Binomial Effect Size Display (BESD) to communicate the practical impact of psychological interventions. The BESD presents experimental or correlational effects as a 2 by 2 matrix where the value of the outcome variable (e.g., GPA) is displayed as a proportion (e.g., % above a certain GPA) based on whether the predictor is above or below a certain value (e.g., median
ACT score). Similarly, the Expectancy Chart uses bar graphs to illustrate the expected outcome at a different level of a predictor (Lawshe & Bolda, 1958) and can be used to communicate the validity of selection tests (Cucina et al., 2017; Schrader, 1965). Lastly, the Common Language Effect Size (CLES) describes the probability that a random score on the outcome variable (e.g., GPA) from one group (those with an above average ACT) will differ from the other (below average ACT) (Dunlap, 1994; Krauskopf et al., 2018).

When judging an effect, Kirk (2001) noted that researchers are generally interested in knowing: 1) is the effect significantly different from zero? 2) how large is the effect? and 3) is the effect large enough to be useful? Whereas the first two questions can be addressed statistically using standardized metrics, an answer to the third question relies on context and is thus subject to human interpretation. Alternative effect sizes are ideal for communicating the usefulness of research findings to a lay audience for several reasons (Aguinis et al., 2010; Kuncel & Rigdon, 2012). First, whereas standardized effect sizes rely on technical statistical indices (e.g., correlation coefficient), a "consumer-centric" index uses common expressions such as proportions or frequencies that lay consumers and non-expert decision-makers can better comprehend, thereby, enhancing the understandability of the validity evidence (Brooks et al., 2014). Secondly, the central purpose of validity evidence is to allow decision-makers to ascertain the probability that an individual test taker will reach some level of achievement or performance based on test information (Kane, 2015). Traditional effect sizes only produce a standardized point estimate of the effect size (e.g., $r = .30$), rendering it uninterpretable to evaluate the practical impact on selection outcomes. In contrast, alternative effect sizes present the expected outcomes of selection either numerically or graphically using proportions, which allows lay consumers to more directly see the practical value of selection tests as it relates directly to the outcome of interest, thereby improving the interpretability of the validity evidence.

Given that much of the existing criticisms against standardized tests revolve around predictive efficacy, we anticipate that the improved understandability and interpretability of alternative effect sizes will magnify the perceived validity of the test. Accordingly, we first expect that people will perceive the standardized admission test as more predictive when validity evidence is presented as one of the alternative effect sizes than a traditional effect size metric (e.g., correlation coefficient and coefficient of determination) (Hypothesis 1). Second, we expect that the perceived validity of standardized tests will be associated with more favorable attitudes toward test-use in higher education and more negative attitudes toward its removal. Specifically, we hypothesize perceived validity will be positively associated with people’s endorsement for standardized admission tests (Hypothesis 2a) and negatively associated with people’s endorsement of a test-optional policy (Hypothesis 2b). We also expect an indirect effect of validity presentation on admission test policies where perceived test validity serves as a mediator for the effect of validity presentation on test policy attitudes (Figure 1).

### The potential moderating role of socioeconomic status

The effects of perceived test validity on test acceptance may not hold for everyone. In addition to how validity evidence is presented, individual differences may also affect how individuals make decisions based on that evidence. Research from motivated reasoning suggests that people attain aspects of their self-esteem based on the group to which they belong (i.e., in-group favoritism, Dasgupta, 2004; Kunda, 1990). Furthermore, in-group favoritism may affect how people interpret statistical information (Schaller, 1992). Schaller and colleagues find that group membership is related to people’s ability to engage in “intuitive ANCOVA,” where inferences about linear associations (e.g., gender and leadership competence) depend on the individual’s group membership (Schaller, 1992; Schaller & O’Brien, 1992). For example, when presented with evidence of a gender difference in leader competence, males were more likely to consider other confounds as explanations for the relationship (e.g., gender difference in leader representation). And as a result, females were less likely to conclude that males were better leaders when presented with correlational evidence that links gender to competence.

Building on this work, we posit that the perceived association between test scores and academic performance (i.e., perceived test validity) will differentially impact attitudes toward test-related policies based on the receiver’s socio-economic status. Standardized tests are often criticized for their financial burden to test takers, which may adversely impact individuals with lower SES (Atkinson & Geiser, 2009; c.f. Sackett et al., 2009). Some critics have labeled admission tests as “wealth tests,” implying that the tests reflect gaps in wealth, rather than developed ability or achievement. The use of standardized tests for admissions, therefore, is believed by many to be disadvantageous to people with lower SES. Accordingly, we expect people with lower SES will be more likely to discount that evidence linking test scores to achievement by considering the presence of confounds (e.g., SES difference in test scores) as explanations when forming their attitudes toward test adoption. In contrast, we expect people with higher SES will be more likely to attribute validity evidence as support for the true predictive efficacy of test scores and academic achieve-

![Figure 1. Theoretical model of validity perceptions and admission test policy preferences.](http://online.ucpress.edu/collabra/article-pdf/7/1/24087/465758/collabra_2021_7_1_24087.pdf)
ment driven by developed ability, thereby legitimizing the test validity as evidence in favor of its use in admission decisions. Therefore, we hypothesize SES to moderate the association between perceived validity and attitudes toward tests. The positive (negative) association between perceived test validity and general admission test endorsement (test-optional policy endorsement) will be weaker for individuals with lower SES than those with higher SES. Our hypotheses are also tested as a moderated mediation within our theoretical model such that the indirect effect of the validity presentation manipulation on subsequent dependent variables via an increase in test validity will be stronger for higher (vs. lower) SES individuals (Hypothesis 3a & 3b, see Figure 1).

Overview of experiments

We conducted two randomized experiments to test our hypotheses. Study 1 was conducted using an adult online sample. We presented participants with the validity of the ACT, a standardized admission test, using either a traditional ($r$ and $r^2$) or alternative (CLES, BESD, Expectancy Chart) statistic. Participants indicated their perceived validity of the ACT as well as their endorsement for universities using the ACT for admission decisions and the test-optional policy adoption, which removes the requirement for students to submit standardized test scores when applying to a university. Study 2 was a pre-registered field experiment conducted with students and community members at a large southern university campus. In addition to replicating the main hypotheses in a field sample, we incorporated a behavioral measure of policy endorsement by asking participants if they were willing to sign a petition for their university to adopt a test-optional policy for undergraduate admissions. Following best open science practices, study materials, data, and details of pre-registration can be found at https://osf.io/xfe6h/?view_only=c9e5329f1486f785a3b0aaac9e91b138. We also confirm that all measures, conditions, data exclusions, and sample size planning is accurately reported for both experiments.

Study 1: Methods

Sample

We obtained data using Prolific Academic, an online research crowd-sourcing platform based in the U.K. (Peer et al., 2017). Because our experiment focused on people’s attitudes toward the ACT, which is used primarily in the U.S., we only allowed participants from the U.S. to participate. Based on a conservative estimate of small treatment effect ($f = 0.25$), the recommended sample size for an experiment with five groups, 95% power, and $\alpha = 0.05$ was 305. We also used the R package: semPower following the Neyman-Pearson approach outlined by Moshagen & Erdfelder (2016) to perform a power analysis of the full hypothesized model. Following this approach, we specified a null model where the hypothesized experimental effect and mediation effect were constrained to zero. The recommended sample size to achieve 80% power was 348. Data was obtained from 355 participants (Mean Age = 31.1 [SD = 10.7]; 56% Male; 71% Caucasian; 49% Democrat; 12% Republican; 27% Independent). Fifty-eight percent of the participants had at least an Associate degree or higher. The median yearly income bracket reported by the participants was between $40,000 to $59,999.

Stimuli and procedure

We conducted a between-subjects randomized experiment to examine the effects of validity presentation on validity inferences as well as subsequent attitudes toward standardized admission test policies. At the start of the study, participants were presented with a short statement describing standardized admission tests (Appendix A). Next, participants were told that they would be reviewing some data collected on ACT scores and college GPA from a large southern university.

Participants were randomly presented with one of five effect size presentation conditions ($r$, $r^2$, CLES, BESD, and Expectancy Chart) illustrating the association between ACT scores and GPA. To create the effect size statistics stimulus, we used archival data from students’ self-reported grades and ACT scores gathered for a previous study. The bivariate correlation between GPA and ACT in the sample dataset was 0.50, which is consistent with meta-analytic estimates typically found in the literature (Kuncel & Hezlett, 2010; Sackett et al., 2008). The same sample data was then used to create the stimuli used for the experiment using the Alternative Effect Size Display Calculator (Zhang, 2018). Table 1 contains the exact stimuli and descriptions used. Next, participants responded to dependent variable measures, which are described in the next section.

Measures

Perceived test validity. We used a 4-item measure developed by Kim & Berry (2015). Table 2 contains a list of items. An example item is "The ACT score is a strong predictor of college GPA." Participants responded to each item on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The reliability of the measure was $\alpha = 0.89$.

Admission test endorsement. We used a 6-item measure developed by Kim & Berry (2015). Table 2 contains a list of items. An example item is "I believe the ACT should be used for college admissions." Participants responded to each item on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The reliability of the measure was $\alpha = 0.91$.

Test-optional policy endorsement. We created a 6-item measure for test-optional policy endorsement (Table 2). Before responding to the items, participants read a short description of the test-optional policy. To ensure that participants understood what the policy entailed, we added a comprehension check question after the description (see Appendix B). Participants had to provide the correct answer to the attention check before responding to the items. Participants responded to the first four items on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). For the remaining two items, participants responded to item 11 in Table 2 on a 5-point Likert scale ranging from 1 (much less favorable) to 5 (much more favorable); and
they responded to item 12 on a 5-point Likert scale ranging from 1 (much less likely) to 5 (much more likely). The reliability of the measure was \( \alpha = 0.95 \).

**Subjective socioeconomic status (SES).** We used a common measure of subjective SES developed by Adler et al. (1994). Participants were shown a 9-rung ladder and asked to indicate – based on their subjective judgment – where they saw themselves relative to other people in their community in terms of a combination of income, education, and occupation.

**Study 1: Results and Discussion**

We used a two-phase Structural Equation Modeling approach to test our model. Before data analysis, we created \( k-1 \) dummy-coded variables to denote the experimental condition assignment with the correlation coefficient group serving as the referent (J. Cohen et al., 2013).

**Measurement models**

We first used confirmatory factor analysis (CFA) to examine the measurement model based on the items used in the study. Specifically, we examined the model fit of the three-factor model based on the hypothesized constructs as well as alternative plausible models. The fit indices for the hypothesized three-factor model were all satisfactory based on common cut-off criteria (RMSEA < 0.08, SRMR <
Table 2. Factor loadings of the dependent variable items.

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>There is a strong association between ACT scores and college GPA</td>
<td>0.803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2*</td>
<td>The ACT is an accurate tool for predicting college GPA</td>
<td>0.863</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3*</td>
<td>The ACT score is a strong predictor of college GPA</td>
<td>0.892</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Students who perform well on the ACT tend to do better in college</td>
<td>0.706</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5*</td>
<td>I believe the ACT should be used for college admissions</td>
<td></td>
<td>0.837</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>I think the ACT should not be so important in college admissions (-)</td>
<td></td>
<td>-0.776</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ACT scores provide very useful information about an applicant</td>
<td></td>
<td>0.713</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Schools should stop using the ACT to evaluate applicants (-)</td>
<td></td>
<td>-0.737</td>
<td></td>
</tr>
<tr>
<td>9*</td>
<td>ACT scores should be an important factor when deciding whether an applicant is admitted to college</td>
<td></td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td>10*</td>
<td>I feel satisfied knowing that the ACT is used in college admissions</td>
<td></td>
<td>0.864</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>I endorse the “test-optional” policy</td>
<td></td>
<td>0.935</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>The “test-optional” policy is a good idea</td>
<td></td>
<td>0.931</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>The “test-optional” policy should be adopted at more universities</td>
<td></td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>The “test-optional” policy is fair</td>
<td></td>
<td>0.848</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>If a university or college has a ‘test optional’ policy, how would this affect your general attitude toward this university?</td>
<td></td>
<td>0.816</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>If a university or college has a ‘test-optional’ policy, how would this affect your likelihood of encouraging your family or friends to apply to this university?</td>
<td></td>
<td>0.728</td>
<td></td>
</tr>
</tbody>
</table>

Notes. * denotes the items used in the shortened version of the scales used in Study 2. (-) denotes the negatively worded item.

0.05, CFI > 0.95; Hu & Bentler, 1999; Kline, 2015. The fits of alternative measurement models were all well outside of the criteria for model fit (see supplemental materials). The three-factor model was retained in the subsequent test of hypotheses.

Latent path model

We conducted an SEM with the three latent variables and dummy-coded experimental variables to examine the theoretical model depicted in Figure 1. The hypothesized latent path model had good fit, RMSEA = 0.065; SRMR = 0.050; CFI = 0.949; BIC = 12097.110. Path estimates were made based on 5,000 bootstrapped resamples (Table 3). The effect of extent size communication on perceived predictive validity is illustrated in the regression weights between each dummy-coded experimental variable and the mediator. Because the correlation coefficient condition was used as the reference group, regression weights are interpreted as the difference in the dependent variable between each dummy-coded condition and the correlation coefficient group. A significant regression weight means a significant difference in perceived test validity between the reference group (correlation coefficient) and the dummy-coded group. Based on the results, there was a significant difference in the perceived validity of the ACT between participants who saw a correlation coefficient versus each of the three alternative effect sizes.

Post-hoc comparisons revealed that, compared to the correlation coefficient condition (M = 2.80, SD = 1.04), people judged the ACT to be significantly more predictive when validity information was presented as BESD (M = 3.58, SD = .81, Cohen’s d = .84, p < .001); CLES (M = 3.46, SD = .86, d = .69, p < .001); and expectancy chart (M = 3.83, SD = .67, d = 1.17, p < .001). However, there was no significant difference in the perceived validity of the ACT between people presented with a correlation coefficient and a coefficient of determination (M = 3.10, SD = .95, d = .26, p = .24) (Figure 2). Taken together, Hypothesis 1 was fully supported.

Test of mediation. As hypothesized, we found that the perceived validity of the ACT positively predicted people’s endorsement for the ACT as an admissions test (Hypothesis 2a) and negatively predicted their endorsement of the test-
optional policy (Hypothesis 2b). The estimated indirect effects of validity presentation on both dependent variables via an increase in perceived validity were statistically significant (Table 3). To further establish evidence for mediation, we tested an alternative structural model where the direct effects between the experimental manipulation variables and the two outcome variables were estimated. The direct effect model – despite added model complexity – had worse relative fit than the indirect effects only model (ΔBIC = -21.82).

Taken together, the results suggest that perceived validity fully mediated the effect of validity presentation on subsequent attitudes toward the ACT as an admission test as well as the endorsement of the test-optional policy. In other words, through an increase in perceived test validity, people who viewed one of the alternative validity presentations showed more support for the ACT as an admissions test and lower endorsement for the test-optional policy than those presented with a traditional effect size display.

Test of moderated mediation. Hypotheses 4a and 4b were tested using bootstrapped mediation with the PROCESS macro in SPSS (Hayes, 2017). For simplicity in interpretation and analysis, we created a new dummy-coded variable to dichotomize the experimental factor such that participants in the three alternative validity presentation conditions (CLES, BESD, and Expectancy Chart) were combined into a single group and participants in the two traditional validity groups (r and \( r^2 \)) were also combined. We feel this aggregation was justified because of the clear difference in validity perceptions between the alternative versus traditional validity presentation groups and the similarities between variants of alternative (and traditional) validity statistics.

Table 3 contains the result of the moderated mediation analysis. We examined the indirect effects for each DV conditioned on varying levels of subjective SES. First, we found indirect effects of validity presentation (traditional vs. alternative) on test admission to be positive for all three levels of SES. The index of moderation was not statistically significant. Therefore, we did not find support for Hypothesis 4a.

We found the indirect effect of validity presentation mode on the test-optional policy was significant when SES was high (indirect effect = -0.21) or moderate (indirect effect = -0.11), but not when subjective SES was low (indirect effect = -0.01). The index of moderated mediation was also statistically significant. Therefore, Hypothesis 4b was supported. To further aid interpretation, we plotted the linear associations between perceived test validity and test-optional endorsement across different levels of subjective SES (Figure 3, McCabe et al., 2018) using multiple regression. Results are consistent with the expected directions of the effect. The negative association between perceived validity and test-optional policy endorsement was stronger for individuals with high SES, but not low or moderate SES.

Supplemental analysis

Although not hypothesized, we also examined the experimental effects of validity presentation on admission test endorsement and test-optional policy endorsement directly. Between-subjects ANOVA revealed a significant main effect of validity presentation on admission test endorsement, \( F(4, 350) = 3.75, p = .005, \eta^2 = .04 \), but not test-optional policy endorsement, \( F(4, 350) = 2.05, p = .09, \eta^2 = .02 \).

In the first experiment, we found that people who viewed one of the three alternative validity presentations perceived the admission test to be more predictive. Furthermore, the increase of validity perception was associated with 1) higher endorsement of college admission tests and 2) lower endorsement for the test-optional policy. These findings provide initial evidence that perceived test validity is malleable and depends on how validity evidence is communicated.
Table 4. Summary of indirect effects and conditional indirect effects of validity presentation for Study 1.

<table>
<thead>
<tr>
<th>Paths</th>
<th>Estimate</th>
<th>SE</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV = Admission Test Endorsement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirect Effect</td>
<td>0.15</td>
<td>0.03</td>
<td>[0.10, 0.20]</td>
</tr>
<tr>
<td>Conditional Indirect Effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1 SD SES</td>
<td>0.17</td>
<td>0.03</td>
<td>[0.11, 0.24]</td>
</tr>
<tr>
<td>Mean SES</td>
<td>0.15</td>
<td>0.03</td>
<td>[0.10, 0.21]</td>
</tr>
<tr>
<td>+1 SD SES</td>
<td>0.13</td>
<td>0.03</td>
<td>[0.08, 0.20]</td>
</tr>
<tr>
<td>Index of Moderated Mediation</td>
<td>-0.01</td>
<td>-0.01</td>
<td>[0.03, 0.01]</td>
</tr>
<tr>
<td>DV = Test-Optional Policy Endorsement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirect Effect</td>
<td>-0.12</td>
<td>0.04</td>
<td>[-0.20, -0.04]</td>
</tr>
<tr>
<td>Conditional Indirect Effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1 SD SES</td>
<td>-0.01</td>
<td>0.05</td>
<td>[-0.12, 0.08]</td>
</tr>
<tr>
<td>Mean SES</td>
<td>-0.11</td>
<td>0.04</td>
<td>[-0.20, -0.04]</td>
</tr>
<tr>
<td>+1 SD SES</td>
<td>-0.21</td>
<td>0.07</td>
<td>[-0.34, -0.09]</td>
</tr>
<tr>
<td>Index of Moderated Mediation</td>
<td>-0.05</td>
<td>0.02</td>
<td>[-0.09, -0.01]</td>
</tr>
</tbody>
</table>

Notes: IV = Validity Presentation Condition (0 = Traditional, 1 = Alternative). Mediator = Perceived Test Validity. C.I. = confidence interval.

Figure 3. Simple slopes (b) between perceived test validity and test-optional endorsement for Study 1 at 1SD/ 2SD above and below the mean for subjective SES.

Each graphic shows the computed 95% confidence region (shaded area), the observed data (gray circles), the maximum and minimum values of the outcome (dashed horizontal lines), and the crossover point (diamond). The x-axes represent the full range of the focal predictor. CI = confidence interval; PTCL = percentile.

Furthermore, the perception of test validity is meaningfully related to admission test-related policy preferences.

Study 2: Methods

Sample

We collected data from participants at a large, public, and moderately selective southern university in the U.S.

1 The undergraduate acceptance rate at this university is 75%. The average ACT score of the accepted students is 26.
We followed the pre-registered plans to obtain a minimal sample of 200 participants. We oversampled and stopped collecting data at 215 recorded responses. We removed surveys where the respondent did not complete any dependent variables. After exclusions, the sample size was 198 participants: 55% female, 64% Caucasian, 47% Democrat, 22% Republican, 17% Independent. The mean age was 20.90 (SD = 3.83). Seventy-seven percent had a bachelor’s degree or higher. Ninety-seven percent of the respondents were students and included 22% freshmen, 26% sophomores, 18% juniors, 23% seniors, and 8% graduate students.

**Stimuli and procedure**

To collect data for this field study, research assistants asked students and staff passing by the university library if they would participate in a survey about standardized admission tests. Those who agreed were presented with an iPad to participate in the study. The overall procedures were identical to Study 1. After completing the study on an iPad, we asked the respondents if they would endorse a test-optional policy by physically signing a petition.

**Measures**

**Perceived test validity.** To minimize survey time, we used a shortened 3-item version of Kim and Berry’s (2015) scale to measure participant’s perceptions of test validity (see Table 2). An example item includes, "There is a strong association between ACT scores and college GPA." Participants responded to each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The reliability of the measure was α = 0.87.

**Admission test endorsement.** We measured test endorsement using a 3-item shortened version of Kim and Berry’s (2015) scale (see Table 2). An example item is "ACT scores should be an important factor when deciding whether an applicant is admitted to college." Participants responded to each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The reliability of the measure was α = 0.88.

**Test-optional policy endorsement.** We used a single-item measure to assess test-optional policy endorsement. Like in Study 1, before responding, participants read a description of the test-optional policy. The participants were then asked whether they would endorse the university’s decision to have a test-optional policy. The participants responded to the item on a 5-point Likert type scale (1 = strongly oppose; 5 = strongly favor).

**Test-optional policy petition.** We obtained a behavioral measure of test-optional support by asking participants if they would be willing to physically sign a petition for their university to adopt the test-optional policy. We recorded participants’ willingness to sign the petition as either yes (1) or no (0).

**Subjective socioeconomic status.** We used the same measure as Study 1 to measure subjective SES.

**Stimuli and procedure**

To collect data for this field study, research assistants asked students and staff passing by the university library if they would participate in a survey about standardized admission tests. Those who agreed were presented with an iPad to participate in the study. The overall procedures were identical to Study 1. After completing the study on an iPad, we asked the respondents if they would endorse a test-optional policy by physically signing a petition.

**Measures**

**Perceived test validity.** To minimize survey time, we used a shortened 3-item version of Kim and Berry’s (2015) scale to measure participant’s perceptions of test validity (see Table 2). An example item includes, "There is a strong association between ACT scores and college GPA." Participants responded to each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The reliability of the measure was α = 0.87.

**Admission test endorsement.** We measured test endorsement using a 3-item shortened version of Kim and Berry’s (2015) scale (see Table 2). An example item is "ACT scores should be an important factor when deciding whether an applicant is admitted to college." Participants responded to each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The reliability of the measure was α = 0.88.

**Test-optional policy endorsement.** We used a single-item measure to assess test-optional policy endorsement. Like in Study 1, before responding, participants read a description of the test-optional policy. The participants were then asked whether they would endorse the university’s decision to have a test-optional policy. The participants responded to the item on a 5-point Likert type scale (1 = strongly oppose; 5 = strongly favor).

**Test-optional policy petition.** We obtained a behavioral measure of test-optional support by asking participants if they would be willing to physically sign a petition for their university to adopt the test-optional policy. We recorded participants’ willingness to sign the petition as either yes (1) or no (0).

**Subjective socioeconomic status.** We used the same measure as Study 1 to measure subjective SES.

**Study 2: Results and Discussion**

We followed the pre-registered analytical plans to test our hypotheses. First, we used an independent t-test to examine the effect of alternative (versus traditional) validity display on the perceived validity of standardized admission tests. Consistent with Hypothesis 1, people that viewed one of the three alternative validity displays (CLES, BESD, and Expectancy Chart) reported the standardized test as more valid (M = 2.74, SD = 1.10) than people that viewed a traditional effect size (M = 2.26, SD = 1.00), t(172) = 3.19, p = .001, d = 0.46. Between-subject ANOVA revealed similar results, F(4, 193) = 4.44, p = .002, η2 = .08 (Figure 4). When these findings are taken together, Hypothesis 1 was fully supported.

**Latent path analysis**

Next, we examined the indirect effect of validity presentation on the respondent’s endorsement of standardized tests for admissions (Hypothesis 2a) and endorsement for the test-optional policy (Hypothesis 2b). Like Study 1, we used SEM to test our hypotheses. Because a categorical indicator (petition signature) was included, we used WLSMV estimation for our model testing, which is robust against non-normally distributed and categorical indicators (Li, 2016). We first examined the fit of the measurement model. Perceived test validity and general test endorsement were formed with the three respective items. We used the single-item measure of test-optional policy endorsement as well as people’s willingness to sign the petition as indicators of the test-optional endorsement latent construct. The three-factor measurement model had good fit (RMSEA = 0.00, SRMR = 0.018, CFI = 1.00) and was retained for path analysis. Alternative models were also explored and exhibited significantly worse fit than the three-factor model (see supple-

---

2 https://osf.io/wgnip
mental materials).

Path estimates were made based on 5,000 bootstrapped resamples. Consistent with Study 1, we found that the perceived validity of the ACT positively predicted people's endorsement for the ACT as an admissions test (Hypothesis 2a) and negatively predicted their endorsement of the test-optional policy (Hypothesis 2b) (Table 3). The estimated indirect effects of validity presentation on the policy endorsement, however, depended on the specific type of validity display. Specifically, the Expectancy Chart had the largest indirect effect on the perception of validity and indirect effects on subsequent policy preferences. The effect of the CLES condition on policy preferences was mixed. The indirect effect on general test endorsement was statistically significant, but not with the test-optional policy endorsement. The BESD condition did not produce a more favorable perception of test validity nor higher endorsement for admission tests, in general.

**Perceived Validity and Petition Signatures.** We conducted logistic regression to examine the association between perceived test validity and the likelihood of signing the petition to adopt the test-optional policy. We found that perceived validity was negatively associated with the likelihood of signing the petition, $\beta = -0.66$ [95% C.I.: -0.97, -0.36], Wald = 27.8, $p < .001$. The estimated odds ratio suggested that for every unit increase in the perceived admission test validity, participants were roughly half as likely to sign the petition to adopt the test-optional policy, $\text{Exp}(\beta) = 0.51$ [95% C.I.: 0.38, 0.69]. These findings provide more robust support for the association between perceived test validity and policy preferences using multiple methods.

**Test of moderated mediation.** To maximize statistical power for moderation analysis, we chose to combine data from Studies 1 and 2, as planned in the pre-registration. To standardize the measures, we transformed the dependent variables into the percent of maximal possible score (P. Cohen et al., 1999). Using a combined dataset, we did not find evidence for moderated mediation for either general test endorsement (Index of moderation = 0.002, [95% C.I.: -0.02, 0.06]) or test-optional policy endorsement (Index of moderation = -0.004, [95% C.I.: -0.012, -0.005]). Still, the indirect effect of validity presentation format on general test endorsement ($\beta = 0.050$, [0.034, 0.068]) and test-optional policy endorsement ($\beta = -0.022$, [95% C.I.: -0.036, -0.010]) remained statistically significant in the combined dataset.

**Supplemental analyses**

We also examined the experimental effects of validity presentation on admission test endorsement and test-optional policy endorsement directly. Between-subjects ANOVA revealed a significant main effect of validity presentation on admission test endorsement, $F(4, 195) = 2.70$, $\eta^2 = .03, p = .032$; but not test-optional policy endorsement, $F(4,192) = 0.46$, $\eta^2 = .01, p = .77$. Thus, we replicated the effect of validity presentation on perceived test validity. But the experimental effect of validity presentation on test policy preferences was not significant. Nevertheless, these exploratory analyses were not part of the hypotheses nor were they pre-registered.

Overall, we replicated the main findings from the first online experiment. In a community sample of primarily college students, we found more favorable attitudes toward admission tests when validity evidence was presented using an alternative effect size. Perceived test validity was the highest when participants viewed an expectancy chart. We also found that perceived test validity was associated with both general admission test endorsement (positive) and test-optional endorsement (negative). Finally, the self-reported support for test-optional endorsement extended to a behavioral measure of signing a physical petition.

**General Discussion**

Despite imperfections, standardized college admission tests (i.e., ACT and SAT) are construct-valid measures of developed general mental ability and one of the best predictors of an array of future academic, career, and life achievements (e.g., Brown et al., 2021; Kuncel et al., 2004; Robertson et al., 2010) and these tools have been carefully calibrated due to decades of criticism to minimize bias and adverse impact (e.g., Cronbach, 1975; Phelps, 2005; Sackett et al., 2008). Nevertheless, test critics maintain that standardized tests are not predictive of academic performance.

In this paper, we examined three alternative effect size presentations and their effects on validity perception and test-related policy preferences. We found strong support for our hypotheses that people will perceive the admission test as more predictive when validity evidence is presented as one of the alternative effect size displays rather than a traditional metric (i.e., correlation coefficient). The effects are most pronounced when an Expectancy Chart is used. We also found support that validity presentation indirectly affects subsequent policy-related preferences through validity perception. Specifically, the perceived validity of standardized tests was positively associated with endorsement of using the test for admission decisions and negatively associated with endorsement of the test-optional policy. Of the three alternative effect size presentations, we found that the Expectancy Chart produced the largest effect on validity understanding and admission test endorsement. These findings are consistent with and replicate aspects of previous studies that examined graphical visual aids on statistical comprehension and validity perception (Zhang et al., 2018). Overall, these results suggest that 1) perception of validity is malleable and depends on how evidence is presented; and 2) validity perceptions are associated with admission test related policy preferences.

Although we found that validity presentation had indirect effects on admission test endorsement and test-optional policy endorsement through an increase in validity perception, the main experimental effects of validity presentation on policy preferences were mixed. In both experiments, we found a modest effect of validity presentation on general test endorsement. People reported more favorable attitudes toward colleges using admission tests when validity evidence was presented using an alternative effect size display. We did not, however, find conclusive evidence for effect size presentation directly affecting test-optional policy endorsement. Whereas the first study using an online sample provided some support, it was not replicated in...
the field experiment. In general, we also found weaker experimental effects in the field experiment, which is not surprising given the added environmental noise. Nevertheless, we still found support for our primary hypotheses related to validity perception and general admission test endorsement in both Studies 1 and 2.

We found that the effect of validity perception on test-optimal policy endorsement – but not general test admission – was moderated by subjective SES in Study 1, which provides partial support for our hypotheses. The moderation effect was not replicated with the addition of the second field experiment. It is worth mentioning that the field experiment was conducted on a sample of mostly college students in a large public university, who might be more socio-demographically homogenous than a general sample. Furthermore, students on a university campus are all selected based, at least partly, on their performance on the ACT. Therefore, the sample fails to include students who were rejected based on their test scores or for other reasons.

Practical implications

Our findings have clear implications for how academics communicate their findings—especially their effect sizes—with the public (e.g., media, executives, policymakers) (Lewis & Wai, 2021). As noted by Camara & Shaw (2012), many academics are reluctant to engage with the public on issues about psychometrics and measurement due to the complexity of statistical prediction and psychometric theory and inherent challenges in conveying that complexity in plain language. Additionally, academics who understand the full body of evidence surrounding standardized tests are often not consulted by the media (Phelps, 2005). Consequently, the public may often under-value the practical impact of psychometric findings or more technical findings generally when expressed as a simple correlation or proportion of variance. As we showed in this paper, however, alternative effect size metrics such as the CLES may be ideal when summarizing the predictive efficacy of testing in simple statements. And where possible, Expectancy Charts are ideal if the goal is to communicate the practical impact of statistical prediction as it relates to real-world outcomes. Of course, depending upon one’s political aims (either to illuminate or obfuscate), our findings also illustrate which types of statistics are useful towards those ends. Nevertheless, we maintain that when possible, researchers should present evidence in a way that the public can most readily understand. Only then can we ensure that policies related to admission tests are enacted based on the proper evidence that has accumulated to date.

Our research also has implications for statistical education in applied social sciences (e.g., education, public policy, management). Considering that most research findings rely on statistical associations between variables or treatments, we believe that statistical and methodological curriculums should include alternative effect size metrics described in this paper to help educate about the malleability of statistical perception and the importance of comparing statistical information using the same communication display mechanisms. Relatedly, practitioners should be instructed on how to present and interpret research evidence as it relates to their practice (e.g., management, educational policy). Arguably, the gap between different research fields with different theoretical and methodological traditions as well as the gap between research and practice can be reduced if everyone has a greater understanding and appreciation for statistical evidence and how to accurately interpret it in the appropriate context.

Limitations and future directions

Our experiments focused on testing ways of communicating the bivariate correlation between a single predictor and criterion. Future research might examine the communicative efficacy of alternative validity presentations in the context of multiple predictors (e.g., Krasikova et al., 2018), incremental predictions (e.g., Mattern et al., 2011), or effect sizes in meta-analyses or other contexts where they are widely used and there are varying standards (Kraft, 2020). Relatedly, the presentation of alternative effect sizes requires a selection of cut-off values for the criteria (e.g., proportion of students with GPA > 3.5). Depending on the choice of cut-off, the resulting evidence will likely differ. Future research, therefore, should more comprehensively examine the different ways in which alternative effective sizes can be presented and how these choices may affect the interpretation of evidence. Second, future research could also extend our theoretical model to include other individual and situational factors that may influence how effect sizes are perceived and utilized. Cognitive factors such as numeracy, spatial reasoning, or graph literacy may, for example, influence the perception of effect size statistics. And non-cognitive factors such as personality and worldview may influence how validity evidence is utilized in policy preferences and decisions (Highhouse & Rada, 2015). Additionally, the samples used in this study may not fully represent those who are responsible for policymaking (e.g., administrators). Nevertheless, our sample represents a wide range of constituents (parents and students) whose voices and preferences do indeed affect public policy (Chen, 2019; Henderson et al., 2019). Still, the generalizability of our findings needs to be explored further. Future research should examine other contextual moderators such as the test itself (undergraduate versus graduate admission tests) and test context (education versus employment). Our study relied primarily on ad-hoc measures that were adapted for our research, which may potentially suffer from reduced construct validity. Although we took care to ensure that the items used in our study meaningfully distinguished the constructs of interest, future research could improve on the measures used here. Finally, our experiments focused primarily on how the interpretation of validity evidence affects policy preferences related to admission testing. In practice, attitudes toward standardized testing and decisions toward its use are multi-determined. Other factors such as the cost of the test and adverse impact are also relevant for understanding whether standardized tests are used. Although an experimental design allows us to rule out these factors as potential confounds for our study, future research should take these factors into consideration when modeling the decision processes related to admission test policies.
Conclusions

The practical impact of psychological research on policy relies on clear communication of research evidence to the public. Despite the belief held by many test critics that standardized admission tests are not useful for predicting future performance, our research suggests that the perception of test validity is malleable and is dependent on how effect size evidence is presented. We showed in our paper that when the validity of standardized tests is presented using alternative effect size displays instead of traditional statistics (e.g., Pearson’s r), people perceive the test to be more predictive and are more willing to endorse its use for admission decisions.

Contributions

Contributed to conception and design: DZ
Contributed to acquisition of data: DZ
Contributed to analysis and interpretation of data: DZ, JW

Acknowledgements

We would like to thank following research assistants for their instrumental help in conducting this research: Sheryl Lobo, Olivia Kjellsten, Kristen Waguespack.

Funding Information

We did not receive any funding for this research.

Competing Interests

The authors have no competing interests to declare.

Data Accessibility Statement

All data and material can be found on OSF at https://osf.io/xf6gh/?view_only=c9c5329f14864785a3b0aaac9e91b138.

Submitted: January 12, 2021 PDT, Accepted: May 13, 2021 PDT
REFERENCES


Appendices

Appendix A

General Instructions

"Standardized aptitude tests such as the ACT and SAT are frequently used by college admission committees in the United States. The ACT, specifically, consists of four sections: math, reading, science, and English. The scores range from 1 to 36.

One reason why these tests are used for college admission decisions is because students with higher scores are more likely to have a higher college GPA, and thereby more likely to graduate and get hired."

Appendix B

Test-optional Policy Description

"In the past several years, a number of American universities have adopted a "test-optional" policy. These colleges and universities do not require applicants to submit their standardized test scores (e.g., ACT or SAT) when applying. Instead, applicants are generally evaluated based on other factors such as high school grades, extracurricular activities, and curriculum difficulty."

Please answer the following questions about this policy.

Test-optional Comprehension Check

What is the "test-optional" policy?

a. Universities do not require applicants to submit their standardized test scores
b. Universities do not use any tests in their classrooms
c. Universities do not enforce drug test policies on campus
SUPPLEMENTARY MATERIALS

Peer review history

Supplemental material