education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 33 Number 28 April 22, 2025 ISSN 1068-2341

A Validation Review of the SAT and ACT for College and University Admissions Decisions

Audrey Amrein-Beardsley
Arizona State University
Zarrina T. Azizova
University of North Dakota
Norman P. Gibbs
Mesa Public Schools
Chukwu(emeka) Ikegwuonu
St. Cloud State University
Jeongeun Kim
University of Maryland, College Park
Deborah Michele La Torre
University of California, Los Angeles
Matthew R. Lavery
South Carolina Education Oversight Committee
Margarita Pivovarova
&
Yi Zheng
Arizona State University
United States

Citation: Amrein-Beardsley, A., Azizova, Z. T., Gibbs, N. P., Ikegwuonu, E., Kim, J., La Torre, D. M., Lavery, M. R., Pivovarova, M., & Zheng, Y. (2025). A validation review of the SAT and ACT for college and university admissions decisions. *Education Policy Analysis Archives, 33*(28). https://doi.org/10.14507/epaa.33.8734

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 24/5/2024
Revisions received: 10/2/2025
Accepted: 21/3/2025

Abstract: In response to a call for research on using the SATs and ACTs for U.S.-based college and university admissions, researchers systematically interrogated the literature surrounding both tests, using a framework for validity evidence built upon the *Standards of Educational and Psychological Testing* and Kane's contemporary view of validity. Authors of the 72 peer-reviewed articles selected for this review provided mixed validity evidence. Results indicated that both tests are psychometrically strong. However, when they are put into practice, especially when important decisions are to be made using test output, the validity of the decisions derived using SAT and ACT output is, especially for students from racial minority and socioeconomically disadvantaged backgrounds, at issue.
Keywords: admissions; assessment; educational policy; higher education; law/legal; test theory/development; validity/reliability

Una revisión de validación del SAT y ACT para decisiones de admisión en universidades y colegios

Resumen: En respuesta a una convocatoria de investigación sobre el uso del SAT y el ACT para la admisión en universidades y colegios en Estados Unidos, los investigadores analizaron sistemáticamente la literatura relacionada con ambas pruebas, utilizando un marco de evidencia de validez basado en los Estándares para pruebas educativas y psicológicas y la perspectiva contemporánea de validez de Kane. Los autores de los 72 artículos revisados por pares seleccionados para esta revisión proporcionaron evidencia de validez variada. Los resultados indicaron que ambas pruebas son psicométricamente sólidas. Sin embargo, cuando se aplican en la práctica, especialmente cuando se toman decisiones importantes basadas en sus resultados, la validez de dichas decisiones, en particular para estudiantes de minorías raciales y con desventajas socioeconómicas, se pone en cuestión.
Palabras-clave: admisiones; evaluación; política educativa; educación superior; derecho/legal; teoría del test/desarrollo; validez/confiabilidad

Uma revisão de validação do SAT e ACT para decisões de admissão em faculdades e universidades

Resumo: Em resposta a um chamado por pesquisas sobre o uso do SAT e ACT para admissão em faculdades e universidades nos Estados Unidos, os pesquisadores analisaram sistematicamente a literatura relacionada a ambos os testes, utilizando uma estrutura de evidência de validade baseada nos Padrões para Testes Educacionais e Psicológicos e na visão contemporânea de validade de Kane. Os autores dos 72 artigos revisados por pares selecionados para esta revisão apresentaram evidências de validade mistas. Os resultados indicaram que ambos os testes são psicometricamente robustos. No entanto, quando aplicados na prática, especialmente quando decisões importantes são tomadas com base em seus resultados, a validade dessas decisões, particularmente para estudantes de minorias raciais e contextos socioeconômicos desfavorecidos, é questionável.
Palavras-chave: admissões; avaliação; política educacional; ensino superior; direito/legal; teoria/desenvolvimento de testes; validade/confiabilidade

A Validation Review of the SAT and ACT for College and University Admissions Decisions

The SAT is a standardized test used for college admissions throughout the United States (US). Since the not-for-profit College Board debuted in 1926, its SAT acronym evolved from the Scholastic Aptitude Test to the Scholastic Assessment Test; however, the acronym SAT now stands for nothing more than the name of the test (i.e., the SAT). The SAT tests writing, reading, and mathematics. The ACT is also a standardized test used for college admissions throughout the US. Introduced in 1959 and formerly known as the American College Test, the ACT is currently administered by a nonprofit organization called ACT Incorporated. The ACT tests English, mathematics, reading, scientific reasoning, and writing, although the writing test is optional.

The SATs and ACTs have played key roles in college and university admissions decisions in the US and many other countries (Edwards et al., 2012) over this time. As such, scholars have repeatedly examined the validity of these tests, especially when used for such consequential purposes (Choi & Park, 2013; Huysamen, 1997; Koljatic et al., 2013; Liu & Wu, 2006; Yamamoto, 2016).

At the same time, developers of both tests continue to claim to have carefully developed and improved both tests over time, as per current cognitive and behavioral science theories over time (see, for example, Croft & Beard, 2021). However, much concern still surrounds both tests with regards to the adequacy of the evidence supporting how and to what extent both tests meet standards of validity, reliability, and fairness, for example, as defined by the American Educational Research Association ([AERA] et al., 2014). Such inconclusive and oft-contradictory evidence has limited the field from conceptualizing how to better (or best) utilize these tests for admissions.

From their inception, the SATs and ACTs have also been situated in, as well as complicated by multifaceted socio-political-legal contexts. Complicating the uses of these tests further include ranges of stakeholders involving test-developers, psychometricians, education policy researchers, international education partners, policymakers, test-users, and test-takers themselves, many of whom have varying intentions surrounding the SATs and ACTs (Chatterji, 2013).

Indeed, the SATs and ACTs have faced significant criticisms over the years given their reliance on standardized testing as predictive measures of students' academic potential in college. Critics, including many of the stakeholders just mentioned, view these tests as poor predictors of college success, arguing both overemphasize rote memorization and test-taking strategies over other important aspects of student's abilities, such as creativity, problem-solving, and personal character. Critics argue that these tests perpetuate inequality, as students from lower-income families or marginalized communities often lack access to resources like test prep, tutors, and extracurricular activities, perpetuating score disparities. Moreover, critics posit that both tests contribute to "test-centric" cultures that have greater albeit less observable, tangible, or measurable effects.

However, the SATs and ACTs endure due to a plethora of counterpoints and reasons. These include that both tests provide standardized, objective measures of students' academic abilities, allowing for relatively fairer comparisons across students from different schools, regions, and backgrounds on a common scale. This, proponents argue, is a better approach than relying on potentially more subjective factors like high school grade point averages (GPAs), which can greatly vary within and across high schools, districts, and states. The essays, interviews, and recommendation letters that often accompany SAT and ACT scores for college and university admissions are often positioned as more subjective by these same proponents. Proponents also argue that the SAT and ACT can help identify talented students who may not have access to advanced coursework or extracurricular opportunities, giving them a relatively better chance to succeed given their test performance. Additionally, they contend that both tests better encourage

students to focus on the core academic skills (e.g., in mathematics, reading, and writing) they will need in college, and beyond. Perhaps the strongest reason that both the SAT and ACT persist across U.S. colleges and universities is that these tests help college and university personnel make more efficient admissions decisions.

Enter COVID/Enter California

After the COVID outbreak began in spring 2020, however, more than 1,600 (60%) of U.S. colleges and universities (Strauss, 2020) dropped[1] SAT and ACT requirements for admissions (Adams, 2020; Belkin, 2020; FairTest, 2020; McDonnell Nieto del Rio, 2021; Strauss, 2020; Watanabe, 2020a). Many of these institutions decided not to continue the use of the SAT or ACT post-COVID, while others reinstated them, with most rationales justifying such continuations based upon arguments about the predictive validity of both tests (Jaschik, 2022a, 2022b). In California, many colleges and universities, including the University of California (UC) system, which is of primary interest in this study, moved toward eliminating the requirement for both of these standardized tests for all UC admission purposes. More specifically, in May 2020, two months after public and private schools and higher education institutions around the world were closed due to COVID, UC system leaders made the landmark decision to suspend the SAT and ACT requirements for all applicants due to the widespread disruptions caused COVID (e.g., test cancellations, health concerns). One year later, UC system leaders completely removed SAT and ACT requirements from all admissions process, effectively making all UC schools *test-blind*[2], meaning they would not consider these scores even if submitted. This decision was influenced by then even more growing concerns about both tests' fairness, accessibility (i.e., especially during COVID), and roles in perpetuating inequalities across admissions processes.

Around this same time, in July 2020, the case *Smith et al. (Plaintiffs) v. Regents of the University of California, Janet Napolitano* was filed. In this case, hereafter referred to as *Smith v. Regents* (2020; see also Hartocollis, 2019), plaintiffs brought charges against UC Regents that the SATs and ACTs were biased against students of color and students from low-income backgrounds, were not aligned with the needs of California schools and the state of California, were being used to deny applicants equal protection under the California Constitution, were not helping to create diverse student bodies representative of California, and were exclusionary, discriminatory, classist, and racist (Strauss, 2020; see also Atkinson & Geiser, 2009; Hartocollis, 2019, 2021; Jaschik, 2020; Watanabe, 2020a, 2020b). While pre-COVID, UC Regents were moving towards revising their admissions policies and procedures as mentioned, so as also "to gradually reduce and eliminate the role of the SAT and ACT in admissions," UC Regents were planning to do so only if they could develop a different admissions test (e.g., the Smarter Balanced tests) by 2025 (UC Faculty Report, 2020; see also Jaschik, 2020; McDonnell Nieto del Rio, 2021; Watanabe, 2020a, 2020b). UC system leaders have since opted to move away from using the Smarter Balanced tests as part of their college admission criteria, focusing more on high school GPA, coursework, and other factors. Notwithstanding, the allegations presented by the plaintiffs in *Smith v. Regents* (2020) were the focus of this study.

---

[1] We do not differentiate or debate the differences between test-optional policies, whereby college and university applicants get to choose whether they submit their SAT or ACT scores, and "test-free" or "test-blind" policies, whereby colleges and universities may not consider applicants' SAT or ACT scores even if submitted.

[2] See comment above.

Purpose of the Study

More specifically, there were five allegations at issue. First, plaintiffs argued that "over the past six decades," UC leaders commissioned multiple studies "on what, if at all, SAT and ACT scores contribute[d] to the prediction of first-year grades, and [researchers] repeatedly arrived at the same answer: almost nothing" (*Smith v. Regents,* 2020, p. 5). Second, plaintiffs charged that researchers consistently demonstrated that "high school grades are consistently the best predictor of college success – a finding that ha[d] also been established by hundreds of studies at other colleges and universities," and UC System leaders had "not only admitted the 'considerable redundancy' of using both high school grade point average (GPA) and SAT or ACT scores as admissions criteria, but [they] also recognized that high school GPA 'ha[d] less adverse impact on disadvantaged groups" (*Smith v. Regents,* 2020, p. 5). Third, plaintiffs noted that researchers found that more privileged "students [could] be coached, to advantage," given the fact that "affluent students [could more] effectively purchase higher scores through expensive private tutoring services [which] diminish[ed] the already limited predictive value of the tests…further" (*Smith v. Regents,* 2020, p. 5-6). Fourth, they wrote that "rather than measuring individual merit, SAT and ACT results [given they are both norm-referenced tests] artificially compare[d] students against one another in a way designed to produce high and low scores," to repeatedly produce a normal score distribution (i.e., a bell curve) which also tended to iteratively discard test items on which "underrepresented minority students perform[ed] well," creating bias (*Smith v. Regents,* 2020, p. 6). Finally, plaintiffs alleged that all of these factors yielded a "starkly disparate [set of] student outcomes" (p. 6).

Accordingly, the purpose of this systematic literature review was to methodically situate the allegations put forth by plaintiffs within the literature surrounding these allegations from the top, peer-reviewed journals in the field from 1969 to the end of 2019. This study was prompted by the role played by the first of nine authors on this study as an expert witness in this case who had the responsibility to situate the plaintiffs' allegations within the research literature. This work was to, ideally, give both sides in this case a clear understanding of what the research had demonstrated regarding the overall validity of using the SATs and ACTs for college admissions decisions.

Framework for Validity Evidence

The *Standards of Educational and Psychological Testing* (AERA et al., 2014; hereafter referred to as the *Standards*) describe validity as "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (p. 14). Validation is "a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (p. 11). More particularly, Kane (2013) defines a test's validity as per any test's Interpretation/Use Argument (IUA), whereby the IUA framework can be used to justify (e.g., *argue*) the validity of the inferences made *using* test scores by demonstrating that their *interpretations* are supported by the research evidence. Accordingly, the IUA of focus in this study was that: *SAT and ACT tests accurately, consistently, and fairly measure college readiness and can predict college success; thus, SAT and ACT scores support valid college admissions decisions*. Figure 1 provides an overview of the framework we developed and used to help organize the validity evidence in the literature reviewed.

Illustrated in Figure 1 is that the *Standards* (AERA et al., 2014) include five distinct sources of validity evidence which should be examined to support (or challenge) the arguments for (or against) this or any other IUA. These sources of validity evidence, illustrated in the outside circle, include evidence based on test content, response processes, internal structure, relations to other variables, and related consequences. Included at the center of the framework is evidence of reliability/precision and aspects of measurement bias/fairness, currently titled "Reliable, Precise, and

Fair Methods & Measurement" and hereafter referred to as *Methods and Measurement: Reliability/Precision* and *Methods and Measurement: Bias/Fairness*.

Figure 1
Framework for Validity Evidence

For *Methods and Measurement: Reliability/Precision,* AERA et al. (2014) define reliability as the consistency of results over time or across different testing conditions. They define precision as the accuracy and consistency of those results, ensuring that any test measures the intended attribute with minimal error. For *Methods and Measurement: Bias/Fairness,* AERA et al. (2014) define bias as the systematic error in the measurement process that consistently skews results in one direction, leading to inaccurate or unfair conclusions about the attribute being measured. The define fairness, related to bias, given measurement bias threatens fairness and equity and can lead to certain groups of test takers being disadvantaged or inaccurately assessed, undermining validity. Correspondingly, fairness is defined as "the equality of testing outcomes for relevant test-taker subgroups" (AERA et al., 2014, p. 54). Other evidence relating to issues of fairness fall under their appropriate sources of validity evidence (e.g., evidence based on consequences might include differential outcomes for subgroups).

Operationalizing the center circle of Figure 1 in this way (i.e., reliability/precision and bias/fairness) brought the five sources of validity evidence outside the circle and these two measurement concepts in the inside up to seven total measurement concepts on which we focused and used to frame our analyses. The instrument we used to code the literature reviewed included these definitions and additional specifications capturing the procedures we used to operationalize these validity-related measurement concepts, again, as foundational to this study (Ribes-Iñesta, 2003; see also Cronbach, 1989).

Methods

Research Questions

Using this validity framework, we set out to answer the following research questions: (1) To what extent does the validity evidence reported in each of the seven areas of the framework support or challenge using the SATs and ACTs in college and university admissions decisions? (2) To what

extent does the validity evidence presented by scholars published in high-impact, peer-reviewed journals support or challenge the use of the SATs and ACTs in college and university admissions decisions?

Systematic Literature Review

To answer these research questions, we systematically reviewed the literature (Dewey & Drahota, 2016; Gough et al., 2012; Munn et al., 2018) to identify, select, and critically examine the peer-reviewed research and arguments concerning the use of the SATs and ACTs in college and university admissions decisions. We searched for and ultimately included in our analyses both quantitative evidence (e.g., reliability coefficients, internal consistency indicators, predictive validity coefficients) and qualitative evidence (e.g., how authors positioned the evidence derived using their written text, how authors responded to others' research findings in their written text).

Journal Selection and Inclusion Criteria

We searched studies published in top-tier peer-reviewed journals, which we defined as peer-reviewed journals with the highest impact factors as per the *2018 Journal Citation Reports*® (JCR; Clarivate Analytics, 2020); we included journals from the top quartile (i.e., 25%) of each journal's JCR category. We then used the EBSCOhost online database to search for all articles containing any variation of the names of the SAT or ACT (e.g., "SAT," "Scholastic Aptitude Test," "Scholastic Achievement Test") along with mention of college or university admissions.

We selected articles published any time between 1969 to the end of 2019 (i.e., prior to the COVID-pandemic[3]). While a less arbitrary starting point may have been to review articles from 1926 and beyond, when the SAT was first administered to students, or from 1959 and beyond, when the ACT was first administered to students, both tests have remained similar in purpose, function, form(at), and scoring schema from their beginnings[4]. Even though both tests have remained relatively similar over time, however, they have likely changed enough to *not* treat them as similar, or constant, as we did here. We were not able to feasibly address this limitation in this study (see more study limitations forthcoming).

After removing duplicates, our initial search returned 915 records. We then limited the results to articles published only in our qualified journals, which left 151 articles (17% of records returned) to be screened for inclusion in the review (see Figure 2 for the number of articles considered at each stage of the review). Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009), we evaluated the titles and abstracts of each article to determine which ones met an additional set of inclusion criteria: (a) author(s) of the paper discussed or directly investigated the (un)suitability of the SAT or ACT as a basis for college and university admissions decisions, or (b) author(s) of the paper included evidence that supported or challenged the IUA as per our validity framework.

Thereafter, each team member (n=9) independently screened the 151 articles and then met in teams of two to resolve discrepancies and further clarify our inclusion/exclusion criteria. In case of disagreement (38 articles or 25% of those screened) a third team member was involved. We then

---

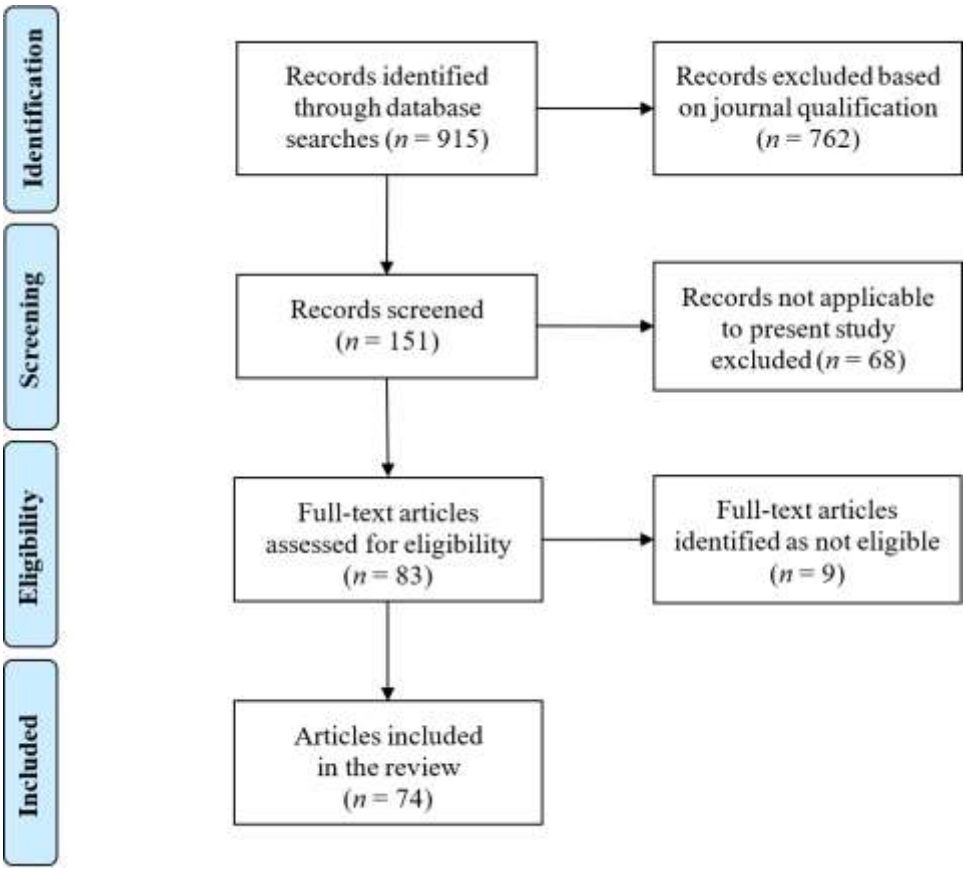
[3] We acknowledge that there are and will likely continue to be plenty of papers published after 2019 that still rely on pre-COVID data. As just one example, see: Goodman, J., Gurantz, O., & Smith, J. (2020). Take two! SAT retaking and college enrollment gaps. *American Economic Journal: Economic Policy, 12*(2), 115-158. https://doi.org/10.3386/w24945

[4] For example, writing tests were added to both tests, score reporting changes were made to the ACT in 2015 and to the SAT in 2016.

identified 83 articles (55% of those screened) for final inclusion. Two of the team members reviewed the full text of each article d reconciled discrepancies using the instrument described next.

Figure 2

PRISMA Flow Diagram



Review Instrument

For our review instrument, we adapted an online instrument from another study during which we developed, tested, validated, and subsequently used the same validity framework (Amrein-Beardsley et al., 2023). In this instrument, we included items to collect information about each of the studies reviewed (e.g., title, author(s), publication year), the extent to which the authors studied each area of validity evidence, and the positions that the author(s) of each article took on the IUA under investigation. See the instrument used for this study in the Appendix.

Detailed in this instrument are the response options we used to capture how the authors of each reviewed publication positioned their evidence as supportive or challenging the IUA. First, we recorded the IUA per study as: a) directly studied, whereby author(s) contributed original validity evidence related to the area; b) mentioned with evidence, whereby author(s) did not contribute new validity evidence to the literature but mentioned the area while citing evidence provided in other papers; c) mentioned, whereby author(s) mentioned the area, but provided no external evidence; or d) not addressed, whereby author(s) did not mention the area of validity evidence. For those areas of the framework in which author(s) *directly studied* or *mentioned the IUA with evidence*, we rated the degree to which the author(s) of the paper positioned the validity evidence as *supportive* or *challenging* of the

IUA on a five-point scale from 5-to-1, with 5=*Supportive* (author[s] of the publication proposed or supported the IUA); 4=S*omewhat Supportive* (author[s] generally supported the IUA, though they may have advised some caution), 3=*Neither supportive nor challenging* (author[s] provided evidence or discussion in support of the IUA as well as evidence or discussion against it), 2=*Somewhat Challenging* (author[s], challenged, or offered caution about the IUA), and 1=*Challenging* (author[s] opposed or argued against the IUA). Our instrument also included text boxes for each area in which we inserted notes, comments, and pasted passages (with page numbers) to collect and retain more qualitative data. Please see these and other items with response options included in our instrument in the Appendix.

Data Analyses

When conducting systematic literature reviews, assessing agreement between reviewers is also crucial for ensuring reliability, validity, and reducing bias and increasing fairness, as defined prior and illustrated at the center of Figure 1. Accordingly, we calculated measures of inter-rater agreement among each of our nine reviewers per article. We also calculated descriptives statistics to illustrate our findings, with our descriptive statistics including means (*M*), standard deviations (*SD*), and coefficients of variation (*CV*).

In terms of our inter-rater levels of agreement, we had 65% initial exact agreement (78% initial agreement within one on all Likert-scale items) as based on all 15 categorical items for the 83 articles we reviewed. During reconciliation, we reached 100% agreement and excluded 11 more articles (12% of the full-text articles read) as *not applicable*, leaving a final list of *n*=74 articles (marked with asterisks in our references).

It was this set of articles (*n*=74) on which we focused for our final analyses. These articles came from education, psychology, and sociology journals, in that order, with occasional entries from journals in other fields (e.g., nursing and health) *if* studies pertained to using the SAT or ACT for college and university admissions purposes. We also included a few studies about using the SAT or ACT for graduate admissions purposes (e.g., medical school).

Study Limitations

Before we proceed, however, it is important to note what we deem as our most important study limitations. First, our inclusion of journals with the highest impact factors likely excluded some articles in which authors used less traditional (or *objective* methods). Second, and related, it is important to underscore that most of the research that we included in our review, especially given the nature of our research questions, was about the psychometric and statistical properties of these tests and their numerical test scores, which naturally led us towards empirical works that were written using more traditional and quantitative research methods (e.g., positivist methods).

Third, and related, again, is that there were also likely articles that we excluded in which other authors may have documented, for example, these tests' unintended consequences in more depth. This is especially important to underscore given *validity evidence based on related consequences* was one of our seven key areas in our validity framework. We may have (and likely) missed some of the more qualitatively oriented studies on this topic simply due to the (unfortunate) fact that sometimes these types of studies with these types of methods are relatively less likely to be published in the journals we selected for inclusion in this study (i.e., peer-reviewed journals with the highest impact factors).

Finally, it is important to note that we treated each study as essentially the same (i.e., beyond the delimiters we used, as mentioned, regarding whether author(s) directly studied any part of our validity framework, whereby author(s) contributed original validity evidence related to the area, and

so on). We did not conduct any deeper analyses of any study's quality, otherwise, trusting that our inclusion criteria (i.e., peer-reviewed journals with the highest impact factors) yielded a similar (and hypothetically solid) indicator of each study's quality prior to the analyses we conducted. It is important to note this is a limitation, regardless, given the event to which one or more relatively more rigorous studies may have been more compelling than one or more relatively less powered or well-designed studies in our sample.

Results

Overall Support or Challenge

On average, author(s) of each of the reviewed papers provided evidence for two or three areas of the framework for validity evidence. Authors of 60 papers (81%) directly investigated at least one area of our validity framework. Authors of 33 papers (45%) provided validity evidence taken and cited from other sources, and author(s) of 19 papers (26%) directly studied and provided evidence from other studies, again, for at least one area of our validity framework. Table 1 includes this information, including additional information about the positions that authors of the studies we reviewed took on the IUA as well as authors' affiliations.

Table 1

Author Positions, Types of Validity Evidence Provided, and Author Affiliations

	n (%)	*M*	*SD*	*CV*
Authors Took a Position on the IUA	74 (100%)	2.9	1.4	.48
Authors advanced an empirical position on the IUA	61 (82%)	2.9	1.6	.55
Authors provided peripheral evidence about the IUA	13 (18%)	*	*	
Authors generally supported the IUA	16 (22%)	4.3	1.0	.23
Authors generally challenged the IUA	15 (20%)	1.3	0.3	.23
Type of Validity Evidence Author(s) Presented				
Authors directly studied areas of the validity framework	60 (81%)	3.1	1.4	.45
Authors provided validity evidence cited from other sources	33 (45%)	2.2	1.1	.50
Authors directly studied and provided evidence from other sources	19 (26%)	2.9	1.4	.48
Authors' Affiliations				
Authors affiliated with the College Board, ACT Inc., or ETS	20 (27%)	3.1	1.3	.42
Other authors	54 (73%)	2.8	1.4	.50

Note. *Indicates that no ratings were provided for evidence which were coded as peripheral.

The overall weight and direction of all validity evidence provided, whether original or cited from other sources, was positioned as neither supportive nor challenging to the valid use of SAT and ACT scores in college and university admissions decisions. Authors who directly studied areas of the framework positioned their evidence as neither supportive nor challenging to the IUA, while authors who cited evidence from other sources positioned it as somewhat challenging. We found a statistically significant ($t(91)=3.07$, $p=.003$) and relatively large ($d=0.67$) difference between the two ratings.

More specifically, authors of 61 of the 74 papers reviewed (82%) advanced a position on the IUA. Authors of the other 13 papers (18%) did not advance a position but provided evidence relevant to its validation and were coded as *peripheral*. On average, authors who took a position on the IUA neither supported nor challenged it, but authors of 16 papers (22%) supported the use of SAT and ACT scores in admissions decisions (i.e., rating=5), and authors of 15 papers (21%) challenged the same practices (i.e., rating=1). Not surprisingly, authors who supported the IUA positioned the validity evidence presented as somewhat supportive to the use of SAT and ACT scores in admissions decisions, while authors who challenged the IUA positioned the validity evidence as challenging the IUA, yielding a statistically significant difference, $t(29)=11.13$, $p<.001$. Interestingly, the literature reviewed included 20 papers (28%) for which the authors or the study itself were supported by or connected to the College Board, ACT Inc., or the Educational Testing Service (ETS), which develops and administers the SAT on behalf of the College Board. No significant differences were observed in how the validity evidence was positioned when comparing the 20 papers with these authors to the rest of the sample written by authors with no apparent connection. Overall, though, our general results suggested that within the body of literature we reviewed, a reader could find evidence to support the IUA under investigation as well as evidence to challenge it.

Validity Evidence by Area

Next, we discuss our findings by each validity area. Table 2 displays mean support ratings, capturing how authors of the reviewed literature positioned the evidence provided across all 74 articles.

Table 2

Number of Articles and Mean Rating of Validity Evidence by Area

Area of Validity Evidence	*n* (%)	*M*	*SD*	*CV*
Overall (any of the areas)	74 (100%)	2.9	1.4	.48
Methods and Measurement	38 (53%)	2.0	1.0	.50
Reliability/Precision	5 (7%)	3.8	0.8	.21
Measurement Bias/Fairness	37 (51%)	1.9	0.9	.47
Test Content	17 (24%)	2.3	1.4	.61
Response Processes	18 (25%)	2.6	1.3	.50
Internal Structure	8 (11%)	3.6	0.9	.25
Relations to Other Variables	58 (81%)	3.2	1.6	.50
Related Consequences	43 (60%)	2.2	1.3	

Note. For each area of the framework for validity evidence, evidence is rated on a scale from 5 to 1, where 5=*Supportive*, 4=S*omewhat Supportive*, 3=*Neither Supportive nor Challenging,* 2=Somewhat Challenging, and 1=*Challenging*

Methods and Measurement

For *Methods and Measurement*, we included evidence pertaining to reliability and precision, as well as random or systemic error evident in the scores. Authors of 38 articles (53%) presented validity evidence that fit in this area, positioning the evidence on average as somewhat challenging to the IUA. Of these 38 papers, authors of nearly all papers (n=37, 97%) presented evidence related to

Measurement Bias/Fairness, and authors of five papers (13%) presented validity evidence on *Reliability/Precision* with four papers containing evidence of both.

Reliability/Precision. Overall, evidence related to the reliability/precision of SAT and ACT test scores was positioned as somewhat supportive for use in college and university admissions decisions. Authors of four articles (Aguinis, 2016; Coyle, 2006; Mattern & Patterson, 2013; Stricker et al., 1993) reported reliability values of the SAT or ACT or their subtests which ranged between .89 and .93. Authors of two articles gave the specific types of reliability estimates reported, one test-retest (Coyle, 2006) and the other internal consistency (Stricker et al., 1993). Otherwise, authors of five articles (Bridgeman & Wendler, 1991; Freedle, 2003; Hanford, 1985; Sackett et al., 2009; Zwick & Sklar, 2005) mentioned the reliability of the SAT or ACT without giving direct evidence or citations. Those authors considered the tests highly reliable, with three mentioning expected values to exceed .90 (Bridgeman & Wendler, 1991; Sackett et al., 2009; Zwick & Sklar, 2005). Our review suggested that there was solid evidence that the reliability of SAT and ACT scores typically exceeded .90, which as per Aguinis et al. (2016), exceeds "recommended reliability standard[s]" (p. 5).

Bias/Fairness. Evidence related to measurement bias and fairness in SAT and ACT test scores was positioned as somewhat challenging regardless of whether the authors directly studied the area or cited evidence from others. We identified two main themes in how the evidence of test fairness was studied. The first theme concerned the roles that test-taker characteristics (e.g., race, ethnicity, gender identity, age, socioeconomic status) play throughout all stages of test development, administration, scoring, interpretations, predictions, and uses for admissions and other decisions (e.g., scholarships). The second theme addressed associations between environmental factors and test-takers' abilities and test scores.

Authors of studies who challenged the IUA reported adverse effects of the SAT or ACT by various student demographics. Aldag and Rose (1983) found a negative correlation between ACT score and age, disfavoring older test-takers; a similar disadvantage was reported for English language learners (ELLs) who took the SAT (Alderman, 1982). Other authors reported that test score disparities existed between males and females (Oswald et al., 2004; Stricker et al., 1993) and students who identified as White or from minoritized racial/ethnic groups (Mercer, 1984; Sternberg, 2006; Zwick, 2007). The alleged cultural bias of the SAT and ACT's scoring techniques also negatively affected African American test-takers as per Freedle (2003; 2004) and Santelices and Wilson (2010). Socio-economic status and parents' education (Atkinson & Geiser, 2009; Crisp et al., 2009; Mercer, 1984), race/ethnicity (Berry et al., 2014; Freedle, 2003; 2004), and gender identity (Breland & Griswold, 1982; Keiser et al., 2016) were also associated with SAT scores, which translated to predictive biases for these groups (Aguinis et al., 2016; Bridgeman & Wendler, 1991; Buchman et al., 2010a; Fischer et al., 2013; Isaacs, 2001; Mattern & Patterson, 2012; Sternberg, 2006; Zwick & Sklar, 2005). Overall, these findings add to concerns about both tests being "metric[s] of rationalized fairness" (Furuta, 2017, p. 249).

On the opposite side, some authors reported that the tests were culturally fair and not biased by students' socio-economic backgrounds, gender identity, or race. This provided support to the validity evidence in favor of the IUA (Dorans, 2004; Fagan & Holland, 2009; Linn, 2009; Schmitt et al., 2009; Wainer & Steinberg, 1992).

Along the second theme – test-takers abilities and environmental factors in relation to test scores - authors addressed the effects of special preparation programs on test score improvements. While authors of two studies reported insignificant gains from coaching programs (Becker, 1990a; DerSimonian & Laird, 1983), other authors evidenced positive relationships between special

preparation and test scores (Alderman & Powers, 1980; Slack & Porter, 1980a, 1980b), which raises concerns about access to resources, training materials, and programs and availability of extra time to all students (Jackson, 2010; Walpole et al., 2005). As resources, such as access to high-speed internet in households and availability of special preparation programs in underserved schools, also play a role in students' opportunities to prepare for and perform on the tests, school contexts might also affect such scores as predictors of college acceptance and success (Espenshade et al., 2005; Walpole et al., 2005).

Overall, the reviewed evidence herein challenged the notion that SAT and ACT scores are a fair measure of ability and merit given test score disparities along student demographics and test-taker characteristics that should be unrelated to tested constructs. Additionally, authors showed that material resources, various forms of capital, and pre-college circumstances are also significant and non-trivial factors contributing to students' test performance on both tests.

Validity Evidence Based on Test Content

Originally positioned as content-related evidence of validity, evidence based on test content is concerned with the relationships between test content and the constructs that tests are purported to measure (Cronbach, 1971; Cureton, 1951; Kane, 2013). Validity evidence here may include analyses of content aligned to the tested domains; comprehensiveness, relevance, or sufficiency of both domains as per desired interpretations; or content-area experts' recommendations and testimonials on constructs and domains (AERA et al., 2014; Kane, 2013).

Initially designed to provide an assessment of test-takers' general aptitude, independent of college preparatory coursework (Atkinson & Geiser, 2009), the SAT was repeatedly subject to critique. Since SAT scores may be significantly improved by coaching, the test may not, in fact, be measuring intrinsic aptitude (Slack and Porter (1980a, 1980b), challenging the IUA on the grounds that the test is neither a fair (since only some students would have access to coaching) nor valid measure (since the test is coachable).

Authors who supported the test content validity of the SAT argued that the test was not meant to be a measure of immutable aptitude, but a measure of cognitive skills learned both in and outside of high school (Jackson, 1980; Messick & Jungeblut, 1981). The SAT was a better measure of writing aptitude than students' GPAs from high school (Breland & Griswold, 1982), and both the SAT and ACT remained the best options for assessing college readiness in the absence of a national curriculum (Linn, 2009; see also Isaacs, 2001).

Authors who did not support test content validity presented evidence which challenged the IUA. For instance, the SAT excluded a variety of factors which loaded onto college readiness, including: (1) time management (Britton & Tesser, 1991); (2) computational skills and quality of workmanship (Bridgeman & Wendler, 1991); (3) a set of "Big Five" (McCrae & Costa Jr., 2008) personality traits (see also Keiser et al., 2016); (4) creativity and pragmatic decision-making (Sternberg, 2006); and (5) abilities to process new knowledge (Fagan & Holland, 2009). Contradicting the IUA, these researchers pointed out that the SAT underestimated college grades for Black and Hispanic test-takers (Berry et al., 2014), and that differential item functioning (DIF) in easier test content led to underestimation of minority test-takers' college readiness (Freedle, 2003, see also Becker, 1990b; Walpole et al., 2005). Finally, Atkinson and Geiser (2009) argued that differing educational priorities between states and universities made the SAT and ACT poor measures of either a high school program of study or the knowledge and skills valued in college readiness.

Overall, authors of 17 papers (24%) provided validity evidence related to the contents of the SAT and ACT, which they positioned as somewhat challenging to the use of these tests in college

admissions decisions. Most of them (*n*=12) challenged the content validity component of the IUA based on (1) both tests' coachability, (2) the incompleteness of both tests in measuring college preparedness, and (3) the unfeasibility of assuming that a single test can address the priorities of states and universities. Others (*n*=5) acknowledged these concerns but considered them overstated, commending the tests.

Validity Evidence Based on Response Processes

Sources of evidence related to *Response Processes* capture how well the construct or performance of any target domain matches the responses or performances measured (AERA et al., 2014). Evidence of such validity may include investigations into how test takers interpret and respond to tests, logical or empirical analyses of matches between response processes and target domains, or studies of raters and graders and the processes they employ (AERA et al., 2014).

Authors of 18 reviewed papers (25%) provided validity evidence related to the response processes for the SATs and ACTs, which they positioned as neither supportive nor challenging to the use of either test for college or university admissions. There were no significant differences reported between studies in which authors directly studied this area or citied evidence from other papers.

The most common theme among articles in which researchers either challenged or remained neutral was about test preparation. Authors noted exposure to test taking skills, such as being exposed to item types, reading skills to help understand or decode problems, and being equipped for pacing, guessing, and using partial information to work backwards through problems as troublesome (Alderman & Powers, 1980; Messick & Jungeblut, 1981; see also Grodsky (2010). Some authors also critiqued using practice sessions and taking tests more than once to increase test and item exposure (Becker, 1990a; DerSimonian & Laird, 1983; Isaacs, 2001; Slack & Porter, 1980a; Walpole et al., 2005). Authors noted differences by student ethnicity in using private and public coaching and training as well (Alon, 2010; see also Jackson, 2010).

Authors who were supportive of the IUA promoted simple drill and practice (Jackson, 1980) and blending content review with using and discussing sample items (Byrnes & Takahira, 1994). Authors also acknowledged that students may experience growth anyway (e.g., via maturation) regardless of coaching or training (Brody & Benbow, 1990; see also Alderman & Powers, 1980).

Another, more traditional theme discussed by authors in the context of response processes included issues of cognitive demand (Aguinis et al., 2016; Bridgeman & Wendler, 1991; Fagan & Holland, 2009), response patterns (Becker, 1990b; Fagan & Holland, 2009; Freedle, 2004), task completion (Becker, 1990b), and the use of cognitive interviews (Freedle, 2004). Authors pointed to gender identity differences in test performance by item and content type (Becker, 1990b) and race (Freedle, 2004). In summary, the reviewed literature offered mixed positions about the SAT and ACT, with emphases on the impacts of coaching or training on SAT and ACT performance.

Validity Evidence Based on Internal Structure

Traditionally, evidence based on internal structure is related to construct validation (Cronbach & Meehl, 1955) and concerns the degree to which scores generated by tests or analytical processes reflect the theories on which tests are based. This includes evidence of how faithfully sub-scales, sub-tests, or other test components yield their hypothesized relationships. Internal structure also includes evidence related to the internal structure of a test, such as the evidence of the relationships among the sub-components of a test, or statistical evidence of tests' dimensional structures. While some types of reliability and DIF also fall within the scope of internal structure, we discussed these finding in our sections on reliability and test fairness.

In our sample, authors of eight papers (11%) provided evidence related to the internal structure of the SATs and ACTs, which they positioned as somewhat supportive of the use of these tests in college and university admissions, regardless of whether the authors directly studied the area or provided evidence from other studies. Authors in this group focused on differences or similarities between performances of the same individuals on different test domains (Fischer et al., 2013; Hanford, 1985; Wasserman, 1978). They reported, for example, that the two components of the SAT—verbal (SAT-V) and mathematical (SAT-M)—provided different information about students' skills (Hanford, 1985), that the correlation between these scores was r=.70 (Wasserman, 1978), and that the SAT-M compared to SAT-V underpredicted college readiness for women (Fisher et al., 2013). Other authors studied individuals whose performance on the SAT contradicted its notion of an ideal test (Freedle, 2004); however, given what the SATs and ACTs are intended to measure, it is unreasonable to expect the tests or their resultant data to conform to such an ideal. Overall, though, authors of the studies in this group did not challenge the validity of the IUA.

Validity Evidence Based on Relations to Other Variables

The relationships of test scores with external variables serve as an important source of validity evidence (AERA et al., 2014), historically discussed as criterion validity (Cronbach, 1971; Kane, 2013; Moss, 1992, 1995). Such evidence may be experimental or correlational; concurrent or convergent, indicating the degree to which two measures of constructs that theoretically should be related are related; discriminant, indicating the degree to which concepts or measurements that are supposed to be unrelated are unrelated; or predictive, indicating the degree to which measurement output can be used to predict other outcomes later in time.

Most of the reviewed studies (n=58, 81%) contained validity evidence in that area, which they positioned as neither supportive nor challenging to using these tests for college and university admissions decisions. Authors of the 46 papers who directly studied this area of validity evidence positioned their evidence as significantly less challenging to the IUA compared to authors of the other 12 papers who positioned the evidence they cited from other sources. The evidence broadly fell into four categories capturing factors external to the test variables – gender identity, coaching and interventions, admissions, and academic performance.

Gender Identity. While the SAT and ACT can help predict success in college, university, and major, there is debate about differential predictions as based on gender identity. While the SATs and ACTs were found to be predictive of major choice in sciences regardless of gender identity (Goldman & Hewitt, 1976), other authors suggested that women were frequently underestimated and tended to perform better in college mathematics than their male peers, countering their test-based predictions (Breland & Griswold, 1982; Bridgeman & Wendler, 1992; Keiser et al., 2016; Stricker et al., 1993). Women also tended to earn better grades when they scored the same on both tests (Fischer et al., 2013). These authors did not support the IUA.

Authors of a number of studies asked important questions about the influence of varying contexts along gender identity lines (Breland & Griswold, 1982; Bridgeman & Wendler, 1991; Keiser et al., 2016; Stricker et al., 1993; Wainer & Steinberg, 1992), adding that women might have underperformed on the SATs and ACTs due to differences in access to role models, family supports, career expectations, institutions which offer supportive environments for women in STEM majors, and the like (Kuchynka et al., 2018; McGee & Bentley, 2017).

Coaching and Interventions. Admission decisions based on SAT and ACT scores motivate students to engage in coaching or training programs to increase their test scores in

relatively short periods of time (Fremer & Chandler, 1971). Authors of the reviewed empirical investigations about such coaching and intervention programs challenged the IUA.

Authors in this group of studies reported positive impacts of Advanced Placement (AP) courses on students' scores, with the most growth found for Black and Hispanic students (Jackson, 2010), ranges of score gains resulting from participation in summer school courses (Brody & Benbow, 1990); and varying growth resulting from coaching interventions and test exposure (e.g., Kaplan and Princeton Review; see, for example, Alderman & Powers, 1980; Slack & Porter, 1980). However, female and Black students were more likely to participate in such programs, and students who participated scored higher and entered more prestigious institutions (Bachmann et al., 2010a).

Admissions. Authors of studies in this group provided mixed evidence supporting and challenging the IUA. While some critiqued the use of the SATs and ACTs in admissions processes due to these tests scores' associations with student demographics (e.g., Sackett et al., 2009), authors of earlier studies found predictive effectiveness or "predictive utility" (Linn, 2009, p. 677; see also Sackett et al., 2009) in both tests (Crouse, 1985a; Crouse & Trusheim, 1991; Schaffner, 1984). Evidence included strong predictions of performance after controlling for socioeconomic background (Sackett et al., 2009; 2012) and significant correlations between college grades and the combination of high school GPAs and SAT scores (Crouse, 1985a; Crouse & Trushiem, 1991), but the predictive value-added was small (Crouse, 1985b).

Authors who challenged the IUA provided evidence of differential predictions and discriminatory effects of SAT and ACT scores, again, on students from racial minority and socioeconomically disadvantaged backgrounds, ranging from empirical findings to philosophical and ethical arguments about college selectivity (Aguinis et al., 2016; Atkinson & Geiser, 2009; Espenshade et al., 2005; Zwick, 2007). Empirical evidence included findings of a one standard deviation difference between race/ethnic minority and white students' scores (Mercer, 1984), as well as an overprediction of college GPAs for Black and Hispanic students and an underprediction for females (Mattern & Patterson, 2013).

Academic Performance. Authors in this group of studies investigated the predictive power of the SATs and ACTs on academic performance measured by but not limited to college GPAs, graduation rates, and graduate educational outcomes. Authors of three studies reported significant correlations between SAT and ACT scores and students' first-semester college GPAs (Schnieder & Overton, 1983; Coyle, 2006; Coyle & Pillow, 2008). Others found that SATs and ACTs better predict first-year college GPAs compared to high school records (Crouse 1985a; 1985b) and that students with the same class rank in high school were more likely to attain a B or C average in college as SAT scores increased (Hanford, 1985). In addition, SAT and ACT scores combined with high school rank significantly predicted four-year cumulative college GPAs for students (Berry & Sackett, 2009; Butler & McCauley, 1987; Stenberg, 2006), and they predicted GPA outcomes better as compared to other non-cognitive data (Schmitt et al., 2009). It is worth noting that other variables (e.g., attendance and absenteeism) also influenced such predictions (Crede et al., 2010; Oswald et al., 2004; see also Fagan & Holland, 2009; Wasserman, 1978).

Other researchers provided evidence which challenged the IUA noting, for example, that high school GPAs were better at predicting college GPA than the SAT (Zwick & Sklar, 2005), that SAT scores added little to the prediction of college grades compared to students' high school records (Slack & Porter, 1980a), and that high school GPAs were better at predicting college graduation rates, especially among White and Hispanic students (Zwick & Sklar, 2005).

Yet, another group of authors yielded statistically significant correlations between SAT scores, final college GPAs, and the same students' Medical College Admissions Test (MCAT) scores

(Hesser et al., 1998; Montague & Frei, 1993; Thurmond & Lewis, 1986). The predictive power of SAT-M scores on these students' performance in medical school was also the same between students from different racial backgrounds, namely from Asian and White backgrounds (Xu et al., 1993). Students' SAT scores over the score of 900 also predicted students' on-time medical school progression for similar students (Edelin & Ugbolue, 2001). Other researchers, citing these same studies, however, cautioned other researchers and practitioners against interpreting these findings as generalizable across populations, especially including ELLs (Alderman, 1982) and older samples of students (Aldag & Rose, 1983).

Validity Evidence Based on Related Consequences

Whether an IUA achieves its intended outcomes and whether it yields any unintended outcomes is another important consideration in reviewing the validity evidence. This evidence is presented next.

Authors of 43 reviewed papers (60%) provided evidence based on the consequences of using SAT and ACT scores in college and university admissions decisions, which they positioned as somewhat challenging to the IUA. Among those studies, authors who directly studied this area of evidence (n=23) positioned the evidence they advanced as less challenging to the IUA than authors of the other 20 papers who cited evidence from other sources.

Authors who discussed intended consequences of the IUA concluded that both tests inform valid admissions decisions as based on their predictability of college performance, with caveats noted. Others pointed out that the use of SAT or ACT can yield different outcomes depending on students' gender identity, age, language, family background, race/ethnicity, etc., potentially altering the chances of being admitted for students from racial minority and socioeconomically disadvantaged backgrounds.

Intended Consequences. While authors of the reviewed studies discussed the actual intents of the SATs and ACTs as they evolved over the past century (Atkinson & Geiser, 2009; Buchman et al., 2010a; R. Jackson, 1980; Slack & Porter, 1980a, 1980b), we considered the consequences that followed our operational definition of the IUA in this section as follows. Of the 13 articles reviewed in which authors directly addressed whether using the SAT or ACT accomplished its intended outcomes, authors of six articles supported the IUA (Berry & Sackett, 2009; Edelin & Ugbolue, 2001; Jackson, 1980; Sackett et al., 2009, 2012; Schaffner, 1985). These authors argued that SAT and ACT scores inform valid admissions decisions. Authors who challenged the IUA (Aguinis et al., 2016; Buchman et al., 2010a; Freedle, 2003; Mattern & Patterson; 2013; Slack & Porter, 1980a; Zwick, 2007) argued that evidence called into question the accurate, consistent, and fair nature of these tests.

Among negative unintended consequences, authors identified the use of SAT or ACT scores as one of the criteria for college admission as a limiting factor for otherwise qualified applicants (Fischer et al., 2013; Mercer, 1984; Schaffner, 1985); this also excluded students from racial minority and socioeconomically disadvantaged backgrounds who generally tend to score lower on such tests (Aguinis et al., 2016; Aldag & Rose, 1983; Alderman, 1982; Atkinson & Geiser, 2009; Mercer, 1984). Additional negative outcomes of using the SATs and ACTs for admissions was the high emotional, psychological, time, and monetary costs of taking both admissions exams. Some argued that such costs coupled with the statistical redundancy of both tests (Crouse, 1985a; Crouse, 1985b) might not justify their value-added for colleges, universities, or test-takers themselves (Slack & Porter, 1980).

In addition, authors pointed out that an overreliance on SAT and ACT scores might prevent capable students from racial minority and socioeconomically disadvantaged backgrounds from entering higher education institutions altogether (Aguinis et al., 2016), particularly at selective

institutions (Posselt et al., 2012; Schmitt et al., 2009). Specifically, authors discussed the admissions of female students (Keiser et al., 2015; Mattern & Petterson, 2013; Wainer & Steinberg, 1992) and racial/ethnic minority students (Long, 2015), including Hispanic, African American, and Native American students (Schmitt et al., 2008; Zwick, 2007). These and other others advanced suggestions regarding using other measures (e.g., entrance essays, GPAs, class ranks) in college and university admissions (Schmitt et al., 2008; see also Hiss & Neupane, 2004; Zwick, 2007).

Ultimately, as per inconsistent prediction patterns of academic and non-cognitive collegiate outcomes for students from racial minority and socioeconomically disadvantaged backgrounds (Aldag & Rose, 1983; Espenshade et al., 2005; Keiser et al., 2015; Oswald et al., 2004, Mattern & Patterson, 2013), authors concluded that students' socioeconomic and language backgrounds (Zwick & Sklar, 2005), as also related to international students (Alderman, 1982, were mechanisms by which both the SAT and ACT had differential consequences in admissions. Students without access to quality curricula aligned with these tests, and students whose families could not afford coaching and test-preparation programs, courses, or special preparatory schools, may have been (and may continue to be) deprived of admissions to their choice colleges and universities as per these tests and the extent to which scores are heavily weighted in admissions decisions (Slack & Porter, 1980b).

Conclusions

As a result of this systematic literature review, we found evidence that both supported and challenged the uses of the SATs and ACTs for college and university admissions decisions. More specifically, there appears to be clear consensus that both are psychometrically strong; that is, they have strong technical and statistical properties. Both are reliable and yield consistent results over time, which is a hallmark of good tests as per the *Standards* (AERA et al., 2014). Further, both tests' subcomponents (e.g., sub-scales, sub-tests, various components of their testing procedures) are strong, yielding accurate inferences about what students know and can do, by subdomain and overall. At the same time, however, in terms of the actual content of both tests, authors of some studies did suggest that the tests' restricted focus on reasoning and core academic domains makes them incomplete measures of college preparedness and, thus, presents a key measurement issue with which users of both tests must contend, especially when either test is weighted more than other admissions measures.

Related to this latter concern are our findings that the predictive values of SAT and ACT scores are the strongest when used alongside high school GPAs, as well as class ranks upon graduation. While both the SATs and ACTs can be used as acceptable or adequate predictors of students' academic success in college, that predictive power is smaller than often assumed, and often smaller than that of high school GPAs. This has important implications for college and universities nationwide.

Perhaps more troubling is that both the SAT and ACT yield differential predictions for sets of dissimilar students. These differential predictions were most often negatively biased against racial minority and socioeconomically disadvantaged students, with varying effects also observed by gender identity, age, and ELL status. Such differential predictions seemingly yield discriminatory effects when using SAT and ACT scores for admissions, especially when capable students from racial minority and socioeconomically disadvantaged backgrounds are competing for admission spots, especially at selective institutions. That the differential score disparities observed exist between test-taker demographics and characteristics that are unrelated to that which both tests are to measure (e.g., aptitude or future success in college) is also problematic.

We also know from this body of research that SAT and ACT scores can yield unfair measures of academic abilities and merit, especially when considering the potential impacts of coaching or training by different groups of students. An additional concern here is the predictive power of the SATs and ACTs on college success given that coaching and training can artificially inflate the indicators of interest, so much as to yield a false indicator of students' abilities and probabilities critical for the long-term success.

There may also be some other unintended, negative consequences at play. These may include but not be limited to students being denied the anticipated use of, perhaps, the high SAT or ACT scores they might yield. Put differently, that these scores may not count for some students (e.g., students defying odds, for example, as based on low GPAs) could also have negative impacts on their access to scholarships, admittance into honors programs, entry into selective institutions, and the like.

All of these findings should be discussed, by us as scholars, but also by the leaders of higher education institutions, especially given the charters, missions, and purposes they have deliberated, embraced, and ultimately put into place as their institutions' motivating and driving forces behind their admissions decisions. Likewise, these leaders must weigh the intended versus unintended consequences underscored as per the literature we reviewed herein to (perhaps more) critically examine their admissions policies, especially as they surround the appropriate and valid uses of the SAT and ACT tests for college and university admissions. Reimagining such admissions, perhaps positioning or valuing them for what they can and cannot do, especially when traditionally marginalized students are being considered for college or university admissions, would be the ultimate end. Although we cannot predict how these tests will continue to work over time given the myriad issues still in play (e.g., current politics throughout the US), this manuscript offers ample and multifaceted evidence about that which should be considered when contemplating the potential or actual uses of either test. Put differently, these research-based and -situated assertions should not go ignored or be overlooked but, rather, serve as a guide for college- and university-level decision makers.

In the End

As for the lawsuit that spurred this study, in August 2020 (one month after *Smith v. Regents* (2020) was filed, plaintiffs filed an injunction given UC Regents still opted to continue using both the SAT and ACT tests post-COVID, "for scholarship and statewide eligibility determinations," which still constituted consequential uses of both tests (*Smith v. Regents,* Preliminary Injunction, 2020, p. 8). This injunction was granted in plaintiffs' favor, forcing UC system leaders to subsequently halt all uses of the SATs and ACTs, "until Defendants [could] demonstrate that the tests [were] equally accessible to all students," including students with disabilities, students for whom access to either test was "impossible or impaired," and students who would be denied equal consideration [across UC] admissions and scholarship policies and processes (*Smith v. Regents*, Preliminary Injunction, 2020, p. 3).

In May 2021, plaintiffs were ultimately declared victorious in an historic ruling in plaintiffs' favor, to the tune of more than $1.2 million, after UC system leaders reached a settlement "to scrap even optional testing from [all UC System] admissions and scholarship decisions" (McDonnell Nieto del Rio, 2021). This settlement made the UC system "the largest and best-known American institution of higher education to distance itself from the use of the[se] two major standardized tests" (McDonnell Nieto del Rio, 2021).

References

Note: References with an "*" indicate the 74 articles researchers included in this review.

Adams, S. (2020). How the SAT failed America. *Forbes*. https://www.forbes.com/sites/susanadams/2020/09/30/the-forbes-investigation-how-the-sat-failed-america/#14d3696753b5

*Aguinis, H., Culpepper, S. A., & Pierce, C.A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology*, *108*(7), 1045-1059. https://doi.org/10.1037/edu0000104

*Aldag, J., & Rose, S. (1983). Relationship of age, American College Testing scores, grade point average, and state board examination scores. *Research in Nursing & Health*, *6*(2), 69-73. https://doi.org/10.1002/nur.4770060206

*Alderman, D. L. (1982). Language proficiency as a moderator variable in testing academic aptitude. *Journal of Educational Psychology*, *74*(4), 580-587. https://doi.org/10.1037/0022-0663.74.4.580

*Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-Verbal scores. *American Educational Research Journal*, *17*(2), 239-251. https://doi.org/10.3102/00028312017002239

*Alon, S. (2010). Racial differences in test preparation strategies. *Social Forces*, *89*(2), 463-474. https://doi.org/10.1353/sof.2010.0053

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

Amrein-Beardsley, A., Lavery, M. R., Holloway, J., Pivovarova, M., & Hahs-Vaughn, D. L. (2023). Evaluating the validity evidence surrounding the use of value-added models to evaluate teachers: A systematic review. *Education Policy Analysis Archives, 31*. https://doi.org/10.14507/epaa.31.8201

*Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665-676. https://doi.org/10.3102/0013189X09351981

*Becker, B. J. (1990a). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, *60*(3), 373-417. https://doi.org/10.3102/00346543060003373

*Becker, B. J. (1990b). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal*, *27*(1), 65-87. https://doi.org/10.3102/00028312027001065

Belkin, D. (2020). College admissions in a Covid year: SATs are out, personal stories are in. *Wall Street Journal*. https://www.wsj.com/articles/college-admissions-in-a-covid-year-sats-are-out-personal-stories-are-in-11600315272?st=ysr94g0v7ehmjyj&reflink=article_email_share

*Berry, C. M., Cullen, M. J., & Meyer, J. M. (2014). Racial/ethnic subgroup differences in cognitive ability test range restriction: Implications for differential validity. *Journal of Applied Psychology*, *99*(1), 21-37. https://doi.org/10.1037/a0034376

*Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science*, *20*(7), 822-830. https://doi.org/10.1111/j.1467-9280.2009.02368.x

*Breland, H. M., & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology*, *74*(5), 713-721. https://doi.org/10.1037/0022-0663.74.5.713

*Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology*, *83*(2), 275-284. https://doi.org/10.1037/0022-0663.83.2.275

*Britton, B. K., & Tesser, A. (1991). Effects of time-management practices on college grades. *Journal of Educational Psychology*, *83*(3), 405-410. https://doi.org/10.1037/0022-0663.83.3.405

*Brody, L. E., & Benbow, C. P. (1990). Effects of high school coursework and time on SAT scores. *Journal of Educational Psychology*, *82*(4), 866-875. https://doi.org/10.1037/0022-0663.82.4.866

Brown v. Board of Education, 347 U.S. 483 (1954). United States Reports (Official Opinions of the U.S. Supreme Court) (36,622).

*Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010a). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces*, *89*(2), 435-461. https://doi.org/http://dx.doi.org/10.1353/sof.2010.0105

*Buchmann, C., Condron, D. J., & Ruscigno, V. J. (2010b). Shadow education: Theory, analysis and future directions. *Social Forces*, *89*(2), 483-490. https://doi.org/10.1353/sof.2010.0073

*Butler, R. P., & McCauley, C. (1987). Extraordinary stability and ordinary predictability of academic success at the United States Military Academy. *Journal of Educational Psychology*, *79*(1), 83-86. https://doi.org/10.1037/0022-0663.79.1.83

*Byrnes, J. P., & Takahira, S. (1994). Why some students perform well and others perform poorly on SAT Math items. *Contemporary Educational Psychology*, *19*(1), 63-78. https://doi.org/10.1006/ceps.1994.1007

Chatterji, M. (2013). Bad tests or bad test use? A case of SAT use to examine why we need stakeholder conversations on validity. *Teachers College Record, 115*(9).

Choi, H. J., & Park, J. (2013). Historical analysis of the policy on the college entrance system in South Korea. *International Education Studies*, 6(11), 106-121. http://dx.doi.org/10.5539/ies.v6n11p106

Clarivate Analytics. (2020). *2019 Journal Citation Reports® Social Sciences Edition.* https://clarivate.com/

*Coyle, T. R. (2006). Test-retest changes on Scholastic Aptitude Tests are not related to *g*. *Intelligence*, *34*(1), 15-27. https://doi.org/10.1016/j.intell.2005.04.001

*Coyle, T. R., Pillow, D. R. (2008). SAT and ACT predict college GPA after removing *g*. *Intelligence*, *36*(6), 719-729. https://doi.org/10.1016/j.intell.2008.05.001

*Crede, M., Roch, S.G., & Kieszczynska, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, *80*(2), 272-295. https://doi.org/10.3102/0034654310362998

*Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An Analysis of students attending a Hispanic serving institution. *American Educational Research Journal*, *46*(4), 924-942. https://doi.org/10.3102/0002831209349460

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147– 171). University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Croft, M., & Beard, J. J. (2021). Development and evolution of the SAT and ACT. In B. E. Clauser, M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (1st ed., pp. 22-41). Routledge.

*Crouse, J. (1985a). Does the SAT help colleges make better selection decisions? *Harvard Educational Review*, *55*, 195-219. https://doi.org/10.17763/haer.55.2.b3q411p04222l175

*Crouse, J. (1985b). This time the College Board is wrong. *Harvard Educational Review*, *55*(4), 478-486. https://doi.org/10.17763/haer.55.4.e21j24l5441418q5

*Crouse, J., & Trusheim, D. (1991). How colleges can correctly determine selection benefits from the SAT. *Harvard Educational Review*, *61*, 125-147. https://doi.org/10.17763/haer.61.2.d6h08n28m54g7434

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed., pp. 621-694). American Council on Education.

*DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, *53*(1), 1-15. https://doi.org/10.17763/haer.53.1.n06j5h5356217648

Dewey, A., & Drahota, A. (2016). *Introduction to systematic reviews* [Online learning module]. Cochrane Training. https://training.cochrane.org/interactivelearning/module-1-introduction-conducting-systematic-reviews

*Dorans, N. J. (2004). Freedle's Table 2: Fact or fiction? *Harvard Educational Review*, *74*(1), 62-72. https://doi.org/10.17763/haer.74.1.8729105044552127

*Edelin, K. C., & Ugbolue, A. (2001). Evaluation of an early medical school selection program for underrepresented minority students. *Academic Medicine*, *76*(10), 1056-1059. https://journals.lww.com/academicmedicine/fulltext/2001/10000/evaluation_of_an_early_medical_school_selection.17.aspx

Edwards, D., Coates, H., & Friedman, T. (2012). A survey of international practice in university admissions testing. *Higher Education Management and Policy*, 24(1), 1-18. https://doi.org/10.1787/hemp-24-5k9bdck3bkr8

*Espenshade, T. J., Hale, L. E., & Chung, C. Y. (2005). The frog pond revisited: High school academic context, class rank, and elite college admission. *Sociology of Education*, *78*(4), 269. https://doi.org/10.1177/003804070507800401

*Fagan, J. F., Holland, C. R. (2009). Culture-fair prediction of academic achievement. *Intelligence*, *37*(1), 62-67. https://doi.org/10.1016/j.intell.2008.07.004

FairTest. (2020). *1,685+ accredited, 4-year colleges & universities with ACT/SAT-optional testing policies for fall, 2021 admissions.* https://www.fairtest.org/university/optional

*Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478-488. https://doi.org/10.1037/a0031956

*Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1-43. https://doi.org/10.17763/haer.73.1.8465k88616hn4757

*Freedle, R. O. (2004). The truth and the truthful sages that spin it: A review of Dorans. *Harvard Educational Review*, *74*(1), 73-79. https://doi.org/10.17763/haer.74.1.8147212057g74635

Fremer, J., & Chandler, M. (1971). Special studies. In W. Agnoff (Ed.), *The College Board admissions testing program* (pp. 147-181). College Entrance Examination Board.

Furuta, J. (2017). Rationalization and student/school personhood in U. S. college admissions: The rise of test-optional policies, 1987 to 2015. *Sociology of Education*, *90*(3), 236-254. https://doi.org/10.1177/0038040717713583

*Goldman, R. D., & Hewitt, B. N. (1976). The Scholastic Aptitude Test 'explains' why college men major in science more often than college women. *Journal of Counseling Psychology*, *23*(1), 50-54. https://doi.org/10.1037/0022-0167.23.1.50

Gough, D. A., Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. SAGE.
*Grodsky, E. (2010). Learning in the shadows and in the light of day. *Social Forces*, *89*(2), 475-481. https://doi.org/10.1353/sof.2010.0063
*Hanford, G. H. (1985). Yes, the SAT does help colleges. *Harvard Educational Review*, *55*(3), 324-331. https://doi.org/10.17763/haer.55.3.g31g503336228721
Hartocollis, A. (2019). University of California is sued over use of SAT and ACT in admissions: A group of students and advocacy groups says the standardized testing requirement is biased and unconstitutional. *New York Times*. https://www.nytimes.com /2019/12/10/us/sat-act-uc-lawsuit.html
Hartocollis, A. (2021). After a year of turmoil, elite universities welcome more diverse freshman classes. *New York Times*. https://www.nytimes.com/2021/04/17/us/minority-acceptance-ivy-league-cornell.html
*Hess, T. G., & Brown, D. R. (1977). Actuarial prediction of performance in a six-year A.B.-M.D. program. *Academic Medicine*, *52*(1), 68-69. https://journals.lww.com/academicmedicine/Fulltext/1977/01000/Actuarial_prediction_of_performance_in_a_six_year.11.aspx
*Hesser, A., Cregler, L. L., & Lewis, L. (1998). Predicting the admission into medical school of African American college students who have participated in summer academic enrichment programs. *Academic Medicine*, *73*(2), 187-191. https://doi.org/10.1097/00001888-199802000-00018
Huysamen, G. K. (1997). Potential ramifications of admissions testing at South African institutions of higher education. *South African Journal of Higher Education, 11*(1). https://hdl.handle.net/10520/AJA10113487_543
*Isaacs, T. (2001). Entry to university in the United States: The role of SATs and Advanced Placement in a competitive sector. *Assessment in Education: Principles, Policy & Practice*, *8*(3), 391-406. https://doi.org/10.1080/09695940120089161
*Jackson, C. K. (2010). A little now for a lot later: A look at a Texas Advanced Placement Incentive Program. *Journal of Human Resources*, *45*(3), 591-639. https://doi.org/10.3368/jhr.45.3.591
*Jackson, R. (1980). The Scholastic Aptitude Test: A response to Slack and Porter's "critical appraisal." *Harvard Educational Review*, *50*(3), 382-391. https://doi.org/10.17763/haer.50.3.f414873706v61867
Jaschik, S. (2020). Dropping the SAT and ACT – for good: University of California plan could change the role of standardized testing in admissions – and not just for the UC System. *Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2020/05/18/university-california-president-proposes-dropping-satact
Jaschik, S. (2022a). MIT reinstates SAT/ACT requirement. *Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2022/04/04/mit-reinstates-satact-requirement
Jaschik, S. (2022b). Will test optional become the 'new normal'? *Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2022/01/24/will-test-optional-become-new-normal-admissions
*Jones, L. V. (1981). Achievement test scores in mathematics and science. *Science*, *213*(4506), 412-416. https://doi.org/10.1126/science.213.4506.412
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000
*Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and

course-taking patterns. *Journal of Applied Psychology*, *101*(4), 569-581. https://doi.org/10.1037/apl0000069

Koljatic, M., Silva, M., & Cofré, R. (2013). Achievement versus aptitude in college admissions: A cautionary note based on evidence from Chile. *International Journal of Educational Development*, 33(1), 106-115. https://doi.org/10.1016/j.ijedudev.2012.03.001

Kuchynka, S. L., Salomon, K., Bosson, J. K., El-Hout, M., Kiebel, E., Cooperman, C., & Toomey, R. (2018). Hostile and benevolent sexism and college women's STEM outcomes. *Pyschology of Women Quarterly*, 42(1), 72-87. https://doi.org/10.1177/0361684317741889

*L., J. F. (1989). SAT Scores. *Pediatrics*, *83*(6), 939-949. https://pediatrics.aappublications.org/content/83/6/939

*Linn, R. L. (2009). Considerations for college admissions testing. *Educational Researcher*, *38*(9), 677-679. https://doi.org/10.3102/0013189X09351982

Liu, H., & Wu, Q. (2006). Consequences of college entrance exams in China and the reform challenges. *KEDI Journal of Educational Policy*, *3*(1), 7-21. file:///Users/ala3171/Dropbox%20(ASU)/Mac/Downloads/Journal_Haifeng%20Liu.pdf

*Long, M. C. (2015). Is there a 'workable' race-neutral alternative to affirmative action in college admissions? *Journal of Policy Analysis and Management*, *34*(1), 162-183. https://doi.org/10.1002/pam.21800

*Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*(1), 134-147. https://doi.org/10.1037/a0030610

McCrae, R. R., & Costa, Jr, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). The Guilford Press.

McDonnell Nieto del Rio, G. (2021). University of California will no longer consider SAT and ACT scores. *The New York Times*. https://www.nytimes.com/2021/05/15/us/SAT-scores-uc-university-of-california.html

McGee, E. O., & Bentley, L. (2017). The trouble success of Black women in Stem. *Cognition and Instruction*, 35(4), 265-289. https://doi.org/10.1080/07370008.2017.1355211

*Mercer, W. A. (1984). Teacher education admission requirements: Alternatives for Black prospective teachers and other minorities. *Journal of Teacher Education*, *35*(1), 26-29. https://doi.org/10.1177/002248718403500108

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

*Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*(2), 191-216. https://doi.org/10.1037/0033-2909.89.2.191

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Annals of Internal Medicine, 151*(4), 264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

*Montague, J. R., & Frei, J. K. (1993). A twelve-year profile of students' SAT scores, GPAs, and MCAT scores from a small university's premedical program. *Academic Medicine*, *68*(4), 306-308. https://doi.org/10.1097/00001888-199304000-00021

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*, 229–258. https://doi.org/10.3102/00346543062003229

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*(2), 5-13. https://doi.org/10.1111/j.1745-3992.1995.tb00854.x

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

*Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187-207. https://doi.org/10.1037/0021-9010.89.2.187

*Posselt, J. R., Jaquette, O., Bielby, R., & Bastedo, M. N. (2012). Access without equity: Longitudinal analyses of institutional stratification by race and ethnicity, 1972–2004. *American Educational Research Journal*, *49*(6), 1074-1111. https://doi.org/10.3102/0002831212439456

Ribes-Iñesta, E. (2003). What is defined in operational definitions? The case of operant psychology. *Behavior and Philosophy, 31*, 111-126.

*Royer, J. M., Abranovic, W. A., & Sinatra, G. M. (1987). Using entering reading comprehension performance as a predictor of performance in college classes. *Journal of Educational Psychology*, *79*(1), 19-26. https://doi.org/10.1037/0022-0663.79.1.19

*Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin*, *135*(1), 1-22. https://doi.org/10.1037/a0013978

*Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., Shen, W., & Kiger, T. B. (2012). The role of socioeconomic status in SAT-grade relationships and in college admissions decisions. *Psychological Science*, *23*(9), 1000-1007. https://doi.org/10.1177/0956797612438732

*Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardized approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106-133. https://doi.org/10.17763/haer.80.1.j94675w001329270

*Schaffner, P. E. (1984). Competitive admission practices when the SAT is optional. *Journal of Higher Education*, *55*(6), 55-72. https://doi.org/10.1080/00221546.1985.11778704

*Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., &Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*(6), 1479-1497. https://doi.org/10.1037/a0016810

*Schneider, L. J., & Overton, T. D. (1983). Holland personality types and academic achievement. *Journal of Counseling Psychology*, *30*(2), 287-289. https://doi.org/10.1037/0022-0167.30.2.287

*Schult, J., & Sparfeldt, J. R. (2016). Do non-*g* factors of cognitive ability tests align with specific academic achievements? A combined bifactor modeling approach. *Intelligence*, *59*, 96-102. https://doi.org/10.1016/j.intell.2016.08.004

*Sesnowitz, M., Bernhardt, K. L., & Knain, D. M. (1982). An analysis of the impact of commercial test preparation courses on SAT scores. *American Educational Research Journal*, *19*(3), 429-441. https://doi.org/10.3102/00028312019003429

*Shewach, O. R., Shen, W., Sackett, P. R., & Kuncel, N. R. (2017). Differential prediction in the use of the SAT and high school grades in predicting college performance: Joint effects of race and language. *Educational Measurement: Issues and Practice, 36*(3), 46-57. https://doi.org/10.1111/emip.12150

*Slack, W. V., & Porter, D. (1980a). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review*, *50*(2), 154-175. https://doi.org/10.17763/haer.50.2.d755627708652757

*Slack, W. V., & Porter, D. (1980b). Training, validity, and the issue of aptitude: A reply to Jackson. *Harvard Educational Review*, *50*(3), 392-401.

Smith et al. v. Regents of the University of California, Janet Napolitano. (2020). No. RG19046222. Superior Court of the State of California. Alameda County.

Smith et al. v. Regents of the University of California, Janet Napolitan. (2020). Preliminary Injunction filed August 20, 2020. No. RG19046222. Superior Court of the State of California. Alameda County.

*Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, *34*(4), 321-350. https://doi.org/10.1016/j.intell.2006.01.002

Strauss, V. (2020). It looks like the beginning of the end of America's obsession with student standardized tests. *The Washington Post*. https://www.washingtonpost.com/education/2020/06/21/it-looks-like-beginning-end-americas-obsession-with-student-standardized-tests/

*Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from scholastic aptitude test scores. *Journal of Educational Psychology*, *85*(4), 710-718. https://doi.org/10.1037/0022-0663.85.4.710

*Thurmond, V. B., & Lewis, L. (1986). Correlations between SAT scores and MCAT scores of Black students in a summer program. *Journal of Medical Education*, *61*(8), 640-643. https://journals.lww.com/academicmedicine/Abstract/1986/08000/Correlations_between_SAT_scores_and_MCAT_scores_of.2.aspx

University of California (UC) Faculty Report. (2020). *Report of the UC Academic Council Standardized Testing Task Force (STTF).* University of California Systemwide Academic Senate, University of California.

*Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review*, *62*(3), 323-336. https://doi.org/10.17763/haer.62.3.1p1555011301r133

*Walpole, M., McDonough, P. M., Bauer, C. J., Gibson, C., Kanyi, K., & Toliver, R. (2005). This test is unfair: Urban African American and Latino high school students' perceptions of standardized college admission tests. *Urban Education*, *40*(3), 321-349. https://doi.org/10.1177/0042085905274536

*Wasserman, M. (1978). An evaluation of a compensatory introductory sociology section. *Journal of Experimental Education*, *47*(2), 162-171. https://doi.org/10.1080/00220973.1978.11011676

Watanabe, T. (2020a). UC experts offer new ammunition against the SAT and ACT as an admissions requirement. *Los Angeles Times*. https://www.latimes.com/california/story/2020-04-22/uc-admissions-experts-drop-ammunition-against-sat-act-requirement

Watanabe, T. (2020b). UC must immediately drop use of the SAT and ACT for admissions and scholarships, judge rules. *LA Times*. https://www.latimes.com/california/story/2020-09-01/uc-may-not-use-the-sat-or-act-for-admissions-scholarship-decisions-for-now-judge-rules

*Xu, G., Veloski, J. J., Hojat, M., Gonnella, J. S., & Bacharach, B. (1993). Longitudinal comparison of the academic performances of Asian-American and White medical students. *Academic Medicine*, *68*(1), 82-86. https://doi.org/10.1097/00001888-199301000-00013

Yamamoto, B. A. (2016). Diversifying admissions through top-down entrance examination reform in Japanese elite universities: What is happening on the ground? In A. Mountford-Zimdars & N. Harrison (Eds.), *Access to higher education: Theoretical perspectives and contemporary challenges.* Routledge. https://doi.org/10.4324/9781315684574

*Zwick, R. (2007). College admissions in twenty-first-century America: The role of grades, tests, and games of chance. *Harvard Educational Review*, *77*(4), 419-428. https://doi.org/10.17763/haer.77.4.u67n84589527t80v

*Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, *42*(3), 439-464. https://doi.org/10.3102/00028312042003439

About the Authors

Audrey Amrein-Beardsley
Arizona State University
Email: audrey.beardsley@asu.edu
ORCID: 0000-0002-1250-2281
Audrey Amrein-Beardsley, PhD., is a professor in the Mary Lou Fulton College of Teaching and Learning Innovation at Arizona State University. Dr. Amrein-Beardsley's research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

Zarrina T. Azizova
University of North Dakota
Email: zarrina.azizova@und.edu
ORCID: 0000-0001-9237-0970
Zarrina Talan Azizova, PhD., is an associate professor in the College of Education and Human Development at University of North Dakota. Dr. Azizova uses qualitative and quantitative research methods and social theory for her research centering on the issues of postsecondary access and student success to advance equity-conscious educational practices for historically marginalized students.

Norman P. Gibbs
Mesa Unified School District
Email: npgibbs@mpsaz.org
ORCID: 0000-0002-5064-2816
Norman P. Gibbs, Ph.D., is **a** program evaluator for the Mesa Unified School District, Mesa, Arizona. Dr. Gibbs' research has centered on assessment and accountability, comparative and international education, and inclusive and participatory decision-making.

Chukwuemeka (Emeka) Ikegwuonu
St. Cloud State University.
Email: emeka.ikegwuonu@stcloudstate.edu

ORCID: 0000-0002-1215-4600
Chukwuemeka (Emeka) Ikegwuonu, Ph.D., is an assistant professor in the College of Education and Learning Design at St. Cloud State University. Dr. Ikegwuonu's research focuses on how organizational decisions at higher education institutions shape student behaviors.

Jeongeun Kim
University of Maryland, College Park
Email: jkim0217@umd.edu
ORCID: 0000-0002-3736-1446
Jeongeun Kim, Ph.D., is an associate professor of higher education at the University of Maryland, College Park. Dr. Kim's research focuses on how institutions of higher education use their autonomy to organize policies and practices to remain competitive and the consequences of these strategies on students and faculty in terms of access, affordability, and quality.

Deborah M. La Torre
University of California, Los Angeles
Email: dlatorre@ucla.edu
ORCID: 0009-0000-4333-8712
Deborah La Torre is completing her Ph.D. in social research methodology in the School of Education and Information Studies at UCLA. As an associate research scientist, her research focuses on STEM learning and assessment, cognitive complexity, educational measurement, and afterschool education.

Matthew R. Lavery
South Carolina Education Oversight Committee.
Email: lavery.matthew.r@gmail.com
ORCID: 0000-0002-4208-7277
Matthew R. Lavery, Ph.D., is Director of Research for the South Carolina Education Oversight Committee. Dr. Lavery's research focuses on the valid use of educational assessments to inform instructional decisions and improve student outcomes.

Margarita Pivovarova
Arizona State University
Email: margarita.pivovarova@asu.edu
ORCID: 0000-0002-2965-7423
Margarita Pivovarova, Ph.D., is an associate professor in the Mary Lou Fulton College of Teaching and Learning Innovation at Arizona State University. Dr. Pivovarova explores policy-relevant education problems including teacher retention, student achievement, and school performance.

Yi Zheng
Arizona State University
Email: Yi.Isabel.Zheng@asu.edu
ORCID: 0000-0003-2671-0820
Yi Zheng, Ph.D., is an associate professor in the Mary Lou Fulton College of Teaching and Learning Innovation at Arizona State University. Dr. Zheng is a psychometrician who studies how educational and psychological measurement instruments (e.g., tests, scales, assessments) are best designed and validated.

education policy analysis archives

Volume 33 Number 28 April 22, 2025 ISSN 1068-2341

Appendix

Validation Article Review Instrument

Introduction

Based on the *Standards for Educational and Psychological Testing*, developed by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) and hereafter referred to as the *Standards* (AERA et al., 2014), the work of Kane (2013) and Messick (1995), and other scholars in the field of validity, this instrument is designed to support article reviews with an eye toward issues of validity in general, and specifically toward the validation of particular interpretations and uses of SAT and ACT test scores.

The sections and items in this instrument follow the framework shown in Figure 1. Simply stated, each of the wedges represent different sources of evidence that either support or challenge the specific interpretation/use argument (IUA; see Kane, 2013) being validated. The IUA relies on the quality of the scores that inform it; thus, the five sources of evidence pictured support the IUA through the measurement and methodology issues on which the IUA rests.

Figure 1

Framework for Validity Evidence

Reviewer

Which member of the research team is completing this entry?

- ○ Reviewer 1 (Blinded for peer review)
- ○ Reviewer 2 (Blinded for peer review)
- ○ Reviewer 3 (Blinded for peer review)
- ○ Reviewer 4 (Blinded for peer review)
- ○ Reviewer 5 (Blinded for peer review)
- ○ Reviewer 6 (Blinded for peer review)
- ○ Reviewer 7 (Blinded for peer review)
- ○ Reviewer 8 (Blinded for peer review)

First Author

Copy/paste or type the last, first name of the sole or first author here.

Potential Conflict of Interest (COI)

Is any contributing author affiliated with the SAT/College Board or the ACT?

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

Year of Publication

Copy/paste or type the four-digit year of the publication here.

Full Title

Copy/paste or type the full title of the publication here.

Type of Publication

What type of publication is the reviewed piece?

- ○ Report of Empirical Research (1)
- ○ Systematic Review or Meta-Analysis (2)
- ○ Program Evaluation Report (3)
- ○ Methodological, Theoretical, or Conceptual Piece (4)
- ○ Other (type publication classification below) (5)

Purpose of the Study

What is the original purpose of the study reviewed (which may or may not be directly related to the IUA)?

This should be 1) copied/pasted from the abstract, 2) copied/pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

Methodological Approach

What general methods were used in the study reviewed?

This should be 1) copied/pasted from the abstract, 2) copied/pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

General Findings

What are the findings reported in the paper (which may or may not pertain to the IUA under investigation, but should pertain to the purpose listed in 0.6)?

This should be 1) copied/pasted from the abstract, 2) copied/pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

1.0 Interpretation/Use Argument (IUA)

"To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores. Validation therefore requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores. Public claims require public justification" (Kane, 2013, p.1). The Interpretation/Use Argument (IUA) "includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (Kane, 2013, p. 2). "The kinds of evidence required for validation are determined by the claims being made, and more-ambitious claims require more evidence than less-ambitious claims" (Kane, 2013, p. 3). "More-ambitious interpretations (e.g., a construct interpretation or a causal claim) tend to be more useful than less-ambitious interpretations, but they are harder to validate" (Kane, 2013, p. 37).

Via this review we will address how the reviewed article contributes to the following IUA:

SAT and ACT tests accurately, consistently, and fairly measure college readiness and can predict college success; thus, SAT and ACT scores support valid college admissions decisions.

- College readiness is defined as having mastered the prerequisite knowledge and skills to participate in traditional college coursework without remediation.
- College success is defined as persistence (i.e., students return the semester following their first semester), retention (i.e., students return the fall following their first year), program completion (i.e., students complete all requirements for their degree program), and degree attainment (e.g., graduation).
- The SAT and ACT scores that inform the IUA are defined as those that are used (or to be used) to predict a student's college readiness and aptitude for college success, collectively and particularly in the following subject areas: English/reading and mathematics, as well as science, writing, history, and foreign languages, which are often optional.
- Valid college admissions decisions are defined as follows: For first-time, freshmen, seeking entrance into the University of California (UC) System, as well as other selective institutions, indicators pursued for admissions emphasize goals pertaining to:
    - Selectivity (e.g., keeping admission and matriculation rates lower/higher than previous years, as weighted at least in part by lower/higher high school grade point averages (GPAs), SAT/ACT scores, and the like); and
    - Equity & Access (e.g., enrollment percentages of students by first-generation status, race/ethnicity, Pell-grant recipiency, out-of-state or international status). If UC (and other selective institutions) try to balance and appropriately weigh admission indicators around these two goals, for purposes of this study we define a "valid decision" to be based on one's demonstrated "academic excellence" as balanced by applicants' relevant background factors such as those listed prior.
        - Note that the IUA concerns the validity of individual admissions and/or

scholarship decisions; decisions that are made at the level of the applicant. Selectivity and diversity, as concerns at the institutional level, are therefore most likely to be represented in the Validity Evidence Based on Related Consequences area of this instrument. Equity and access are terms most often used to reflect the applicant-level side of the diversity coin. For example, when an individual student from a traditionally underrepresented group is inappropriately or unfairly denied admission, it is described as an equity and access problem but influences the institution's diversity (or lack thereof). Similarly, college readiness and college success (defined above) are terms that describe the applicant-level of the selectivity coin. When students who are college ready and likely to be successful in college are inappropriately or unfairly denied admission, it may preserve the institution's selectivity at the expense of the individual qualified to attend. Because the commonly accepted definition of selectivity includes consideration of SAT/ACT scores, selectivity must only be considered in the Validity Evidence Based on Related Consequences area of this instrument.

- More specifically, as per UC Regents Policy 2102 (see here):
  - The undergraduate admissions policy of the UC system is guided by the UC system's commitment to serve the people of California and the needs of the state.
  - The entrance requirements established by the UC system requires that the top one- eighth of the state's high school graduates, as well as those transfer students who have successfully completed specified college work, be eligible for admission to the UC system. These requirements are designed to ensure that all eligible students are adequately prepared for university-level work.
  - Mindful of its mission as a public institution, the UC system has an historic commitment to provide places within each UC campus for all eligible applicants who are residents of California. The UC system seeks to enroll, on each of its campuses, a student body that, beyond meeting eligibility requirements, demonstrates high academic achievement or exceptional personal talent, and that encompasses the broad diversity of cultural, racial, geographic, and socioeconomic backgrounds characteristic of California.
  - Because applicant pools differ among the campuses of the UC system, each campus shall establish procedures for the selection of applicants to be admitted from its pool of eligible candidates. Such procedures shall be consistent with the principles stated above and with other applicable UC policies.
    - Note that it may be helpful to complete the items in this section last.

1.1 Position on IUA being Validated

What kind of position do the author(s) of the publication reviewed take on the IUA above?

Note: This is the author(s)' take on the IUA as revealed through the publication reviewed, not the reviewer's assessment of how well the publication supports the IUA in question.

- ○ (5) SUPPORTIVE (The publication proposes or supports the IUA) (1)
- ○ (4) SOMEWHAT SUPPORTIVE (The publication is generally supportive of the IUA, though it may advise some caution) (2)
- ○ (3) NEITHER SUPPORTIVE NOR CHALLENGING (The publication provides

evidence or discussion in support of the IUA as well as evidence or discussion which
challenges it) (3)

- (2) SOMEWHAT CHALENGING (The publication critiques, challenges, or offers caution about the IUA while acknowledging its appropriate use) (4)
- (1) CHALENGING (The publication challenges, opposes, or argues against the IUA) (5)
- (88) PERIPHERAL (The investigated IUA is not a central aspect of the publication, but the publication contains evidence relevant to the review) (6)
- (99) NOT APPLICABLE (The article reviewed provides no relevant evidence related to the IUA being investigated. Specifically, none of the numbered areas below are addressed) (7)

1.2 IUA-Related Findings

What (if any) findings are reported in the paper that directly pertain to the IUA under investigation?

This should be 1) copied/pasted from the abstract, 2) copied/pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

1.3 IUA-Related Recommendations

What (if any) specific recommendations do the authors make in the paper that directly pertain to the IUA under investigation?

These may be recommendations to use SAT/ACT scores in admissions decisions, recommendations to cease doing so, or cautions to consider. This should be 1) copied/pasted from the abstract, 2) copied/pasted from another part of the article, or 3) written in simple terms by the reviewer if not stated briefly by the author(s).

1.X IUA Notes

Use this space for notes and comments about how the publication reviewed addresses the IUA under investigation that are not captured by any other section in this review instrument.

Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

2.0 Reliability/Precision and Fairness

2.1.0 Reliability/Precision and Stability

The authors of the *Standards* (AERA et al., 2014) point out that the measurement literature has historically used the term "reliability/precision to denote the more general notion of consistency of the scores across instances of the testing procedure" (p. 33). Though discussed "as an independent characteristic of test scores, . . . reliability/precision has implications for validity" (AERA et al., 2014, p. 34), also as disaggregated, whereby "to the extent feasible (i.e., if samples sizes are large enough), reliability/precisions should be estimated separately for all relevant subgroups (e.g., defined in terms of race/ethnicity, gender, language proficiency) in the population" (AERA et al., 2014, p. 37; see also Standard 2.3).

Specifically, evidence that SAT/ACT scores are stable (or unstable) over repeated administrations address this area of the framework. Cronbach's alpha values are measures of internal consistency and are only relevant here if reported for the entire test as an indicator of reliability. Cronbach's alpha values reported for subtests (such as mathematics, reading, science, etc.) are more relevant to area 5, validity evidence based on internal structure.

2.1.1 Addressed

To what extent is reliability/precision or stability addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

2.1.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)

- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)
- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

2.1.X Notes

Use this space for notes and comments related to reliability/precision or stability. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

2.2.0 Fairness

The *Standards* (AERA et al., 2014) discusses fairness as "a fundamental issue in protecting test takers and test users in all aspects of testing" and that "fairness is a fundamental validity issue and requires attention throughout all stages of test development and use" (p. 49). In its glossary, the *Standards* define fairness as follows: The validity of test score interpretations for intended use(s) for individuals from all relevant subgroups. A test that is fair minimizes construct- irrelevant variance associated with individual characteristics and testing contexts that would otherwise compromise the validity of scores for some individuals. (AERA et al., 2014, p.219)

Construct irrelevant variance (CIV) is a term used by Messick (1989) to describe factors that falsely inflate or deflate the measurement of a variable and therefore distort its interpretation, or distort its validity (see also AERA et al., 2014, p. 12). CIV is pertinent here in that the presence of CIV "systematically lowers or raises scores for identifiable groups of test takers and results in inappropriate score interpretations," as often based on students' opportunities to learn that "can influence the fair and valid interpretations of test scores for their intended users" (AERA et al., 2014, p. 54). Authors of the *Standards* (AERA et al., 2014) discuss several aspects of fairness, many of which will be captured in other areas of this instrument. *Standards* authors discuss measurement bias as a major threat for fairness, along with issues of accessibility and universal design.

Note: The operational distinction that we make for the purpose of this study is that evidence has been gathered as part of a reviewed study for the explicit purpose of determining whether the test performs similarly for various subgroups of test takers, some of whom might be considered vulnerable, disadvantaged, or traditionally underrepresented in colleges and universities (e.g., first-generation college students, students with disabilities, students from low SES backgrounds, students for whom the English language is not their first language or their home language) would

be captured in section 2.2, Fairness. In these cases, the author(s) will typically use words like "fairness," "equity," "access," "discrimination," or "social justice" in their narrative. If evidence that could be captured in another area of this instrument is collected and analyzed without reference to such issues, then it should be recorded in that portion of the instrument.

2.2.1 Addressed

To what extent is fairness addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

2.2.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)
- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)
- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

2.2.X Notes

Use this space for notes and comments related to fairness. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

3.0 Validity Evidence Based on Test Content

Evidence based on test content is primarily concerned with the relationship between the content of the tests and the constructs the tests are purported to measure (Cronbach, 1971; Cureton, 1951; Kane, 2013). "Content-oriented evidence of validation is at the heart of the process in the educational arena known as alignment, which involves evaluating the correspondence between student learning standards and test content" (AERA et al., 2014, p. 15). Accordingly, validity evidence of this type may include but not be limited to analyses of the content domain as aligned to the tested domain, discussions of the relevance or sufficiency of the tested content domains as pertinent to the proposed interpretations, or recommendations and testimonials taken from content-area experts (AERA et al., 2014; Kane, 2013).

Note: As a validation review, evidence may only be included in this section of the instrument if the author(s) of the paper being reviewed addresses the alignment of the SAT or ACT test scores to effectively measure college readiness and college aptitude as defined prior; that is, the IUA of interest herein. Furthermore, this area of evidence will only be "directly studied" or "mentioned with evidence" if it relies on more than just logical argument but presents empirical or external evidence to support the argument.

3.1 Addressed

To what extent is content-related evidence of validity addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

3.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)
- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)

- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

3.X Notes

Use this space for notes and comments related to content related evidence of validity. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

4.0 Validity Evidence Based on Response Processes

The Response Processes source of evidence captures how well the construct or performance of the target domain matches the responses or performances measured in order to inform the IUA (AERA et al., 2014). When the responses or performances measured by a test are substantially different from that of the target domain, irrelevant method variance may limit the validity of proposed interpretations of the test (Messick, 1989, 1995). Evidence in this area may include but not be limited to investigations into how test takers interpret and respond to the test used, logical or empirical analyses of the match between response processes and the target domain, or studies of raters and graders and the processes that they employ (AERA et al., 2014). In the Response Processes area, we considered Kane's (2013) suggestion that, "if the [measured] tasks seem to involve the same processes as most tasks in the target domain, extrapolation is likely to seem reasonable" (p. 28). Specifically, if the proposed interpretations and uses of SAT and ACT scores are college admissions decisions, then the knowledge, skills, and processes that are necessary to succeed in college should be the same knowledge skills and processes that are necessary to succeed on these tests.

Note: Evidence related to the "coachability" of the SAT and ACT might belong in this source of validity evidence. If SAT/ACT preparation were only to involve reviewing and mastering the content measured by the test, then it would be relevant to area 3 above, validity evidence based on test content. If SAT/ACT preparation activities involve anything other than academic content (e.g., test-taking strategies, time management during the test), then it likely belongs here.

4.1 Addressed

To what extent is response process related evidence addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area

of evidence) (4)

4.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- (5) Evidence SUPPORTS the IUA under investigation (1)
- (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- (1) Evidence CHALLENGES the IUA under investigation (5)
- (99) (NA) Not Applicable (0 or 1 selected above) (6)

4.X Notes

Use this space for notes and comments related to response processes. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

5.0 Validity Evidence Based on Internal Structure

In the historical validity literature, evidence based on internal structure is related to construct validation (see Cronbach & Meehl, 1955) and concerns the degree to which the test that supports the IUA reflects the theory on which it is based. This includes evidence of how faithfully the sub-scales, sub-tests, or various components of the testing procedure follow their hypothesized relationships. "Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16).

Typically, evidence based on internal structure is psychometric in nature and concerns the degree to which the test that supports the IUA reflects the theoretical construct on which it is based. This can include evidence of how faithfully the factor analyses, sub-scales, sub-tests, or various components of the testing procedure follow their hypothesized relationships. Given that the SAT and ACT both produce subtest scores as well as overall scores, this area of validity evidence might include examinations of those subtests and/or their relationships to the other subtests.

5.1 Addressed

To what extent is internal structure addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

5.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)
- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)
- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

5.X Notes

Use this space for notes and comments related to response processes. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

6.0 Validity Evidence Based on Relations to Other Variables (e.g., convergent, concurrent, discriminant, or predictive)

In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and, as a result, analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (AERA et al., 2014, p. 16). Relationships between the measurement of interest and other variables (e.g.,

performance criteria) have been historically discussed as a primary aspect of criterion validity (see Cronbach, 1971; Kane, 2013; Moss, 1992, 1995). Such evidence may can be experimental or correlational; concurrent or convergent, demonstrating the degree to which two measures of constructs that theoretically should be related are, in fact, related; discriminant, demonstrating the degree to which concepts or measurements that are supposed to be unrelated are, in fact, unrelated; or predictive, concerning the degree to which measurement output can be used to predict other outcomes, which are typically assessed or observed at some later point in time. In all cases, researchers typically seek evidence related to other variables to demonstrate that the IUA is consistent with conclusions based on other related measures, observations, or outcomes. The extent to which such evidence can be generalized to new situations or multiple types of test takers is in large measure a function of the accumulated research.

6.1 Addressed

To what extent are relationships with other variables addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

6.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)
- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)
- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

6.3 Types

If DIRECTLY STUDIED or MENTIONED WITH EVIDENCE, what types of evidence based on relationships to other variables did the publication reviewed provide (select all that apply)?

- ☐ CONCURRENT/CONVERGENT EVIDENCE - The degree to which measures of the same or similar constructs are related (1)
- ☐ DISCRIMINANT EVIDENCE - The degree to which measures of different constructs are unrelated (2)
- ☐ PREDICTIVE EVIDENCE - The degree to which a test or observation accurately predicts future performance or future outcomes (these future criteria might not be tests per se, but might be outcomes such as persistence, retention, four-year graduation, etc.) (3)

6.X Notes

Use this space for notes and comments related to relationships with other variables. For each piece of evidence based on relations to other variables, be sure to include (a) the variable for which a relationship with SAT/ACT scores was tested, (b) the reported coefficient (or range of coefficients), (c) the significance or non-significance of the relationship, and (d) the effect size as reported/interpreted by the author(s). Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.0 Validity Evidence Based on Related Consequences

If the inferences and decisions that the scores support must be validated, rather than the scores themselves, then it is appropriate to also evaluate the intended and unintended consequences of those score uses as part of the validity argument. Whether the IUA produces its intended outcomes is an important consideration, as is whether the IUA yields unintended outcomes (that can be positive or negative) as well as, or in lieu of, those which are intended. As per AERA et al. (2014) Standard 1.25: "When unintended consequences result from test use, an attempt should [also] be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct. (p. 30). After all, Kane (2013) writes that "a decision rule that achieves its goals at an acceptable cost and with acceptable consequences is considered a success. A decision rule that does not achieve its goals or has unacceptable consequences is considered a failure" (p. 47). Note, however, that not all consequences will be immediately apparent and that, particularly in the case of unintended consequences, a particular IUA might be in use for some time before the full extent of its consequences are known and understood. Subsequently, validity evidence based on related consequences is the element of the framework most likely to reopen a scholarly discussion that might have previously been seen as settled or resolved. Appropriately, it is often observed that "the validation process never ends" (AERA et al., 2014, p. 21).

To help reviewers calculate the balance of intended and unintended consequences reported by the literature reviewed, the three categories provided by Kane (2013) will be split into four separate notes fields: 7.A) INTENDED OUTCOMES, 7.B) POSITIVE SYSTEMIC EFFECTS

(unintended), 7.C) NEGATIVE SYSTEMIC EFFECTS (unintended), 7.D) DIFFERENTIAL EFFECTS / ADVERSE IMPACT (unintended)

7.1 Addressed

To what extent are intended and/or unintended consequences of the IUA under investigation addressed within the publication reviewed?

- ○ (3) DIRECTLY STUDIED (evidence gathered/reported as part of the publication) (1)
- ○ (2) MENTIONED WITH EVIDENCE (e.g., cites other papers) (2)
- ○ (1) MENTIONED (no citations or evidence provided; mark "NA" below and skip to the next numbered area of evidence) (3)
- ○ (0) NOT ADDRESSED (mark "NA" below and skip to the next numbered area of evidence) (4)

7.2 Evidence

If DIRECTLY STUDIED *or* MENTIONED WITH EVIDENCE, does the publication provide evidence that supports or challenges the IUA under investigation?

- ○ (5) Evidence SUPPORTS the IUA under investigation (1)
- ○ (4) Evidence SOMEWHAT SUPPORTS the IUA under investigation (2)
- ○ (3) Evidence NEITHER SUPPORTS NOR CHALLENGES the IUA under investigation (3)
- ○ (2) Evidence SOMEWHAT CHALLENGES the IUA under investigation (4)
- ○ (1) Evidence CHALLENGES the IUA under investigation (5)
- ○ (99) (NA) Not Applicable (0 or 1 selected above) (6)

7.A INTENDED OUTCOMES

Use this space for notes and comments related to the intended consequences of the IUA under investigation. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.B POSITIVE UNINTENDED EFFECTS (applies to system(s) or population(s))

Use this space for notes and comments related to positive unintended consequences of the IUA under investigation that affect everyone. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.C NEGATIVE UNINTENDED EFFECTS (applies to system(s) or population(s))

Use this space for notes and comments related to negative unintended consequences of the IUA under investigation that affect everyone. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

7.D DIFFERENTIAL EFFECTS / ADVERSE IMPACT (applies to individuals or individuals by subpopulation(s))

Use this space for notes and comments related to unintended consequences of the IUA under investigation that affect certain sub-groups or populations differently than others. Remember to include quotes and page numbers for any text copied and pasted from the article reviewed.

8.0 Other Validity Notes

"A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA et al., 2014, p. 21).

8.1 Overall Rating

According to the **reviewer**, and based upon the readings and discussions of the research team, taken in its entirety, how strong is the evidence of validity provided by this publication for the IUA being investigated in this review?

- ○ (5) SUPPORTIVE (1)

- ○ (4) SOMEWHAT SUPPORTIVE (2)
- ○ (3) NEITHER SUPPORTIVE NOR CHALLENGING (3)
- ○ (2) SOMEWHAT CHALLENGING (4)
- ○ (1) CHALLENGING (5)

8.A Author's Concerns

What concerns or limitations do the author(s) raise about the IUA that have not already been captured in another part of this form?

8.B Reviewer's Concerns

What concerns or limitations does the reviewer see that have not already been captured by another part of this form?

8.X Other Notes

Provide other thoughts, comments, and notes here about the publication in its entirety or about issues not addressed elsewhere in this form.

0.X Star Rating

On a five-star scale, how would you rate the overall value of this publication for inclusion on a list of recommended reading for those interested in the use of SAT and ACT tests in college admissions decisions?

- ○ (5) Five Stars (1)
- ○ (4) Four Stars (2)

- ○ (3) Three Stars (3)
- ○ (2) Two Stars (4)
- ○ (1) One Star (5)