DOCUMENT SUMMARY This comprehensive review investigates the scientific validity of 364 psychological assessment tools used in legal contexts and finds them severely lacking. The study reveals that while most tests have some empirical testing, only 40% have generally favorable psychometric reviews, and a tool's "general acceptance" in the field has only a weak relationship to its actual scientific quality. Critically, the legal system fails as a gatekeeper: this "junk science" is rarely challenged in court (only 5.1% of cases), and when it is, the challenges fail two-thirds of the time, providing powerful evidence for Enlitens' argument that the traditional testing system is broken.

FILENAME
Neal2019_RESEARCH_research_article_CritiqueOfStandardizedTesting_LegalAdmissibility

METADATA **Primary Category:** RESEARCH **Document Type:** research_article **Relevance:** Core **Update Frequency:** Static **Tags:** #StandardizedTesting, #Psychometrics, #JunkScience, #Daubert, #LegalAdmissibility, #Validity, #Reliability, #AssessmentCritique, #MMY, #SystemicFailure **Related Docs:**

Why This Matters to Enlitens

This paper is the single most important piece of evidence we can use to dismantle the credibility of the standardized testing industry. It is a systematic, data-driven indictment that proves our core assertions: many widely used psychological tests are scientifically weak ("junk science"), their popularity is not an indicator of quality, and the legal system that is supposed to vet them completely fails to do so. This research provides the statistical "ammunition" for our whitepaper, marketing materials, and conversations with skeptical parents and professionals. It allows us to state, with backing from a major scientific publication, that the system is broken not just in theory, but in documented practice. The paper's call for high-quality, nonproprietary measures is a direct endorsement of the need for initiatives like the Enlitens Interview.

Critical Statistics for Our Work

- **Widespread Use of Weak Tools:** Of 364 common forensic assessment tools, only about 40% have generally favorable reviews of their psychometric properties in authoritative sources.
- **Many Tools Have Negative Reviews:** Of the tools that have been professionally reviewed, 37% have "mixed" reviews and 23% have "generally unfavorable" reviews.
- **Many Tools Have No Reviews:** 37% of the 364 tools used in legal settings have no professional reviews available in the most comprehensive sources.
- **"General Acceptance" is Meaningless:** The study found only a weak relationship (Cramer's V = 0.17) between a tool being "generally accepted" by clinicians and the actual favorability of its technical properties.
- **Legal System Fails as Gatekeeper:** Legal challenges to assessment evidence occurred in only 5.1% of cases studied.
- **Challenges Rarely Succeed:** When challenges to assessment tools were raised, they only succeeded in having the evidence excluded about one-third (32%) of the time.

- **The Worst Tools Get a Pass:** The most scientifically suspect tools (those with unfavorable reviews that are not generally accepted) were **never challenged** in the 3-year case law sample.
- **Faulty Forensic Science:** Faulty forensic-science evidence has been involved in up to 45% of known false conviction cases.

Arguments Against Standardized Testing

- **Profit-Driven Industry:** Psychological testing is a big business with million- and billion-dollar companies that look to maximize profit, which may not align with producing psychometrically sound tests.
- **Commercial Publication is Not a Quality Guarantee:** Some commercial tests are published without ever having been subjected to scientifically sound testing or surviving the scientific peer-review process. Even when peer-reviewed articles exist, they may only be reviews of the information provided by the publisher in the commercial manual, not independent testing.
- **Lack of Training:** Most graduates of doctoral programs in psychology lack fundamental competencies in measurement science, and practitioners have limited knowledge of psychometric criteria for tests.
- **Invalid Test Batteries:** Clinicians often use a battery of multiple tests, but the specific *combination* of tests in the battery has likely never been validated, making interpretations speculative hypotheses at best.
- **Lawyers and Judges are Ineffective Gatekeepers:** Lawyers are not trained to analyze test validity, and judges struggle to apply legal standards like *Daubert*, routinely admitting evidence with poor scientific foundations.

Quotes We Might Use

- **On the overall findings:** "We find that many of the assessment tools used by psychologists and admitted into legal contexts as scientific evidence actually have poor or unknown scientific foundations. We also find few legal challenges to the admission of this evidence."
- **On legal system failure:** "Attorneys rarely challenge psychological expert assessment evidence, and when they do, judges often fail to exercise the scrutiny required by law."
- **On the big business of testing:** "Psychological testing is a big-business industry, and test publishers—some of them million- and billion-dollar companies publicly traded on the stock exchange—look to maximize profit."
- **On the disconnect between popularity and quality:** "Furthermore, there is a weak relationship between general acceptance and favorability of tools' psychometric properties."
- **On the failure to challenge the worst tests:** "Challenges to the most scientifically suspect tools are almost nonexistent."
- **On the need for non-commercial alternatives:** "We suggest nonproprietary measures with strong scientific underpinnings be prioritized over commercially developed tools that have not been independently tested, especially in criminal cases, given due process concerns..."

FORMATTED CONTENT

Psychological Assessments in Legal Contexts: Are Courts Keeping "Junk Science" Out of the Courtroom?

Abstract

In this article, we report the results of a two-part investigation of psychological assessments by psychologists in legal contexts. The first part involves a systematic review of the 364 psychological assessment tools psychologists report having used in legal cases across 22 surveys of experienced forensic mental health practitioners, focusing on legal standards and scientific and psychometric theory. The second part is a legal analysis of admissibility challenges with regard to psychological assessments. Results from the first part reveal that, consistent with their roots in psychological science, nearly all of the assessment tools used by psychologists and offered as expert evidence in legal settings have been subjected to empirical testing (90%). However, we were able to clearly identify only about 67% as generally accepted in the field and only about 40% have generally favorable reviews of their psychometric and technical properties in authorities such as the Mental Measurements Yearbook. Furthermore, there is a weak relationship between general acceptance and favorability of tools' psychometric properties. Results from the second part show that legal challenges to the admission of this evidence are infrequent: Legal challenges to the assessment evidence for any reason occurred in only 5.1% of cases in the sample (a little more than half of these involved challenges to validity). When challenges were raised, they succeeded only about a third of the time. Challenges to the most scientifically suspect tools are almost nonexistent. Attorneys rarely challenge psychological expert assessment evidence, and when they do, judges often fail to exercise the scrutiny required by law.

Introduction

Psychological tests, tools, and instruments play an increasingly significant role in determining the outcome of legal cases. One might think that, given the stakes involved, the validity of such tests would always be carefully examined. That is not, however, always what happens. Virtually all jurisdictions charge judges with the responsibility of evaluating the admissibility of expert evidence. This gatekeeping function is designed to separate the wheat from the chaff, thus ensuring that only reliable and valid expert-opinion testimony is allowed as evidence. Yet judges frequently have trouble evaluating the scientific merits of various expert methods, and major investigations have revealed that courts routinely admit evidence with poor or unknown scientific foundations. Up to 45% of known cases of false conviction involve faulty forensic-science evidence.

Psychological testing is a big-business industry, and test publishers—some of them million- and billion-dollar companies publicly traded on the stock exchange—look to maximize profit. The public and the courts might assume that psychological tests published, marketed, and sold by reputable publishers are psychometrically strong tests. But not all psychological tests have good technical quality, and the psychometric properties of other tests are unknown. In their systematic review of all 283 psychological assessment test entries in the Sixteenth Mental Measurements Yearbook, Cizek and colleagues (2012) found that 59.5% of the educational and psychological tests were evaluated as either unfavorable, mixed, or neutral by professional reviewers.

Part I: A Systematic Analysis of Psychological Assessment Tools Used in Court

Across 22 surveys, 364 distinct psychological assessment tools were identified as having been used by or acceptable for use by clinicians in forensic settings. We used this set of 364 tools as the basis for the current investigation.

Results

- **Evidence of Testing:** Most of the tools (n=326 of 364, 90%) have been subjected to testing.
- **General Acceptance:** We found insufficient evidence to make a judgment about general acceptance for about half of the tools (n=185 of 364, 51%). Of those for which we found evidence, we were able to clearly identify about two thirds as generally accepted (n=119 of 179, 67%).
- **Professional Reviews of Quality:** For 37% of the tools (n=136 of 364), no professional reviews are available in the comprehensive review sources. Of those for which reviews are available, only 40% have generally favorable reviews (n=91 of 228 with reviews). Nearly the same percentage (n=84 of 228; 37%) have mixed reviews in these professional review sources, and the remaining 23% have generally unfavorable reviews (n=53 of 228).
- **Relationship Between Acceptance and Quality:** The relationship between general acceptance and overall quality was statistically significant, but weak in strength, Cramer's V=0.17, p<.001. Thus, although there appears to be a positive association between the degree to which a tool is generally accepted in the field and the favorability of the tool's technical properties, the relation does not appear as strong as one might expect.

Discussion Among the more concerning findings are that only about 67% of the tools used by clinicians in forensic settings could clearly be identified as generally accepted, and only about 40% received generally favorable reviews in authorities such as the MMY. The relationship between overall quality and general acceptance was weak.

Some psychological assessment tools are published commercially without participating in or surviving the scientific peer-review process and/or without ever having been subjected to scientifically sound testing—core criteria the law uses for determining whether evidence is admissible. On closer examination, peer-reviewed publications about psychometric properties may turn out to be reviews solely of the information published in the commercial manual. Thus, some tools appear to have survived scientific peer review but... may not in fact have been subject to the same level of scrutiny that is accorded psychometric results from independent testing.

Another issue involves structured professional judgments (SPJ). SPJs intentionally eschew traditional psychometrics and norm-based interpretations. Consequently, SPJ tools are challenging to evaluate both from traditional psychometric theories and under the Daubert criteria.

We did not study unaided clinical judgment. About 25% of psychologists providing clinical expert testimony in court continue to rely on unstructured evaluations. The weight of the evidence indicates that structured approaches using psychological assessment tools are more valid and reliable than unaided clinical judgment.

A common practice not evaluated is the use of test batteries. Many psychologists use multiple psychological tests to inform their conclusions. Two psychologists may administer entirely different test batteries to the same examinee, a problem compounded by the likelihood that any given battery has not been subjected to testing in the configuration used by the clinician. As the Standards indicate, it is understood that little, if any, literature exists that describes the validity of interpretations of scores from highly customized or flexible batteries of tests.

Part II: Case-Law Analysis: Are Courts Scrutinizing Psychological-Assessment Evidence?

Hypothesis: Our hypothesis for this part of the study was that psychological tools and assessments are rarely challenged or scrutinized in court (even when they should be).

Method: To get a sense of how often 30 exemplar tools were discussed and challenged by the courts, we searched Westlaw for all judicial opinions and orders from all states and federal courts during 2016, 2017, and 2018. A total of 372 cases were analyzed.

Results

- **Frequency of Legal Challenges:** Out of 372 cases in which more than a mere mention of one of the 30 exemplar tools occurred, only 19 involved a challenge to a tool's admissibility or the admissibility of testimony relying on the tool (5.1%).
- **Success of Challenges:** On those few occasions when challenges did take place, they often failed. Only 6 of the 19 cases challenged (32%) succeeded (that is, the psychological assessment evidence was ruled as inadmissible and excluded from evidence 32% of the time challenges were raised).
- **Relationship Between Validity and Challenges:** Our evidence shows little relation between a tool's psychometric quality and its likelihood of being challenged. Of the three tools fitting the description of the "worst" (i.e., those that received unfavorable reviews and are not generally accepted), none was challenged. The SIT/SFRIT was cited 15 times, but never challenged. Those data suggest that some of the weakest tools tend to get a pass from the courts.
- **Bottom Line:** Our bottom-line conclusion is that evidentiary challenges to psychological tools are rare and challenges to the most scientifically suspect tools are even rarer or are nonexistent.

Discussion Lawyers are generally not trained in how to analyze the validity of a psychological tool. Rather, they are likely to defer to what experts tell them. If they are not alerted to the weaknesses of a tool, lawyers are unlikely to raise a challenge. Experts may not mention issues about a tool's weaknesses to lawyers because they are unaware of them or because they use the tool themselves and prefer not to be challenged.

General Discussion

The purpose of this project is to help lawyers and courts to see psychological assessment evidence as challengeable; to help psychologists see where their assessments are weak and how to select stronger tools for use in high-stakes decisions; and to inspire researchers to help bolster science where needed. We find that many of the assessment tools used by psychologists and admitted into legal contexts as scientific evidence actually have poor or unknown scientific foundations. We also find few legal challenges to the admission of this

evidence. Attorneys rarely challenge the expert evidence and, when they do, judges tend not to subject psychological assessment evidence to the legal scrutiny required by the law.

We hope these findings motivate the public—as well as professionals in the field—to be more critical of psychological-assessment evidence, and to go to primary sources for the most up-to-date information about individual tools as the literature evolves.

A key implication of our findings... is that, even when administered appropriately, tests produce scores that are valid only for specific purposes. Psychologists must recognize that a given measure may be valid for use in some settings yet not in others.

Scientists and practitioners interested in improving the state of the science and state of practice in forensic psychological assessments might work to create a free, online database similar to the PhenX Toolkit, but specific to forensic mental health. If the field moves toward open materials (nonproprietary, and perhaps noncommercial), it could pave the way for better connections between research and practice, and a more cumulative science.