aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
2019, Vol. 20(3) 135–164

DOI: 10.1177/1529100619888860
www.psychologicalscience.org/PSPI
SAGE

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

Tess M. S. Neal[1], Christopher Slobogin[2], Michael J. Saks[3,4], David L. Faigman[5], and Kurt F. Geisinger[6,7]

[1]School of Social and Behavioral Sciences, Arizona State University; [2]Law School, Vanderbilt University; [3]Sandra Day O'Connor College of Law, Arizona State University; [4]Department of Psychology, Arizona State University; [5]Hastings College of the Law, University of California; [6]Buros Center for Testing, University of Nebraska–Lincoln; and [7]College of Education and Human Sciences, University of Nebraska–Lincoln

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.

Keywords

psycholog*, assessment, evaluation, forensic, psychometric, *Daubert*, expert, law, legal

Psychological tests, tools, and instruments play an increasingly significant role in determining the outcome of legal cases. Personality measures of various types are often proffered in cases judging parental fitness for custody purposes (e.g., *Lefkowitz v. Ackerman*, 2017) or in termination of parental rights (e.g., *In re Dayana J*, 2016). Psychological tools are used to determine the causes of legally relevant mental health symptoms such as posttraumatic stress (e.g., *Tardif v. City of New York*, 2018), often arising from military service in combat zones (e.g., *Eichenberger v. Shulkin*, 2017). Various tests measuring intellectual and social functioning may determine the outcome of Social Security and other disability proceedings (e.g., *Hurskin v. Commissioner of Social Security*, 2016). Tests for measuring malingering, or faking, can also be important in such proceedings (e.g., *Cannon v. Commissioner of Social Security*, 2018), as well as in cases in which a criminal defendant is asserting incompetence to stand trial or insanity at the time of a crime (e.g., *People v. Jing Hua Wu*, 2016). Judges may rely on risk assessment instruments in deciding whether an offender should go to prison and, if so, for

Corresponding Author:
Tess M. S. Neal, School of Social & Behavioral Sciences, Arizona State University, 4701 W. Thunderbird Rd., Mail Code 3051, Glendale, AZ 85306
E-mail: tess.neal@asu.edu

how long (e.g., *Wisconsin v. Loomis*, 2016), or whether a convicted criminal is to remain incarcerated despite completing a prison sentence (e.g., *In the Matter of Kristek*, 2016). Most dramatically, intelligence tests have become all but dispositive in determining whether a person should be sentenced to death under the Supreme Court's case law exempting people with intellectual disability from the death penalty (*Moore v. Texas*, 2017).

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 (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993; Fed. R. Evid. 401, 2019; Fed. R. Evid. 702, 2019; *Frye v. U.S.*, 1923). 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 (e.g., National Research Council, 2009; Saks & Koehler, 2005). When poor science is not recognized as such and is used to reach legal decisions, the risk of error rises and the legitimacy of the legal system is threatened (Bell et al., 2018). Consider, for example, the global crisis of confidence about scientific evidence that has erupted in response to damning reports about the scientific validity of many forensic-science techniques (Bell et al., 2018; National Research Council, 2009; President's Council of Advisors on Science and Technology, 2016). There are real-world consequences of poor validity in forensic-science techniques: Up to 45% of known cases of false conviction involve faulty forensic-science evidence (Innocence Project, n.d.).

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. Companies such as Pearson Clinical, Psychological Assessment Resources, Stoelting, Western Psychological Services, Pro-Ed, and Multi-Health Systems sell thousands of psychological assessment tools, many of which are revised and republished with updated versions over time. Many of these tools are sold for hundreds of dollars, and usually there are recurring per-use costs for items such as answer sheets, record forms, or online administration and scoring programs.

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 (e.g., Carlson, Geisinger, & Jonson, 2017; Cizek, Koons, & Rosenberg, 2012), 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 (Spies & Plake, 2005), 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. Likewise, although noting that data on the issue is limited, Slaney (2017) suggested that many tests currently in use have not been sufficiently validated.

Establishing the scientific foundations of psychological assessments used in high-stakes contexts such as legal proceedings is particularly critical. Although the U.S. Supreme Court has attended to the psychometric properties of psychological assessment tools in some settings (see, e.g., *Hall v. Florida*, 2014, focusing on statistics regarding the standard error of measurement in intelligence tests), some of the psychological assessment tools and methods of expert judgment admitted into court might not be admitted if subjected to serious scrutiny (Neal, 2018; Otto & Heilbrun, 2002). The time is ripe for an investigation of the scientific status of psychological assessments used in legal contexts. This project advances that investigation by evaluating both the scientific basis of psychological assessment tools and the courts' evaluation of them.

Psychological Measurements

Sir Frances Galton, the father of modern psychometrics, pioneered efforts to measure physical, psychophysical, and mental abilities in his London Anthropometric Laboratory (Wasserman & Bracken, 2013). Galton quantified everything from fingerprint characteristics and weather patterns to audience boredom in scientific meetings, as measured by fidgets and yawns per minute. Derived from the Greek *psyche* (soul) and *metro* (measure), Galton defined *psychometry* as the "art of imposing measurement and number upon operations of the mind" (Galton, 1879, p. 149). On the basis of these foundations, modern psychometric theories have evolved as a set of scientific rules for creating and measuring the usefulness of psychological tests.

Psychometric theories and considerations

The two leading psychometric theories today are *classical test theory* and *item response theory*. Scholars contrast them, but most test developers use elements from both approaches (Nunnally & Bernstein, 1994). Classical test theory, pioneered by Galton, Pearson, Spearman, and Thorndike (see Gulliksen, 1950), shaped psychological-test development through the second half of the twentieth century and evolved into generalizability theory (or G-theory; Cronbach, Gleser, Nanda,

& Rajaratnam, 1972; Vispoel, Morris, & Kilinc, 2018). Item response theory, pioneered by Rasch (1960) and Lord and Novick (1968), is influential and is evolving with new rules of measurement (Embretson, 1995).

Regardless of the underlying theory, two of the core psychometric concepts are validity (accuracy) and reliability (repeatability). Validity has been consistently endorsed as the major prerequisite to the psychometric health of a tool (Borsboom, Mellenbergh, & van Heerden, 2004; Clark & Watson, 1995, 2019; Cronbach & Meehl, 1955; Loevinger, 1957). Validation is an ongoing effort consisting of collecting, analyzing, and synthesizing various sources of evidence about how a particular tool performs in different sets of circumstances (Kane, 2013; Messick, 1989; M. E. Strauss & Smith, 2009). Substantive inferences can follow from a psychological measurement only when (a) good theory supports the items included in the scale, (b) the measure has acceptable psychometric properties (e.g., reliability and dimensionality), and (c) the measure relates as hypothesized to other constructs in ways that capture group differences or causal processes as expected (e.g., convergent and discriminant validity; Borsboom, 2006; Borsboom et al., 2004; Clark & Watson, 2019; Flake & Fried, 2019).

Although validity theorists have paid attention to test score *interpretation* for many decades, more recently the field has increasingly attended to how test scores are *used* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014; Kane, 2013; Messick, 1995). Validity is thus considered a property of the proposed interpretations and uses of test scores, rather than a property of the test itself (e.g., Kane, 2013; Messick, 1995). Related to this focus on use is the controversial notion of consequential validity (see Messick, 1995). Messick argued that the content- and criterion-related aspects of validity cannot be separated from the social consequences of test-score use. Whether or not the consequences of test use are labeled as an aspect of test validity, the concern about the social consequences of how tests are used is highly relevant for the current project.

A related point is that reliability and validity are context specific. Clinicians and courts need to make judgments about the scientific acceptability of a tool on a case-by-case basis, weighing the known psychometric qualities of a tool against the demands of the specific case. For example, a tool might have good reliability and validity for measuring intelligence in children but might not be appropriate for measuring an adult's intelligence or measuring other constructs, such as competence to stand trial. In addition, tools that are known to have good psychometric properties with some populations may not be appropriate for use on populations who were not part of the norming samples. For instance, influential cases in Australia and Canada ruled inadmissible the use of risk-assessment tools to evaluate indigenous people who were not part of the tools' validation sample (*Director of Public Prosecutions for Western Australia v. Mangolamara*, 2007; *Ewert v. Canada*, 2018). Furthermore, even if a tool has demonstrated good psychometric properties in the lab, the tool's field reliability and validity in real-world settings might be significantly lower than in controlled settings, where most tools are tested. Several recent articles call into question the utility of these lab-based validity and reliability indices for real-world forensic settings (see, e.g., the 18-article *Psychological Assessment* Special Issue on the Field Utility of Forensic Assessment Instruments and Procedures; Edens & Boccaccini, 2017).

Analyzing the quality of psychological measurements

In evaluating the validity and reliability of psychological tools, several authoritative sources are useful. First, manuals designed by most tool developers typically provide normative data and statistical information demonstrating the psychometric properties of the tool. Peer-reviewed articles published in the primary literature provide psychometric results of tools tested in particular contexts and/or updates of norms. Additional sources that aid in critically evaluating tools include well-regarded standards for what makes a good psychological assessment tool, as well as comprehensive review sources that compile psychometric information about a particular tool from manuals, peer-reviewed articles, and other sources (e.g., websites and unpublished studies) and provide professional reviews and evaluations of tools. Sources relied on in this project are reviewed next.

The Standards. The *Standards for Educational and Psychological Testing* (hereinafter, *Standards*) have provided guidance about appropriate test development and criteria for evaluating tests for more than half a century (AERA, APA, & NCME, 2014; APA, 1954). The result of a long-standing collaboration among three associations—the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education—the *Standards* provide criteria designed to "promote sound testing practices and to provide a basis for evaluating the quality of those practices" (AERA, APA, & NCME, 2014, p. 1). The U.S. Supreme Court has relied on the *Standards* as an authoritative source for answers to technical and psychometric questions about psychological tests (see, e.g., Lerner, 1971).

The *Standards* include sections devoted to foundational issues, including validity, reliability, and fairness

in testing. In the validity chapter, for example, Standard 1.0 requires "clear articulation of each intended test score interpretation for a specified use . . . and appropriate validity evidence in support of each intended interpretation" (p. 23), and then provides 25 additional standards unpacking this overall requirement. In the chapter on reliability/precision and errors of measurement, Standard 2.0 requires "appropriate evidence of reliability/precision . . . for the interpretation for each intended score use" (p. 42), followed by 20 additional reliability-related standards detailing specific requirements. The *Standards* also include criteria for operational issues, such as test design and development; scores, scales, norms, score linking, and cut scores; test administration, scoring, reporting, and interpretation; supporting documentation for tests; and rights and responsibilities of test takers and users.

Comprehensive review sources. The *Mental Measurements Yearbook* (*MMY*) has provided information about the technical quality of psychological tests for more than 80 years (e.g., Buros, 1938; Carlson et al., 2017). It was founded by Oscar Buros, who sought to protect test users and hold test publishers accountable for their claims concerning their products. The *MMY* is considered the most accurate, complete, and authoritative source of information about published psychological tests (Cizek et al., 2012; Plake, Conoley, Kramer, & Murphy, 1991). The American Psychological Association's website highlights the *MMY*, and describes it (and its companion resource, *Tests in Print*), as "two of the most useful and popular reference series" available for finding information about particular psychological tests (APA, 2019). Furthermore, the *MMY* has been recommended as a primary source of authority for psychological tests used in legal proceedings (e.g., Heilbrun, 1992, 1995).

The *MMY* provides timely, consumer-oriented reviews of psychological tests with critically evaluative information for informing test selection. It is updated about every 3 years. Most tests have two independent professional reviews of the test's strengths and weaknesses, as well as reviewer references. To be reviewed in the *MMY*, a test must be published in English, must be commercially available, and must have information available about its technical quality. To be reviewed, test developers request a review and submit various materials and technical documentation for scrutiny, and this information is fact-checked carefully before publication in the *MMY*. Reviews are updated if a test is substantially revised, appears in a new edition, or publishes updated norms.

More than 10,000 tests have been reviewed by the *MMY* since its first edition, with a searchable version of all its published test reviews available through online subscription. Today, almost everyone who accesses the *MMY*—including professional psychologists and members of the public investigating the status of tests—does so online. Subscriptions to these electronic databases are offered to individuals and institutions, primarily libraries, across the globe (J. Carlson, personal communication, 28 June 2019). In addition, the Buros Center has a website called *Test Reviews Online* through which people can purchase reviews of individual tests from the *MMY* for a nominal fee (currently $15.00; https://marketplace.unl.edu/buros/).

Two other comprehensive review sources were used in this project: E. Strauss, Sherman, and Spreen's (2006) *Compendium of Neuropsychological Tests* and Grisso's (2003) *Evaluating Competencies: Forensic Assessments and Instruments*. Strauss and colleagues' compendium provides information (background, norms, reliability, validity, utility) and critical reviews of major neuropsychological tools to aid practitioners' and scholars' evaluation of the psychometric properties of various tools in clinical contexts. Grisso's text provides an evaluative review of 37 specialized forensic instruments designed to assess various legal competencies. As with the other sources, it provides background and psychometric information about each tool, along with a critical review of each, to aid the appropriate selection of and critical evaluation of test use.

Legal Standards for the Admission of Expert Testimony

Legislation and case law define the conditions under which evidence, including expert testimony, is admissible in court. To be admissible, all evidence must be relevant. Rule 401 of the Federal Rules of Evidence (Fed. R. Evid. 401, 2019) defines evidence as relevant if it "(a) has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action."

Expert evidence, including evidence relying on psychological tests, must meet additional admissibility criteria. Under Rule 702 of the Federal Rules (Fed. R. Evid. 702, 2019) and the rules of almost every state, such evidence must "assist the trier of fact." In addition, Rule 702 states that

> a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify . . . if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

This language is meant to implement the U.S. Supreme Court's decisions in a landmark trilogy of cases: *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993; hereinafter, *Daubert*), *General Electric Co. v. Joiner* (1997; hereinafter, *Joiner*), and *Kumho Tire Co. v. Carmichael* (1999; hereinafter, *Kumho Tire*). Numerous states have adopted, to varying degrees, the same rules, or parts of them.

In both the federal rules and in the *Daubert* trilogy, the word "reliable" is meant to refer to both reliability and validity as those concepts are used in science. *Daubert* set forth four factors that courts might consider in evaluating reliability: (a) whether the method used by the expert has been subjected to empirical testing, (b) whether the error rate for the method is known (or potentially known), (c) whether the method used by the expert survived the scientific peer-review process and was published in a peer-reviewed journal (from which a court is expected to find help in evaluating the soundness of a study's methodology), and (d) whether the procedure or test used by the expert is generally accepted within the relevant scientific community. The last of these factors had been the principle criterion for gauging the admissibility of expert testimony in federal court before *Daubert* (see *Frye v. U.S.*, 1923) and still is in about half the states.

Daubert and *Joiner* also stressed that experts must show that the methods relied on in forming their opinions are validly linked to the facts of the specific case, a requirement that in essence repeats the relevance inquiry in the expert context. Judges are to exclude expert testimony that is "connected to existing data only by the *ipse dixit* ['mere assertion'] of the expert" (p. 146). That is, forensic mental health experts must use methods that are appropriate for the specific population and circumstances of the case at hand, and they must be ready to provide relevant information about the "fit" between the method used and the specific case facts. This requirement suggests that courts should be particularly attentive to whether there are data from the field (as opposed to the lab) about the reliability and validity of a psychological tool and whether the tool was designed to address both the issue at hand (e.g., parental fitness, risk of reoffending) and the population in question (e.g., in terms of gender, age of the defendant/litigant).

Finally, in *Kumho Tire*, the Court extended *Daubert*'s validity standard to *all* expert testimony—not just "scientific" testimony but testimony based on "specialized" and "technical" expertise as well. This admonition means that mental health practitioners who consider themselves clinicians and not necessarily "scientists" must still demonstrate the validity of their work. Put simply, in jurisdictions where the *Daubert* trilogy has been adopted, *all* expert testimony, of any kind, is subject to suitable tests of validity.

The Current Project

The current project is a two-part analysis of psychological assessments by psychologists in legal contexts. The first is a systematic review of 364 psychological assessment tools that psychologists use in legal cases, as reported in 22 different surveys of experienced forensic mental health practitioners. The review analyzes the tools from both a scientific and legal perspective, using the review sources and evidentiary rules described above. The second part of the project looks at how the courts have assessed the admissibility of a diverse subset of these tools in actual cases.

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

Forensic psychology is a subfield of psychology in which psychological science or professional practice is applied to the law to help resolve legal, contractual, or administrative matters (Neal, 2018). Forensic psychology has grown steadily in recent years, as evidenced by the APA's recognition of forensic psychology as a specialty area of psychology, its addition of a special section on "forensic activities" to the ethical principles and code of conduct for psychologists, and its publication of specialty guidelines for forensic psychology (see, e.g., APA, 2013; Neal, 2018). As of 2017, 7% of all licensed psychologists in the United States had a primary or secondary area of specialty in forensic psychology, and 8% of all board-certified psychologists in the United States were certified in forensic psychology (Lin, Christidis, & Stamm, 2017). Accompanying this growth is a major increase in the development, marketing, and use of psychological assessment instruments in forensic settings (Edens & Boccaccini, 2017; Grisso, 2003).

Clearly, knowing how to interpret demonstrated psychometric properties is important for psychologists to make case-by-case determinations of scientific appropriateness of a tool. However, studies of the quantitative training offered in doctoral psychology programs reveal "grave concerns about the most fundamental issues for adequate measurement" (Aiken, West, & Millsap, 2008, p. 37) and suggest that most graduates of doctoral programs in psychology (including elite programs) lack fundamental competencies in measurement science (see also Lambert, 1991; Meier, 1993; Merenda, 1996). Perhaps unsurprisingly, data from Furnham (2018) suggest that psychology practitioners are limited in their knowledge of psychometric criteria of tests. Thus, the first step in

evaluating the state of forensic psychology is coming to some understanding of which tools are commonly used.

What psychological assessment tools do clinicians use in forensic settings?

In a major investigation of the assessment practices and expert-judgment methods used by mental health professionals in forensic settings, Neal and Grisso (2014) found that most (74.2%) mental health professionals use psychological assessment tools in their forensic work. They also found that those clinicians who use tools usually rely on more than one, with an average of four different assessment tools for each forensic evaluation. Furthermore, they found that psychologists vary widely in their choice of tools, even when they are evaluating the same type of issue. These findings in the forensic evaluation context are substantively similar to large-scale investigations undertaken by the APA among clinical psychologists and neuropsychologists across other settings (e.g., Camara, Nathan, & Puente, 1998).

In addition to the investigation by Neal and Grisso (2014), there have been at least 21 other surveys of mental health practitioners that ask about tool use in general or specific forensic settings (Ackerman & Ackerman, 1997a, 1997b; Ackerman, Ackerman, Steffen, & Kelley-Poulas, 2004; Archer, Buffington-Vollum, Stredny, & Handel, 2006; Boccaccini & Brodsky, 1999; Borum & Grisso, 1995; Bow & Quinnell, 2001; Keilin & Bloom, 1986; LaFortune & Carpenter, 1998; Lally, 2003; Lees-Haley, 1992; Lees-Haley, Smith, Williams, & Dunn, 1996; Martin, Allan, & Allan, 2001; McLaughlin & Kan, 2014; Naar, 1961; Pinkerman, Haynes, & Keiser, 1993; Quinnell & Bow, 2001; Rogers & Cavanaugh, 1984; Ryba, Cooper, & Zapf, 2003a, 2003b; Slick, Tan, Strauss, & Hultsch, 2004). Across these 22 surveys, 364 distinct psychological assessment tools were identified as having been used by or acceptable for use by clinicians in forensic settings (King, Wade, & Tilson, 2017).

We used this set of 364 tools as the basis for the current investigation. This set of tools includes a diverse set of test types, such as aptitude tests (e.g., general cognitive and ability tests), achievement tests (e.g., tests of specific knowledge or skills), and personality, psychological, and diagnostic tests. It includes measures designed for both adults and youth. And it includes tools that can be used to address a wide range of referral questions, including, for example, competence or fitness to stand trial, violence risk assessment, sexual offender risk assessment, mental state at time of offense, aid in sentencing, disability, child custody, civil commitment, child protection, civil tort, guardianship, competency to consent to treatment, juvenile transfer to adult court, fitness for duty, and capacity to waive Miranda rights.

The characteristics of the mental health experts whose responses are recorded in the 22 surveys we canvassed are worth highlighting, since we relied on the information about tests used in legal settings based on the results from these expert respondents. Overall, they were well-qualified experts, nearly all of whom were doctoral-level psychologists with many years of experience in the field. For example, Neal and Grisso's (2014) survey of forensic mental health professionals—which yielded 286 of the 364 distinct psychological assessment tools we study in the current project—comprised 434 experts in the field, 91% of whom were doctoral-level clinicians. That sample had an average of 16.56 years (SD = 12.01) of forensic evaluation experience, and 16.4% of the clinicians were board-certified. The samples from the other surveys were largely comparable (more details to follow in the Coding General Acceptance section).

Method

Coding scheme. Our authorship team, composed of psychologists and lawyers with diverse sets of expertise relevant to this project, worked together to develop a coding scheme that would capture relevant characteristics of the 364 psychological assessment tools. We relied on legal standards and psychometric theory to guide the formation of the coding scheme. From the legal side, we coded whether each test was generally accepted in the field, whether it had been subjected to testing, and information about peer review.[1] From the psychometric side, coding the technical quality each tool involved reliance on the Mental Measurements Yearbook (e.g., Carlson et al., 2017), E. Strauss and colleagues' (2006) compendium of neuropsychological tests, Grisso's (2003) compendium of forensic psychological tests, the peer-reviewed published literature, and websites and other publication outlets for sources of information.

Coding general acceptance. As construed by the courts, general acceptance is a protean concept. Courts differ on what must be generally accepted, the scope of the "field" that must generally accept it, and how acceptance is to be measured (Giannelli, 1981). However, psychological assessment tools that have been recognized by leading treatises or texts or that are considered valid by the majority of clinicians who practice in the relevant forensic area are highly likely to be found generally acceptable by the courts if used for the purpose for which they are recognized as useful (Epps & Todorow, 2019; Giannelli, 1981).

We searched for evidence of general acceptance in forensic psychology and more broadly across psychological assessment. Our coders searched through nine articles published in the past 15 years that provide a general discussion of psychological tools for the names

and acronyms of each of the 364 psychological assessment tools in our list (Archer et al., 2006; Elhai, Gray, Kashdan, & Franklin, 2005; LaDuke, Barr, Brodale, & Rabin, 2018; Lally, 2003; McLaughlin & Kan, 2014; Neal & Grisso, 2014; Rabin, Barr, & Burton, 2005; Ryba, Cooper, & Zapf, 2003b; Slick et al., 2004). Each of these sources had information about tools that were frequently used and/or endorsed by psychologists, some of them across different settings (e.g., Elhai et al., 2005; Rabin et al., 2005), others specific to forensic settings (e.g., LaDuke et al., 2018; Lally, 2003; Neal & Grisso, 2014). Across these nine surveys that we relied on for coding general acceptance of tools in the field, a total of 2,384 mental health experts responded, almost all of whom were doctoral-level psychologists with an average of about 15 years of experience in the field and among whom about 15% were board certified.[2]

We conducted the initial search to determine whether these nine published surveys of experts' views of psychological assessment tools used in court mentioned the 364 tools or their acronyms through an automatic PDF cross-reference program. We then followed up the automatic search results with a team of coders who reviewed the results, corrected any errors, and coded the data as follows:

> There was not enough information available to determine whether a tool was generally accepted,
>
> the information indicated that the tool was indeed generally accepted in the field,
>
> the information about general acceptance was conflicting or unclear, or
>
> it was clear that the tool was not generally accepted.

Because the nine published articles that our coders relied on as sources when rating the general acceptance variable each used different words, phrases, and measurements to describe whether a particular tool was seen as acceptable or endorsed by clinicians, each coder was asked to indicate the evidence on which they based their coding decision (i.e., enter quotations and citations from sources of information) for each rating. This information was used by the coders when they met to discuss and resolve discrepancies between codes.

Coding whether a tool was subjected to testing. We asked our coders to look for evidence that each of the 364 tools had been subjected to testing. First, we searched for information about each tool's availability, including whether the tool is commercially published (i.e., whether someone is selling and charging money for it), the commercial publisher, and whether the tool has a commercially published manual. We also searched for whether each tool is available for free online (i.e., whether someone is offering it without charging money for it) and whether the tool has a noncommercially published manual. Finally, we compiled manual citations, as well as electronic links to the manual sources.

After this initial information was coded (and after the comprehensive-review source information was coded as described in the following section), we asked our coders whether there was evidence available that each tool had been subjected to testing (no, yes, or unclear), and to enter evidence for their coding decision. The evidence of testing included citations to published peer-reviewed articles, information in manuals, data sources online, and so forth. We did not ask the coders to try to find every instance of testing; rather, if they found evidence of testing and could code "yes" in response to the question, they were asked to move on to the next tool and continue coding new data.

Coding peer review of psychometric quality. Coders were asked to search through the three primary comprehensive review sources described earlier in this article for reviews of the 364 tools. They coded dichotomously for whether each tool was reviewed in the *MMY*, E. Strauss and colleagues' (2006) compendium of neuropsychological tests, and Grisso's (2003) forensic competencies compendium. They also entered quotations, summaries, and citations of information from each of these sources, especially with respect to the psychometric strengths and weaknesses of the various tools. They looked carefully for information about the tool's performance in forensic contexts. They were also asked to enter reviewers' references from the *MMY* because the professional reviewers who wrote those reviews relied on several different sources of information about the development and psychometric qualities of each tool. These reviewer references were then used toward coding the "evidence of testing" variable described in the previous section.

Finally, coders were asked to provide an *overall summary evaluation* of the psychometric and technical quality of each psychological assessment tool and to base this evaluation on whether the professional review sources (i.e., *MMY* database; Grisso, 2003; E. Strauss et al., 2006) concluded with generally favorable, generally unfavorable, or mixed summary evaluations. This overall summary evaluation item was modeled on a similar approach by Cizek and colleagues (2012). This method has significant limitations (detailed in the discussion to follow); however, it has the advantage of relying in large part on the most comprehensive review source available for psychological assessments—and importantly, one that is accessible to the public.

Table 1. Krippendorff's α Reliability Estimates at the Individual and Duo Levels

	Round 1: Individual level			Round 2: Duo level		
Criterion	α	95% CI	No. of pairs	α	95% CI	No. of pairs
Reviewed in Mental Measurements Yearbook?	0.82	[0.80, 0.85]	2,169	0.88	[0.82, 0.92]	363
Reviewed in Strauss et al., 2006 compendium?	0.81	[0.77, 0.84]	2,184	0.83	[0.74, 0.91]	363
Reviewed in Grisso, 2003?	—	—	—	1.0	[1.00. 1.00]	364
Is the tool commercially published?	0.61	[0.58, 0.65]	2,126	0.73	[0.65, 0.80]	364
Does the tool have a commercially published manual?	0.52	[0.49, 0.56]	2,093	0.72	[0.65, 0.79]	364
Is the tool available for free online?	0.49	[0.46, 0.53]	2,117	0.64	[0.55, 0.72]	364
If noncommercial, does the tool have a manual?	0.24	[0.21, 0.28]	2,092	0.62	[0.54, 0.70]	363
Has the tool been tested?	0.39	[0.34, 0.44]	2,178	0.63	[0.51, 0.75]	364
Is the tool generally accepted in the field?	0.46	[0.43, 0.50]	2,175	0.75	[0.69, 0.81]	364
Overall summary evaluation	0.82	[0.80, 0.84]	2,181	0.86	[0.81, 0.90]	365

Note: Each estimate is the product of 10,000 bootstraps at the individual level (four independent coders) and at the duo level (2 teams of 2 coders) for the coded variables. CI = confidence interval.

Coding process and interrater reliability. The coding was carried out by a team of 30 coders, who were trained in applying the foregoing coding scheme. These coders included one postdoctoral scholar in psychology and law, two students pursuing a PhD in psychology and law, 22 students pursuing a master of science in psychology, and five advanced undergraduate psychology majors active in author T. M. S. Neal's psychology research lab. Most of the coding took place during three full-day coding events in the fall of 2018. Each variable was coded separately by four independent coders. Then the coders for each variable were assigned partners to work through each assigned rating and resolve discrepancies. Finally, the duos were asked to meet with the full complement of four coders and resolve discrepancies between the duos at the final integrated level of coding.

We calculated interrater reliability statistics for both the individual level of coding and the duo level of coding. We used a robust and general reliability statistic called Krippendorff's α (Kα), which conservatively generates reliability estimates for judgments with any number of raters, at any level of measurement, across sample sizes, and with or without missing data (Hayes & Krippendorff, 2007). Given our coding process, it is noteworthy that Kα treats raters as freely interchangeable and is unaffected by number of raters. Consistent with Kα as a measure of the reliability of data (as opposed to the reliability of raters), we used a bootstrapping algorithm to estimate the reliability distribution for each coded variable by resampling hypothetical reliability data from pairs of values found in the original data. All of our variables were treated as nominal except the overall summary evaluation, which we treated as ordinal.

As indicated in Table 1, the Krippendorff α coefficients for each coded variable improved from the individual (four separate coders) level—where the values ranged from .24 (poor reliability) to .82 (good reliability)—to the duo integration level (two teams of two coders each), where the Kα values ranged from .62 (questionable reliability) to 1.0 (perfect reliability), with most of the variables having acceptable to good reliability. The item asking coders to determine whether each tool had been reviewed in Grisso (2003) was perfectly reliable with just one round of rating.

Our results section reports the integrated data from the two duo groups *after* resolving differences between the duos for each variable coded. The reliability of the final results is not calculable, because it represents the integration of differences and negotiation of discrepancies in the data between the two duo levels and is thus one additional step beyond that for which we can provide reliability information. However, given that reliability became stronger at each level of coding integration, we believe the final integrated results that we report are sufficiently reliable for drawing meaningful inferences from these data.

Results

The descriptive statistics from our coding results are provided in Table 2 and Figures 1 to 3. More than half of the tools (60%) were reviewed in the *MMY* (along with 89% of the commercially published tools), but only 14% were reviewed in E. Strauss and colleagues' (2006) neuropsychological compendium and only 7% in Grisso's (2003) forensic text, presumably because the latter two volumes are narrower in focus. Most of the tools were published commercially (68%), though a sizeable minority (27%) were available for free online. We found manuals for most of the tools; however, we

Table 2. Descriptive Coding Results

Criterion	Result categories			
	Yes	No	Unclear	Not applicable
Reviewed in Mental Measurements Yearbook?	220 (60%)	144 (40%)	—	—
Reviewed in Strauss et al., 2006 compendium?	52 (14%)	312 (86%)	—	—
Reviewed in Grisso, 2003?	24 (7%)	340 (93%)	—	—
Is the tool commercially published?	248 (68%)	83 (23%)	33 (9%)	—
Does the tool have a commercially published manual?	232 (64%)	94 (26%)	38 (10%)	—
Is the tool available for free online?	98 (27%)	231 (63%)	35 (10%)	—
If noncommercial, does the tool have a manual?	64 (18%)	30 (8%)	48 (13%)	222 (61%)

Note: Values are *n*s with percentages in parentheses. Some tools were both published commercially and available for free online.

could not find manuals for about a quarter of the tools, and for about 10% it was unclear whether there was a manual available.

As Figure 1 shows, most of the tools (n = 326 of 364, 90%) have been subjected to testing. Figure 2 shows 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 about general acceptance, we were able to clearly identify about two thirds as generally accepted (n = 119 of 179, 67%). For 16.8% of the tools, evidence concerning general acceptance was conflicting (n = 30 of 179), and for the remaining 16.8% (n = 30 of 179), the evidence was clear that they are *not* accepted in the field.

The results from the overall summary evaluation are depicted in Figure 3. The data show that 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).

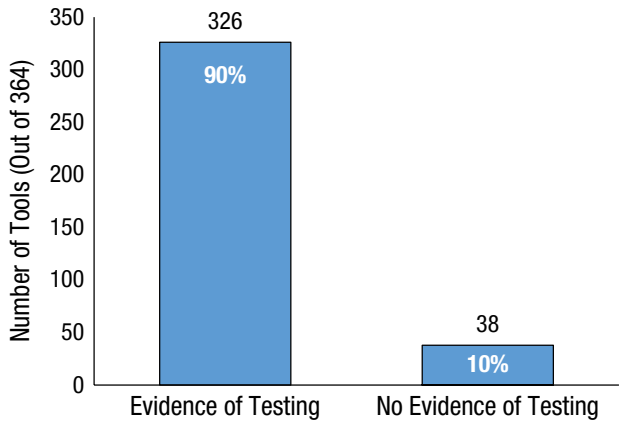

Fig. 1. Evidence that the tool was subjected to testing. Number of tools is graphed according to whether there was any evidence of testing.

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).

We also tested the strength of the relationship between the general acceptance in the field and the overall summary evaluation of quality variables according to the professional review sources (see Fig. 4). Encouragingly, the darkest-colored areas of the bars (associated with tools that are generally accepted) is larger for the tools with generally favorable reviews, and smaller for tools with generally unfavorable reviews. However, this figure also shows other relationships, such tools with generally unfavorable reviews nevertheless being generally accepted. To estimate the strength of the relationship between general acceptance and overall summary evaluation of quality, we used Cramer's V, which produces values ranging from 0 to 1, where a value of 0 indicates no relationship and 1 indicates a perfect relationship (Cohen, 1988). 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

In this article, we avoid making conclusions about which tools are "good" or "bad." This investigation is a snapshot in time, and the evidence base upon which our ratings were made will grow and change over time. Thus, we focus on the major themes of the findings, at a higher level of abstraction than the individual tool level.

Both positive and problematic findings emerged from this analysis. Among the positive findings are that

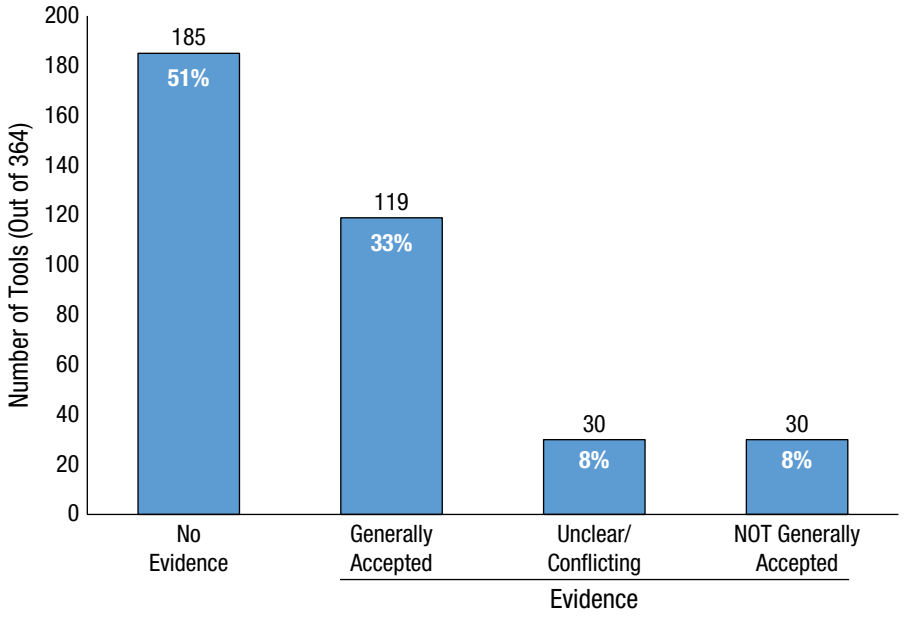

Fig. 2. General acceptance in the field. Number of tools is graphed separately for each coding category.

there are many psychometrically strong tests used by clinicians in forensic practice. And, consistent with their roots in psychological science, psychological assessment tools are nearly all tested—and thus have a known or knowable error rate with respect to at least some outcome measures. In addition, there is a positive relationship between the overall psychometric strength of tools and their general acceptance. However, among the more concerning findings are that only about 67% of the tools used by clinicians in forensic settings could

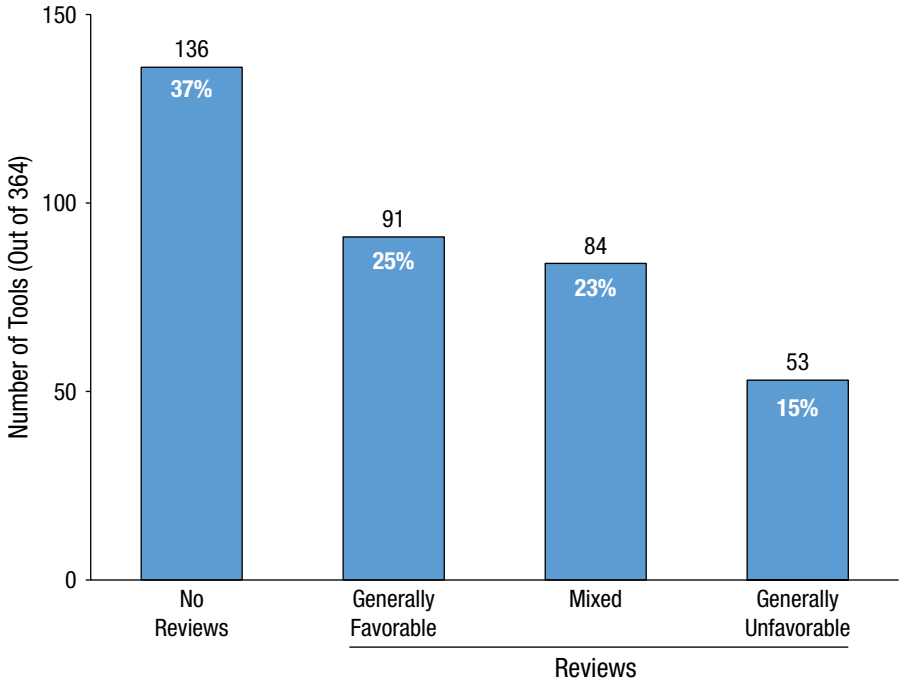

Fig. 3. Overall summary evaluation of quality. Number of tools is graphed separately for each overall summary category.

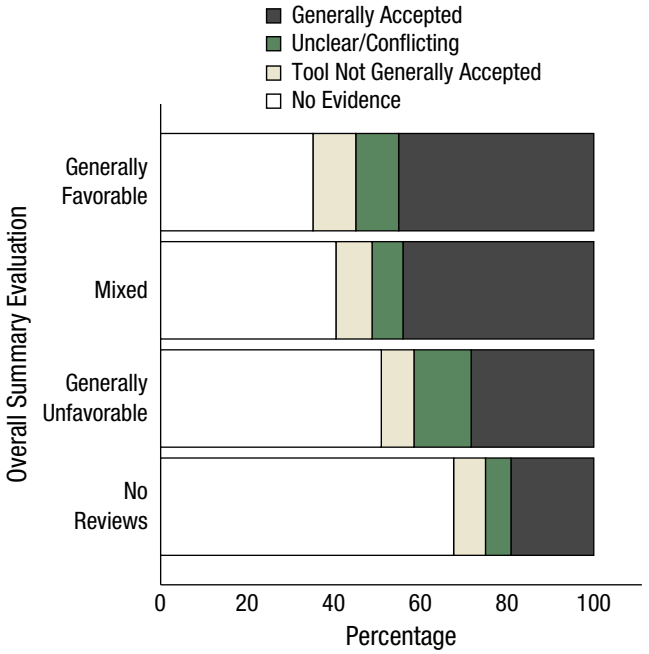

Fig. 4. Relationship between overall evaluation of quality and general acceptance in the field. Overall summary evaluation is graphed as a function of the percentage of tests falling into each coding category.

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. General acceptance can be a desirable attribute of a tool but is not sufficient by itself to establish validity.

The process of coding the data yielded several insights. For instance, 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. The obvious concern this raises about the scientific bona fides of these tests is exacerbated by the possibility that consumers (including practicing psychologists) may not be aware of these facts. Although this observation may reflect market economy pressures, proprietary interests, and intellectual-property concerns, test developers concerned about getting their product to market quickly may lose out on the value of the scientific process of peer review, which could help refine the tools in their developmental process and improve legal admissibility.

Another observation is that 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, given the commercial incentives at stake when the tool developer is the entity that provides the relevant data, may not in fact have been subject to the same level of scrutiny that is accorded psychometric results from independent testing. Furthermore, test reviews are a different category of article and typically are not subjected to the same scrutiny as original empirical articles.

A final issue worth noting involves the class of psychological assessment tools called structured professional judgments (SPJ). SPJ tools are designed as checklists or memory aids and generally rely on evidence-based factors that clinicians are supposed to consider as relevant to particular judgment tasks. SPJs intentionally eschew traditional psychometrics and norm-based interpretations and are "designed with the individual client in mind" rather than for group-based predictions (e.g., Historical Clinical Risk Management–20, or HCR-20; Douglas, Hart, Webster, & Belfrage, 2013). Although clinicians can rate each item, they do not add up the ratings or interpret the meaning of the numerically integrated items. Researchers sometimes study these tools by treating the items actuarially to study reliability and validity, but this research treatment of the tools is substantively different from how the tools are used in practice. Equating research outcomes with clinical use of these tools by assuming the actuarially treated data in the research literature is generalizable to clinical use is a problem for the class of SPJ tools. Consequently, SPJ tools are challenging to evaluate both from traditional psychometric theories and under the *Daubert* criteria.

Limitations

Limitations of scope. One limitation of this project has to do with its scope. This project focuses on the psychometric properties of individual tests known to be used in forensic contexts. Although this focus enabled us to survey the field in a scientific manner, it means that we did not examine two other potentially problematic ways in which psychological-assessment evidence comes into court.

First, we did not study unaided clinical judgment (i.e., evaluations conducted without standardized measurement tools). About 25% of psychologists providing clinical expert testimony in court continue to rely on unstructured evaluations (Neal & Grisso, 2014). Evaluations based on unaided clinical judgment do not lend themselves to testing and thus cannot easily be analyzed from the *Daubert*-type perspective we used here. However, it is worth noting that the weight of the evidence indicates that structured approaches using psychological assessment tools are more valid and reliable than unaided clinical judgment (e.g., Ægisdóttir et al., 2006; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000).

The second type of psychological assessment technique not evaluated by this project—one even more common than unaided clinical judgment—is the use of test batteries. Many psychologists use multiple psychological tests to inform their conclusions. Indeed, an APA Psychological Assessment Work Group recommended that clinicians use a multimethod assessment battery to maximize the validity of assessments (Meyer et al., 2001). Consistent with this recommendation, Neal and Grisso (2014) found that clinicians use on average four different psychological assessment methods in a given forensic evaluation (*SD* = 2.95, range = 1–18).

The tremendous diversity within the field in terms of the various theories and assessment approaches "breeds the divergence in opinion that makes the 'battle of the experts' a regular courtroom occurrence" (Faust & Ziskin, 1988, p. 33). 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* (AERA, APA, & NCME, 2014) indicate (p. 155),

> When psychological test batteries incorporate multiple methods and scores, patterns of test results frequently are interpreted as reflecting a construct or even an interaction among constructs underlying test performance. Interactions among the constructs underlying configurations of test outcomes may be postulated on the basis of test score patterns. The literature reporting evidence of reliability/precision and validity of configurations of scores that supports the proposed interpretations should be identified when possible. However, 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. . . . If the literature is incomplete, the resulting inferences may be presented with the qualification that they are hypotheses for future verification rather than probabilistic statements regarding the likelihood of some behavior that imply some known validity evidence.

Because a number of tests might be relevant to most areas of legal inquiry, and because evaluators might combine these tests in a number of different ways, this project could not analyze that method of assessment.

Methodological limitations. In addition to the scope limitations of this project, several methodological limitations should be noted. First, we did not distinguish between multiple versions of a single tool, instead combining the psychometric and technical information for all of the relevant versions. Doing so obscures some important details, given that newer versions of tests often have improved psychometric and technical properties compared with earlier versions. This limitation is reasonable given our goal of providing a big-picture overview but would be problematic if we tried to make finer judgments about individual tests.

Our heavy reliance on comprehensive review sources such as the *MMY* for much of our coding process has some significant drawbacks. One limitation is that the *MMY* reviews only commercially published tests. Another is that it does not review all published tests. Publishers must choose to have their measures reviewed in the *MMY* and must submit certain test information and technical data for a review of their tests to be published in the *MMY*. Thus, publishers of inadequately documented tests or tests with poor technical quality may not seek review by the *MMY*.[3] Sometimes even tests with strong psychometric evidence are not reviewed in the *MMY*, such as those that are not commercially published or those belonging to publishers that choose, for whatever reason, to not have their tests reviewed.

An additional limitation of the *MMY* is the possibility that the reviewers missed or misinterpreted some technical information about a given test. In the introduction to every yearbook (e.g., Carlson et al., 2017, p. xiv), the *MMY* cautions that

> Active, evaluative reading is the key to the most effective use of the professional expertise offered in each of the reviews. Just as one would evaluate a test, readers should evaluate critically the reviewer's comments about the test. The reviewers selected are competent professionals in their respective fields, but it is inevitable that their reviews also reflect their individual perspectives. The Mental Measurements Yearbook series was developed to stimulate critical thinking and assist in the selection of the best available test for a given purpose, not to promote the passive acceptance of reviewer judgment.

Thus, the reviews in the *MMY* should also be read with a critical eye, and attention to the primary literature is important. This is especially true for psychological-assessment experts. We do not advocate that experts who use psychological assessment tools rely on sources such as the *MMY* as their primary decision criterion for whether to use a test. The bar for experts is clear from the ethics codes; experts are supposed to rely on the manuals and the primary literature to determine whether a given tool is scientifically strong enough for use in a given case (e.g., ethical standards 2.04, 9.01, 9.02, and 9.09 of the *Ethical Principles of Psychologists and Code of Conduct*, APA, 2017; standards 10.2 and

10.5 of the *Standards for Educational and Psychological Testing*; AERA, APA, & NCME, 2014).

An inevitable methodological limitation of our project is that, except with respect to the "tested" variable, we did not do a systematic review of the peer-reviewed literature for all 364 tools (and even with respect to the tested variable we stopped the review once we identified one source that confirmed the tool had been tested). Rather, our "overall quality" rating relied on the content in published review compendiums (i.e., *MMY*, Grisso, 2003; E. Strauss et al., 2006). This method was comprehensive enough for the abstract interpretations we sought to make about the overall quality of tools that mental health professionals use, how use of these tools can improve, and the types of issues the legal system should be aware of in connection with the use of psychological assessments in expert evidence. But it necessarily means that the database cannot serve as a comprehensive source of information about these tools.

Ideally, we would have undertaken a full psychometric review of all information about each and every one of the 364 tools (e.g., in the published and unpublished literature, in test manuals, on test websites) in order to classify and describe the quality of the tests. As we noted above, evidence available in the primary, peer-reviewed literature should be the ultimate arbiter of the scientific worth of assessment tools. However, for this project, we instead relied on comprehensive review sources such as the *MMY* for two reasons.

First, creating a comprehensive psychometric review of all sources for each tool would have been a Herculean (and arguably impossible) task for an article that seeks to generate a general picture of what is going on without claiming specific tools are "good" or "bad." The *MMY* review process involves the full-time effort of a staff of several professionals and the assistance of hundreds of expert reviewers in the field who review tests on an ongoing basis, something we could not replicate. Second, any review database we provided would be almost immediately out of date, unlike the primary literature and the *MMY*, which are regularly updated. Our goal for this article is to emphasize the need to critically evaluate psychological tests and to inform readers how and where to find up-to-date information about the tests they encounter, rather than try to provide comprehensive evaluations ourselves. Our methods are sound for achieving this goal.

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

The second part of the project is a systematic analysis of case-law admissibility challenges to psychological assessment tools used in court. Surveys of judges (e.g., Gatowski et al., 2001) and analyses of judicial behaviors (e.g., Fradella, Fogarty, & O'Neill, 2003; Groscup, Penrod, Studebaker, Huss, & O'Neil, 2002) suggest that judges aim to apply *Daubert*-like admissibility criteria in carrying out their gatekeeping role. But research also shows that judges struggle to apply these criteria (Dahir et al., 2005; Gatowski et al., 2001; Groscup et al., 2002). Consequently, 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).

An evidentiary framework

Judicial discussions of psychological tests are difficult to analyze without a typology of the considerations that might affect judicial analysis. Faigman and colleagues provided such a typology, set out in two law review articles (Faigman, Monahan, & Slobogin, 2014; Faigman, Slobogin, & Monahan, 2016). They make a basic distinction between framework testimony and diagnostic testimony. Framework testimony is testimony about general scientific concepts that can be applicable to more than one case. In contrast, diagnostic testimony seeks to draw conclusions about a particular case by applying the general scientific/framework knowledge to the case at hand. Faigman and colleagues also point out, however, that much diagnostic testimony is not, or at least should not be, the province of experts. They argue that an expert should not be allowed to proffer diagnostic testimony if scientists have not developed a diagnostic methodology that permits valid statements to be made about individual cases.

As to how courts should analyze the admissibility of testimony that is based on science, Faigman and colleagues (2014; Faigman et al., 2016) suggest considering five admissibility criteria, all based on the *Daubert* trilogy and the federal rules of evidence: (a) relevance or "fit," (b) qualifications, (c) helpfulness (added value), (d) prejudicial impact, and (e) validity. *Fit* is the notion that the evidence must answer the legal question presented. *Qualifications* refer to whether the expert has the specialized knowledge or skills necessary to offer the opinion proffered. *Helpfulness* and *prejudicial impact* interact with one another: Even relevant and highly valid expert testimony, offered by a qualified expert, should not be admitted if it does not add to what a lay jury could readily discern for itself or if it would distract the jury because it seems more relevant to an issue not in the case or confuse the fact-finder because of its complexity. Courts often express this idea as a concern that the expert is usurping the province of the jury.

The validity admissibility criterion—or what the Supreme Court in *Daubert* called "reliability"—is the

main concern of this study. The *Daubert*-type factors might be applied quite differently depending on whether the proffered testimony is general or diagnostic in nature. Framework evidence is usually amenable to some type of scientific verification process. For instance, the validity and interrater reliability of a particular tool can be measured through laboratory or field studies involving large sample populations. But the basis of diagnostic testimony can often be very difficult to "test" unless there is some sort of feedback loop with respect to individual cases. Although such feedback loops exist for certain types of forensic evidence (e.g., proficiency testing of ballistic experts), in other cases (e.g., evidence about a defendant's past mental state) the only measure of validity may be an assessment of the method used to reach the diagnosis.

A final consideration is the division between legal questions that should be left to judges in their capacity as gatekeepers of evidence, and factual questions that should be left to the fact-finder (usually a jury). Admissibility questions of general import—that is, those that could apply to other cases besides the one in question—should be subject to *Daubert* analysis (Faigman et al., 2014; Faigman et al., 2016). That would include not only all framework testimony (which by definition is group based) but also diagnostic methods used to say something about the individual that could be used in other cases. For instance, a statement that, according to the Minnesota Multiphasic Personality Inventory (MMPI; Butcher et al., 2001), people with a particular profile tend to be malingering or depressed would be subject to *Daubert* analysis, as would whether the MMPI is the proper tool for addressing those issues. In contrast, whether the expert properly applied an accepted methodology, such as correctly administering the MMPI in the particular case, is generally an issue of fact that should be decided by the fact-finder (Faigman et al., 2016). Also left up to the jury is whether the substance of admissible expert testimony is persuasive (e.g., this defendant is malingering or depressed).

Although the results reported below make use of this typology, including the distinction between substance and methodology, the focus of our analysis is on validity and, to a lesser extent, fit, which focuses on validity in context.

Selecting the tools to study

To focus this analysis, we selected as exemplars 30 tools from the 364 tools we evaluated in the first portion of the project. These 30 exemplars were chosen with two goals in mind: (a) representation of a wide range of legal issues and forensic referral questions and (b) variety in terms of general acceptance and overall summary evaluation ratings of quality. With respect to the former concern, we selected tools used in civil, criminal, juvenile, and administrative cases; tools for past, present, and future psychological symptoms and conditions; and tools that address ultimate and secondary legal issues, to get an approximate snapshot of what is happening in the courts. With respect to the latter concern, we sought tools that fit each combination of the general acceptance and overall quality variables. For example, we included tools that are generally accepted and have generally favorable reviews, tools that are not generally accepted and have generally unfavorable reviews, and other combinations of these variables. Table 3 lists all 30 exemplar tools and gives the tool names and acronyms in the first two columns.[4] The third column of Table 3 gives the number of cases in which each of the 30 exemplar tools was mentioned at all in the legal database during the 3-year sample period. The fourth notes the number of cases in which the tool's admissibility was challenged (as opposed to simply mentioned). Table 4 shows the categories with respect to general acceptance and favorability of reviews of the 30 exemplar tools.

Method

To get a sense of how often the 30 exemplar tools were discussed and challenged by the courts, the three law scholars on the authorship team, along with law-student research assistants, searched a major legal database, Westlaw, for all judicial opinions and orders from all states and federal courts during the calendar years 2016, 2017, and 2018. Westlaw records include appellate cases but also trial-level evidentiary rulings—all of which we included in our search. The search terms consisted of each exemplar tool's acronym as well as its name spelled out, in an effort to catch references regardless of which version of a tool was used. Where a tool's acronym is a common word (e.g., ASPECT), additional search terms, such as "expert witness" or "psych!" (to capture any word containing the root "psych") were used to narrow the results to cases in which the name referred to the tool.

Each case produced by this method was scrutinized to verify that it involved use of the tool. For example, cases referring to use of the Rorschach test were retained, but cases in which a witness was named Rorschach were discarded. A total of 876 cases were screened in on the basis of these criteria. Except for cases discussing the MMPI, the PCL, and the Rorschach, every screened-in case involving a discussion of the tool beyond a mere mention was closely read. Because there were so many cases discussing the MMPI (n = 485), the PCL (n = 50), and the Rorschach (n = 59), we

Table 3. Exemplar Tool Names, Acronyms, Number of Cases Citing Each, and Number of Cases With an Admissibility Challenge

Test/tool name	Acronym(s)	No. of cases citing tool	No. of cases with an admissibility challenge	Reference
Abel and Becker Cognition Scale	ABCS	12	1	Abel, Becker, & Cunningham-Rathner (1984)
Ackerman-Schoendorf Scales for Parent Evaluation of Custody	ASPECT	0	0	Ackerman (2005)
Amsterdam Short-Term Memory Test	ASTM	0	0	Schmand, Lindeboom, Merten, & Millis (2005)
Hamilton Rating Scale for Depression	HRSD, HDRS, HAM-D, RHRSD	1	0	Warren (1994)
Historical Clinical Risk Management–20	HCR-20, HCR-20V3	30	0	Douglas et al. (2013)
Iowa Gambling Task	IGT	1	0	Bechara (2016)
Kaufman Test of Educational Achievement	KTEA	8	0	Kaufman & Kaufman (2014)
Kinetic Family Drawing	KFD	0	0	Knoff & Prout (1985)
Malingering Probability Scale	MPS	0	0	Silverton & Gruber (1998)
Millon Clinical Multiaxial Inventory	MCMI, MCMI-II, MCMI-III, MCMI-IV	24	1	Millon, Grossman, & Millon (2015)
Mini International Neuropsychiatric Interview	M.I.N.I.	0	0	Sheehan & Lecrubier (1997)
Minnesota Multiphasic Personality Inventory	MMPI, MMPI-2, MMPI-A, MMPI-2-RF, MMPI-A-RF, MMPI-2 RC	485[a]	1	Ben-Porath & Tellegen (2008)
Paced Auditory Serial Addition Test	PASAT	1	0	Gronwall (1977)
Pain Patient Profile	P-3	8	0	Tollison & Langley (1995)
Personality Assessment Inventory	PAI, PAI-A	31	3	Morey (2007)
Psychopathy Checklist	PCL-R, PCL:SV, PCL:YV, PCL-R:2nd ed	50[a]	4	Hare (1991)
Rape Myth Acceptance Scale	RMAS	0	0	Burt (1980)
Rogers Criminal Responsibility Assessment Scales	R-CRAS	2	0	Rogers (1984)
Rorschach Inkblot Test	Rorschach, RCS, R-PAS	59[a]	3	Exner & Erdberg (2005)
Rotter Incomplete Sentences Blank	RISB	1	0	Rotter, Lah, & Rafferty (1992)
Slosson Intelligence Test	SIT, S-FRIT	15	0	Slosson, Nicholson, & Hibpshman (1991)
Static-99	Static-99, Static-99R, Static-2002, Static-2002R, Static-2007	30	3	Hanson & Thornton (2000)
Structured Interview of Reported Symptoms	SIRS, SIRS-2	19	2	Rogers, Sewell, & Gillard (2010)
Structured Inventory of Malingered Symptomatology	SIMS	24	0	Widows & Smith (2005)
Substance Abuse Subtle Screening Inventory	SASSI, SASSI-2, SASSI-3, SASSI-4	16	0	Miller & Lazowski (2016)
Thematic Apperception Test	TAT	11	0	Murray (1943)
Trauma Symptom Inventory	TSI, TSI-2	16	0	Briere (2011)
Wechsler Adult Intelligence Scale	WAIS, WAIS-R, WAIS-III, WAIS-IV	30	1	Wechsler (2008)
Wisconsin Card Sorting Task	WCST, WCST–64, M-WCST	1	0	Heaton, Chelune, Talley, Kay, & Curtis (1993)
Word Choice Test (part of ACS for the WAIS-IV and WMS-IV)	WCT	1	0	Wechsler (2009)

[a]A random subsample of 30 cases was analyzed.

Table 4. Categorization of 30 Exemplar Tools by General Acceptance and Quality

	No professional reviews	Generally unfavorable reviews	Mixed reviews	Generally favorable reviews
Generally accepted	ABCS Static-99	ASPECT PASAT SIMS	HCR-20 SASSI WCST	MMPI PAI PCL-R SIRS TSI WAIS
General acceptance debated		P-3 TAT	MCMI R-CRAS Rorschach	
Not generally accepted	M.I.N.I. RMAS WCT	KFD SIT/SFRIT RISB	IGT MPS	ASTM HRSD KTEA

Note: See Table 3 for acronym definitions. The header row contains the overall summary evaluations of psychometric quality (no professional reviews, generally unfavorable reviews, mixed reviews, or generally favorable reviews). Tools listed in the "generally accepted" row were reported as frequently used by clinicians and had received ratings of acceptability by clinicians, or they were frequently used but with no data about perceptions of acceptability. Tools listed in the "general acceptance debated" row include tools that are frequently used by clinicians and yet rated as unacceptable by many Others, or else are endorsed as acceptable for use, yet infrequently used. Tools listed in the "not generally accepted" row were infrequently used by clinicians and rated as unacceptable for use in forensic settings, or infrequently used with no data about perceptions of acceptability.

randomly selected a subsample of 30 cases of each to subject to careful examination. Thus, a total of 372 cases were analyzed. These cases were read closely by the law professors on the authorship team to determine whether the tool's admissibility had been challenged and, if so, on what grounds and with what result.

Results

The results are organized in Tables 3 and 5. The third column of Table 3 gives the number of cases in which each of the 30 exemplar tools was mentioned at all in the legal database during the 3-year sample period. The fourth column notes the number of cases in which the tool's admissibility was challenged (as opposed to simply mentioned). The admissibility challenge results are organized in Table 5. The 9 exemplar tools that received admissibility challenges are listed in the first column of Table 5. The remaining columns provide information about cases in which a challenge to the tool occurred. They are listed under the type of issue addressed in the admissibility discussion, along with a designation as to the focus of the admissibility analysis. The table note provides more detail.

Frequency of test usage. The first observation one might make about these findings concerns the frequency with which these tests—all of which clinicians report are used in actual practice—appear in case law. Many of our 30 exemplars are mentioned dozens of times and one—the MMPI—is mentioned 485 times. But 12 tools (40% of them) were mentioned only once or not at all. Of course, this absence could be due to something other than an expert's decision not to use the test. For instance, use of the tool might not be mentioned in an expert's report or, if mentioned, might be of so little moment that the court does not allude to it. Or the court might simply refer to the expert's conclusions without specifying any of numerous tests the expert used. Finally, there may simply be little litigation on a psychological issue measured by a particular tool. The absence of judicial reference to a test is in itself not necessarily revealing. Still, the complete absence of any mention of a given test suggests that its use, if it does occur, has not ruffled legal feathers, a suggestion that dovetails with the next finding.

Frequency of legal challenges. The second, and less ambiguous, finding is the paucity of legal challenges to the use of tools. Although we found a moderately large number of cases that spent a paragraph or two describing the tool and/or the results obtained using it, admissibility issues were neither raised nor discussed in most of these cases.[5] 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%; see Fig. 5). In the vast majority of the remaining cases, the courts' discussion of the tool involved questions raised

Table 5. Categorized Judicial Discussions of Admissibility for the Exemplar Psychological Tools That Received an Admissibility Challenge Across 372 Legal Cases (2016–2018)

	Type of admissibility discussion			
Tool	Fit	Validity 1	Validity 2	Other
ABCS	***People v. Fortin***[a] **(FS, FM, DM, DI)**	***People v. Fortin***[a]	***People v. Fortin***[b]	
MCMI—all variations	***Tardif v. City of NY***[c] **(FS)**			
MMPI—all variations				*J.K.J. v. Polk Cty.*[d] (qualifications)
PAI—all variations		*Hopey v. Spear*[e] (FM) *Reaes v. City of Bridgeport*[f] (FS)	***Savage v. State***[g] **(DM)**	
PCL—all variations	*In re Commitment of Sternadel*[h] (DM) *In re Commitment of Gomez*[i] (FS) ***United States v. Gamble***[j]	*State v. Gary K.*[k] (FM)		***United States v. Gamble***[j]
Rorschach—all variations	*United States v. Jones*[l] (DI)	*In the Matter of Garcia*[m] (FM)	*In the Matter of Garcia*[m] *Lefkowitz v. Ackerman*[n] (FM)	
SIRS—all variations	*People v. Jing Hua Wu*[o] (FM)	***People v. Howard***[p] **(FM)**		
Static-99—all variations	***State v. Gordon***[q] **(FS)**	***State v. Gordon***[q] *In re Detention of Wygle*[r] (FS)		*In re Commitment of Hood*[s] (DM)
WAIS—all variations			*Cannon v. Comm. of Soc. Sec.*[t] (DI)	

Note: See Table 3 for acronym definitions, number of cases that cited each tool, and number of cases with an admissibility challenge for each tool. Boldface type indicates cases in which exclusion occurred. Cases may appear more than once in a row if more than one admissibility issue was considered. "Fit" refers to relevance to the legal issue. "Validity 1" refers to testability/error rates. "Validity 2" refers to general/peer acceptance. "Other" includes qualifications of the experts, helpfulness, and prejudicial impact. FS = the focus of the admissibility analysis was on the substance of framework evidence; FM = the focus of the admissibility analysis was on the methodology of the framework evidence; DM = the focus of the admissibility analysis was on the methodology used to obtain diagnostic evidence; DI = the focus of the admissibility analysis was on the credibility/rationality of testimony about how the diagnostic methodology in the case at hand was used and the results that the expert reached.

[a]In *People v. Fortin* (2017), the court extensively considered the lower court's exclusion of the Abel Assessment for Sexual Interest ("Abel test") under California's Kelly-Frye test. Test results were proffered by defendant to rebut testimony that he had sexually abused children under the ages of 10 and 14. Although the lower court allowed the defendant's expert "to opine that [the defendant] lacks sexual interest in prepubescent children," it did not allow testimony on the defendant's performance on the Abel test. The lower court found that "the Abel test has not been adequately peer-reviewed; is not accepted in the scientific community; is designed to monitor convicted sex offenders; and is not intended for use in trials to determine a defendant's guilt or innocence" (p. 872). The court added that "the Abel test is a new scientific technique, process or theory. The Abel test has been deemed unreliable by the Supreme Courts of Connecticut, Maine, Montana, North Dakota and Texas. It has also been rejected in federal court" (p. 874). The court concluded as follows (p. 874):

> [Defendant's expert] cast[s] doubt on the use of the Abel test by acknowledging that the test assesses "persisting" sexual interests in convicted offenders, has not been peer reviewed, is not generally accepted in the scientific community to diagnose pedophilia, and can be thwarted by the test-taker. Under the test's user terms, [the expert] is not permitted to analyze or interpret Abel test results: he must send the results to Atlanta, then assumes proper analysis by Dr. Abel or trained staff and assumes the legitimacy of the results. . . . The process of analyzing responses is closely guarded proprietary information that Dr. Abel refuses to share. Admitting the results of [the expert's] Abel test would invite the jury to infer that [defendant] did not molest the victims: despite its scientific name—the "Sexual Interest Assessment"—the test is not designed to determine whether an accused committed a sex crime against children.

[b]See footnote a. In addition, the court in *People v. Fortin* (2017) said in a footnote that "the Abel test is never used to infer whether someone committed a particular sex act; rather, it reveals which categories provoke persisting sexual interest" (p. 873, note 2).

[c]In *Tardif v. City of New York* (2018), the court agreed with defendant's argument that expert could use the MCMI-III to *diagnose* the defendant, but could not use the test to support a conclusion about causation. Such use, the Court held, would not meet the reliability requirements of Rule 702 and *Daubert*.

[d]The court in *J.K.J. v. Polk County* (2017) considered at length the plaintiff's claims that the experts were not qualified to administer the MMPI. The court rejected these claims and was not asked to, and did not, consider the validity of the test, either as a general matter or as applied.

(continued)

Table 5. (continued)

[e]In *Hopey v. Spear* (2016), the court rejected the plaintiff's challenge of defendant's expert opinion regarding the use of several psychological tests. The court noted as follows:

> [Expert] identified the tests he used and explained the results. Further, he attached the entire Personality Assessment Inventory (PAI) test to the report. The PAI contained [the expert's] conclusions, notes, a graph, and a description of the various assessments performed on Plaintiff. Further, the tests performed on Plaintiff appear to be standard psychological tests that are accepted in the scientific community. (p. 5)

[f]In *Reaes v. City of Bridgeport* (2017), the court rejected plaintiff's claim that the PAI was "culturally biased," thus resulting in disparate impact in an employment decision. The court noted that plaintiff failed to "produce any 'statistical evidence showing that the challenged practice causes a disparate impact on the basis of . . . national origin'" (p. 4).
[g]In *Savage v. State* (2017), the court, following extensive analysis, found that the expert had failed to connect his methods and observations—especially the PAI—with his conclusion.
[h]For *In re Commitment of Sternadel* (2018), the court rejected the argument that the PCL-R is not meant to assess risk in Sexually Violent Predator hearings but is merely a research instrument, pointing out that "it was found that, in practice, there is a correlation between PCL-R scores above a certain threshold and rates of sexual-offense recidivism" (p. 19).
[i]For *In re Commitment of Gomez* (2017), the court rejected argument that, in a Sexually Violent Predator hearing, the fact that the PCL-R does not produce a "diagnosis" makes it irrelevant to determining whether the person has a "behavioral abnormality" as required by statute, noting the statute requires testing for psychopathy. Noteworthy that state's expert scored Gomez at 23 on the PCL-R, whereas the defense expert scored him at 12.
[j]In *United States v. Gamble* (2018), the court said "Because they are inapplicable here, these assessments would likely be, at best, an unnecessary expenditure of BOP's finite resources, and, at worst, affirmatively misleading" (p. 8).
[k]In *State v. Gary K* (2016), the court did not directly address admissibility, but stated that the PCL-R "has multiple flaws including problems with inter-rater reliability and the allegiance effect. PCL-R cut-off scores are also not consistently applied" (p. 16).
[l]In *United States v. Jones* (2018), the court rejected the government's argument that the expert's testimony (i.e., that the defendant had significant deficits in cognition, according to Rorschach results) was irrelevant to whether the defendant had specific intent to commit fraud).
[m]For *In the Matter of Garcia*, (2018) the court concluded that, although the defense expert "did not provide any evidence that [the Rorschach is] routinely used for sexually violent predator determinations" (p. 1), and although the state had a plausible argument that the evidence should be inadmissible under *Daubert* or *Frye*, the admission of the testimony was harmless, because the expert described it and other evidence as merely "pieces of the puzzle" (p.9), and there was no indication that the trial court considered the test in making its decision.
[n]In *Lefkowitz v. Ackerman* (2017), the court noted the plaintiff's argument in a custody case that the test was not "generally accepted" but stated that "it is impossible to tease out the effect, if any, of the alleged sub-par testing" (p. 11).
[o]In *People v. Jing Hua Wu* (2016), the defense objected to the relevance of the SIRS in an insanity case but did not preserve the issue, so the issue was not addressed on appeal.
[p]In *People v. Howard* (2018), the court upheld exclusion of SIRS' results because the trial court had found that the experts disagreed about whether the SIRS produced valid and reliable data and also found that the psychologist who administered the test had scored it improperly and thus would need to conduct further testing, which would result in delay of the insanity trial.
[q]In *State v. Gordon* (2018), the court upheld the offender's contention that use of the Static to enhance a sentence was "off-label use" not authorized by Iowa legislature, and further that "nothing in our record indicates the actuarial tools at issue here were designed to calculate risk-of-reoffending scores at an individual level or for sentencing purposes. Nothing in our record indicates the existence of validation studies for these tests or any cross validation for an Iowa population of offenders" (p. 9).
[r]For *In re Detention of Wygle* (2018), the court, in dictum, questioned the validity of the Static because of its small and foreign validation sample and its high number of false positives and false negatives.
[s]For *In re Commitment of Hood* (2016), the court rejected the claim that an expert failed to apply his own methodology, and held that claims that his testimony based on the Static was confusing and did not aid the factfinder were matters for the jury.
[t]In *Cannon v. Comm. of Soc. Sec.* (2018), the plaintiff's expert testified that the defendant's expert's "administration of the WAIS-IV was 'out of the standard of acceptable practice' because he 'leaves out some of the subtests'" (p. 4). The court rejected this challenge, stating that this disagreement merely explained why the experts had come to different conclusions.

by attorneys or their expert witnesses about how a tool was used (e.g., an argument that the tool was not administered consistent with the tool's manual) or about the proper interpretation of test results (what we call *diagnostic implementation*).[6] Very few raised claims about the general propositions of fit (does the tool serve to enlighten about the *kind of* problem at issue?), validity (does the tool measure what it purports to measure?), or helpfulness (does a tool which meets the fit and validity criteria nonetheless fail to assist the fact-finder or unfairly prejudice one of the parties?). As noted above, such questions have a trans-case nature, and therefore call for a decision by a judge, not a jury.

Frequency and type of successful 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). Most challenges were multifaceted, but the challenger's claim usually focused on fit, one or the other type of validity, or both (see Fig. 6). There were 8 fit challenges (4 succeeded), 13 *Daubert*-validity-factor challenges (5 succeeded),[7] 2 unhelpfulness/prejudicial-impact challenges (1 succeeded), and 1 inadequate-qualifications challenge (which did not succeed).[8] These findings are consistent with those reported in O'Brien's (2018) study of challenges to experts in civil rights cases, in which validity challenges likewise did not succeed as often as relevance/fit challenges. However, in contrast to O'Brien's finding that challenges to qualifications were as common

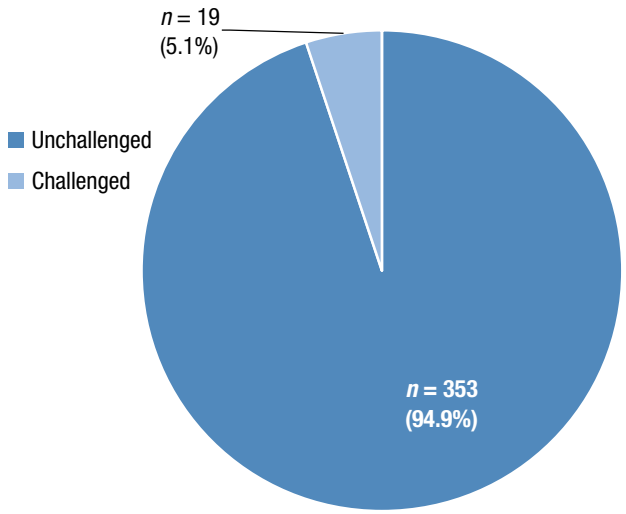

Fig. 5. Frequency of challenges to admissibility of psychological assessment tools in 372 cases involving such evidence. The number of challenged and unchallenged tools is shown.

as validity challenges, challenges to qualifications were much less common than validity challenges in our data.

In many of the cases in which the challenge was rejected, two factors appeared to influence the court's decision. First, where an appellate decision was involved, the court often explicitly deferred to the lower court's or jury's decision (see Table 5, cases cited in notes e, h, and t). Likewise, courts rejecting challenges sometimes stated that any objections to tool-based testimony should be about the weight it is accorded, not its admissibility. Second, even if the court believed the fit or scientific bona fides of a test were somewhat suspect, courts rejecting a challenge often emphasized

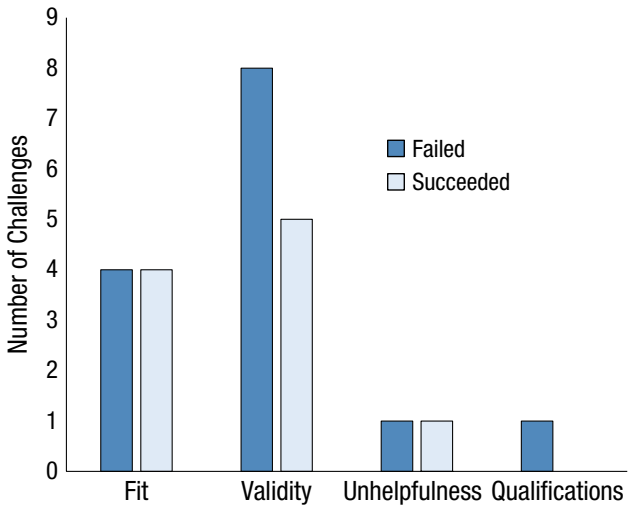

Fig. 6. Types of challenges to psychological assessment evidence in a sample of 372 cases. The number of challenges of each type is presented separately for successful and failed challenges.

that the test's results are only one "piece of the puzzle" in the expert's testimony and thus not dispositive (see Table 5, notes m and n).

Categorizing the type of challenge in terms of legal issues—fit, validity, helpfulness—is occasionally difficult, especially in terms of its scientific category—framework versus diagnostic and methodology versus substance (see, e.g., Table 5, notes a, i, and j). Fortunately, these distinctions may not be crucial if one subscribes to the Faigman et al. (2014, 2016) schema. Under that approach, the most important empirical distinction is between the first three categories, on the one hand, and diagnostic-implementation on the other, which is relatively easy. The validity of framework testimony, and the validity of any diagnostic methodology purportedly used by the expert, is an admissibility issue and should be judged using *Daubert*-like criteria. The issue of whether the expert properly applied a methodology and the persuasiveness of the results reached is a question for the jury, to which the judge should virtually always defer. This distinction between issues that should be decided as a gatekeeping matter and those that should usually go to the jury is most important for determining the proper judicial role in a case involving expert testimony.

The relationship between validity and admissibility challenges. An important question is whether weak tools tend to be challenged more often than scientifically stronger tools. Our evidence shows little relation between a tool's psychometric quality and it's likelihood of being challenged. Five of the nine challenged tools (MMPI, PAI, PCL, SIRS, WAIS) have favorable reviews and are generally accepted; two others (MCMI, Rorschach) have mixed reviews and debatable general acceptance; and the final two (ABCS, STATIC) lack reviews but are generally accepted (for acronym definitions; see Table 3).

Another way to look at the relationship between test validity and legal acceptance is to determine whether the worst of the tools (i.e., those that received unfavorable reviews and are not generally accepted) are especially likely to be challenged. Of the three tools fitting this description, none was challenged. The SIT/SFRIT was cited 15 times, but never challenged. The RISB was cited only once and the KFD was not cited even in a single case, and neither was subject to challenge. Likewise, tools with mixed reviews that are not generally accepted were rarely cited and were not challenged during the 3 years we sampled, and tools with generally unfavorable reviews and debated general acceptance were cited several times but never challenged. Those data suggest that some of the weakest tools tend to get a pass from the courts.

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

We have already discussed the implications of many of our findings. The focus here is on the major finding of this inquiry: the very low level of challenges to psychological tools. Given the questionable validity of many tools, one might have expected more legal controversy over their use. Several explanations for our contrary finding are possible.

First, for some types of cases that we investigated, evidentiary challenges are unlikely simply because prevailing law does not contemplate them. For instance, some of these tools are commonly used in legal contexts in which the rules of evidence do not apply or do not apply as rigidly, such as Social Security adjudications, sentencing hearings, and competency to stand trial assessments (see note 5). At the same time, most of the cases we investigated involved criminal or civil trials in which the rules of evidence do apply. Yet even in these cases challenges are rare.

A second possible explanation has to do with lack of expertise among lawyers and experts. 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. In addition, some experts may feel that, whatever problems afflict a particular test, the alternative (e.g., intuitive judgment unsupported by data) is worse.

Precedent can also play a significant role. If lawyers and experts have always used a particular tool without challenge, then a new challenge is not likely forthcoming. We looked only at cases for a 3-year period. It may be that the tools we examined were challenged in earlier cases. However, in only a few cases was such a previous challenge mentioned (e.g., Table 5, notes m and n). It is likely that many lawyers, experts, and courts simply follow the status quo (Fradella et al., 2003; Shuman, 2001). Finally, in some cases, the state legislature has mandated particular tools, which arguably requires their use regardless of any defects (e.g., Table 5, note i; cf. note a).

Even if lawyers or experts are aware of a tool's weaknesses, they may not push the issue if the tool is one of many bases for an expert's opinion. As noted, appellate courts often adopt the view that, if suspect evidence is merely a "piece of the puzzle," concerns about validity are undercut. Of course, continuing the puzzle analogy, this point of view can fail to consider the influence of the allegedly defective puzzle piece on the opinion as a whole and also fails to consider the validity of the other pieces. Arguably, the more responsible position is the fact that a particular test result is not dispositive does not eliminate concerns about validity (see Table 5, note q).

A less appealing explanation for the dearth of challenges is that lawyers are simply not willing to expend the resources necessary to mount a challenge. At least in cases involving disadvantaged clients (e.g., most criminal cases), lawyers often have little investigative resources or time, given the press of other cases. A *Daubert* challenge can be quite expensive because it usually requires locating the right expert or experts, paying them, and arranging for a separate pretrial hearing. It is possible that challenges are more likely in those few cases in which the financial stakes are high or the litigants are well-financed.

A final, more conceptual explanation for our results has to do with a basic tension between law and science. As Melton et al. (2017) have noted,

> The two disciplines do not conceptualize a "fact" in the same way. Although the sciences [including the science behind psychological assessment tools] are inherently probabilistic in their understanding of truth, the law demands at least the appearance of certainty, perhaps because of the magnitude and irrevocability of decisions that must be reached by law. . . . Because of the law's preference for certainty, experts may feel tempted to reach beyond legitimate interpretations of their data both to appear "expert" and to provide usable opinions. Similarly, legal decisionmakers may disregard testimony properly given in terms of probabilities as "speculative," and may attend instead to experts who express categorical opinions about what did or will happen. (pp. 11–12)

Because of the desire (and need) for certainty in law, lawyers and judges may prefer unstructured clinical judgments that directly address the individual case over psychological test results based on group data that only indirectly provide answers to legal questions. Of course, from a scientific perspective, structured assessments with tested or testable psychometric properties are superior to unstructured clinical judgments.

Limitations

The limitations of our methodology for drawing general conclusions about how courts treat psychological tools are fairly obvious. First, we looked at only 3 years'

worth of case law. Second, we examined what that case law said about psychological testimony only in connection with 30 psychological tools, out of at least 364 that are said to be used by clinicians in forensic contexts. Third, for the three tools we studied that were mentioned in the case law more than 30 times, we closely examined only a random subsample. Fourth, we made no effort to determine independently whether the courts' decisions were right or wrong under the relevant evidentiary standards, for a number of reasons, including the fact that in most of the cases the necessary information was simply not provided.

Limitations such as these were necessary to make the scope of the project manageable. O'Brien (2018), who carried out a similar study, limited his survey to only 2 years and looked solely at civil rights cases. The fact remains that we did not conduct a comprehensive survey of the case law regarding the admissibility of psychological tools; rather, we conducted a limited but organized investigation into a sample of legal cases citing a sample of psychological tools. Our methods provide us a rough nonparametric sense of the population of cases.

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.

Although our data for both parts of this project are available on the Open Science Framework (https://osf.io/qx75p/), we intentionally opted not to focus on individual psychological tests in the results or discussion sections of this article. This decision reflects our intention to provide a big-picture perspective of what is going on in the field (i.e., that psychologists are using some tools in court settings with questionable or poor psychometrics, and the courts are not doing a good job of recognizing it or holding psychologists accountable), rather than offering judgments about individual tools. 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.

Suggestions for psychological scientists and mental health practitioners

Consistent with recommendations from other sources (AERA, APA, & NCME, 2014; APA, 2013, 2017; Flake & Fried, 2019; Heilbrun et al., 2009), organized psychological science should create scientifically strong measures, inform experts about which tools are sound for which tasks, and discourage the use of tools with poor validity or that are unsuitable for the task at hand. Although judges and lawyers are the ultimate arbiters of when and how psychological tools are used in the courts, the onus is on psychology to create sound methodologies and teach its scientists and practitioners to use them. This role is particularly important in those contexts, such as Social Security determinations, in which the rules of evidence do not apply; in settings in which legal gatekeeping is minimal, professional experts have an even greater responsibility to ensure that their instruments are valid. As we have indicated throughout this article, helpful information on tool validity and reliability can be found in tool manuals, in the peer-reviewed literature, and in comprehensive review sources such as the *MMY*. The American Psychological Association's Testing and Assessment website (https://www.apa.org/science/programs/testing) also provides various resources, including frequently asked questions from professionals and the public about psychological testing, information about the *Standards* (AERA, APA, & NCME, 2014), reports from APA committees, and links to testing-related websites and resources.

A key implication of our findings, and one that is in accord with the *Standards*, 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. Moreover, one cannot assess the likely validity of a measure in a given context unless research has been performed previously in a similar context.

Our findings suggest a second important condition for using psychological tools in the legal setting: Their psychometric and context-relevant validation studies should survive scientific peer-review through an academic journal, ideally before publication in a manual. We offer this suggestion to encourage test developers and publishers to ensure their products have gone through rigorous testing and refinement that would pass muster under a *Daubert*-type analysis before bringing them to market. This is a guideline of our own

making and is not currently made by other sources. But given the law's special focus on evidentiary criteria, we think it is a fitting recommendation for selection of psychological testing methods in forensic contexts.

A third means of addressing the import of our findings is to develop consensus-based protocols in research incorporating the best measures available. That is, the scientific knowledge base in forensic psychology could improve if scientists working in forensically relevant contexts made the same measurement choices across different studies toward a more cumulative science. A potential starting point for finding measures is the APA resource PsychTESTS, launched in 2011, which includes 55,000 records of psychological measures, scales, surveys, and other instruments developed for research but not made commercially available (https://www.apa.org/pubs/databases/psyctests). This site serves as a professionally indexed resource for researchers looking for psychological measurement and instrumentation tools and information about their psychometric properties. However, as Flake and Fried (2019) have noted, the wide array of options available offers a stunning number of opportunities for measurement flexibility, with problematic implications for the validity and reliability of conclusions based on such measures. And the problem of measurement in clinical and forensic psychology we are highlighting is by no means solved in the social and personality psychology literature (Flake, Pek, & Hehman, 2017; Fried, 2017; Hussey & Hughes, 2019).

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. The PhenX Toolkit offers free, scientific consensus-based measures for research, including mental health research (Hamilton et al., 2011). The toolkit offers well-established measurement protocols selected by groups of domain experts to help researchers choose protocols likely to be used in additional research, thus facilitating cross-study analysis and strengthening the scientific impact of individual studies. A new forensic mental health-specific platform like this could focus not only on advancing the *measurement science* in the field, but perhaps also on clinical *practice*, sharing a platform for effective cross-pollination of information. Because it would be free and available online, it could also serve as an important source of information for the public about what tools are identified by domain experts as the best available for given types of referral questions.

More specifically, standard batteries for assessment practice could be based on the best clinical tools available. 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 about depriving people of liberty on the basis of proprietary, closed-source assessment tools (e.g., *Wisconsin v. Loomis*, 2016). Standard batteries for various referral questions could be evaluated for validity and interclinician reliability in the field. 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. If the field moved toward standardized batteries, it could open the door for many process-related improvements that could reduce bias and error in forensic psychological assessments.

Suggestions for lawyers and judges for evaluating psychological assessment evidence

A major theme of evidence scholars who write about expert evidence admissibility is the poor job judges do of implementing the dictates of *Daubert*. Although judges generally try to adhere to *Daubert*-like rules, they often struggle and sometimes end up ignoring them entirely (Dahir et al., 2005; Faigman et al., 2018; Gatowski et al., 2001; Groscup et al., 2002). Our data are consistent with these observations. General solutions that have been proposed to deal with this problem include tightening up Federal Rule of Evidence 702, to increase judicial compliance and provide clearer guidance about what factors judges must consider in making admissibility determinations (see e.g., Bernstein & Lasker, 2015). General guidance for judges and lawyers for assessing the validity of experts' methodology is available in Faigman and colleagues' *Modern Scientific Evidence* treatise (Faigman et al., 2018). In addition to recommending *Modern Scientific Evidence* as a resource, below we provide attorneys with more suggestions in connection with the use of psychological tools in the courtroom.

Suppose that attorneys know that the opposing side is likely to rely on psychological assessment tools. Once they understand the case and the part that might be played by the tools, they will need to decide whether to raise a challenge to the evidence. The possibility that the other side's tools are suspect should not automatically trigger a challenge. The attorneys must consider the likelihood that a challenge will be successful, the resources and expertise needed to mount such a challenge, and whether a successful challenge will make a difference to the outcome of the case. In addition, the attorneys may decide—particularly in proceedings in which the rules of evidence do not apply—that any

challenge that occurs will aim to minimize the weight that the fact-finder accords the evidence rather than to exclude the evidence outright.

Lack of resources is often a problem in legal settings. Aside from buying an expert lunch and soliciting "free" advice, low-cost help might be available from a good online resources. Attorneys and judges can begin by looking at sources such as the APA's testing and assessment information website (https://www.apa.org/science/programs/testing) for basic information. In addition, for a small fee, *Test Reviews Online* (https://marketplace.unl.edu/buros/), a web-based service of the Buros Center for Testing, provides published test information and reviews from the *MMY*. Here, users may download information for any of over 3,500 tests that include information about the test's purpose, appropriate populations, score ranges, publication date, admission time, and critical reviews of the tests including review of the technical quality of the tests written by independent experts.

A search of the primary published literature can be accomplished through sources such as Google Scholar or through library search engines. Some published articles in the literature are open access (i.e., available for free), but most are available for purchase. After searching for evidence about the test it should be clear, of course, that if there is no manual and no evidence of testing or peer review in sources such as the *MMY* or in the primary literature, a tool should not be used (and thus it should be challenged if it has been used by an expert).

In analyzing information discovered about a tool, lawyers and judges should begin with the foundation of the expert's testimony. In other words, they should consider the testimony of mental health experts as they would—or, at least, should—approach any expert-opinion evidence that is ostensibly based in science. One can think of this task as involving two basic questions relevant to admissibility (Faigman et al., 2014; Faigman et al., 2016). The first has to do with determining whether the general hypothesized phenomenon exists. Scientists might hypothesize, for example, that extreme trauma causes specific psychological consequences, or that certain people pose a serious risk of future violence. Complex areas of science typically involve many different research teams, different disciplines, and a large body of published research. All of this work is necessary to answer the most preliminary question of all: Does the hypothesized relationship of interest exist?

The next step is to make this information relevant to the individual case at hand. Focusing on the trauma example, it is important to know whether and how trauma causes psychological sequelae, but the ultimate legal issue is whether trauma caused the plaintiff's symptoms. Many people never exposed to trauma develop anxiety, and many people exposed to it never develop anxiety. The challenge in most contexts in which the law uses scientific research is to determine whether a particular case is an instance of the general phenomenon. This determination requires some methodology or test that can be applied to individual cases.

Expert testimony based on psychological tools raises the same two issues. Suppose, for example, a defendant accused of murder claims that a brain injury he suffered a decade before affected his behavior at the time of the alleged crime.[9] The defendant seeks to call a board-certified neuropsychologist who would testify about the psychological and cognitive effects of the defendant's brain injury and trauma stemming from the effects of a gunshot wound he received 10 years earlier. Among the battery of tests the expert administered was the Personality Assessment Inventory (PAI). According to the expert, the PAI supported the following opinion:

> Given the residual cognitive and psychological effects of his Traumatic Brain Injury, under such conditions of chaos and stress [Defendant] would be more likely to perceive himself to be facing an imminent threat and have greater difficulty controlling his reactions. (*Savage v. State*, 2017, p. 187)

Similar to other scientific claims that make it into court, the initial issue presented by this case is whether the claimed phenomenon exists as a general matter, to wit: Does research support the hypothesis that traumatic brain injury can lead to a person perceiving imminent harm more immediately than someone without such trauma? If the answer is no, that ends the inquiry and any expert evidence should be summarily excluded.

If such a phenomenon exists, however, the next question is whether a method exists to identify whether a particular defendant is an instance of this phenomenon. In our admittedly simplified example, this question concerns the PAI's ability to serve this function. Specifically, the PAI as a psychological tool would need to be shown to be valid either for the specific legal question presented, or at least for some subsidiary inference that is relevant to that question. For instance, a test that reliably and validly identifies hypervigilance might be relevant to an ultimate diagnosis in a particular case.

Suggestions for members of the public interacting with psychologists in the legal system

Psychologists have an ethical responsibility to provide the people they evaluate with information about the evaluation process through an informed-consent

procedure, and in most settings, they have an additional ethical duty to provide the evaluee with results of the evaluation (APA, 2013). These pre- and postevaluation meetings present people who are evaluated opportunities to discuss with psychologists the assessment methods and the strengths and weaknesses of the tools used in the assessment process. If litigants or others would like more information about these tools, then, like attorneys and judges,[10] they can look to sources such as the APA's testing and assessment information website for basic information about tests psychologists use, Test Reviews Online or the *MMY* for reviews of specific tests, and sources such as Google Scholar or library electronic databases for published information about specific tests.

Members of the public who gain knowledge about psychological tools from the expert or outside sources, or who are skeptical of the methods used by psychologists in assessment procedures, can share that information with attorneys during the legal process. Attorneys have an ethical responsibility to advocate for their clients' interests. As this article shows, attorneys may underappreciate the limitations of psychological-assessment evidence. Members of the public involved in cases in which psychological-assessment evidence is used might discuss with their attorneys and ask whether a challenge to psychological-expert evidence is feasible.

Implications for other types of psychological assessments in high-stakes contexts

Although this article focuses on clinical assessments, our findings are potentially relevant for a much broader swath of expert psychological test evidence. The *Standards* (AERA, APA, & NCME, 2014) apply in four different testing application contexts: psychological testing and assessment, workplace testing and credentialing, educational testing and assessment, and the use of tests for program evaluation, policy studies, and accountability. This project has focused exclusively on just the first—psychological testing and assessment—but other testing applications raise issues similar to those we have discussed.

For example, there are rich intersections with the law in workplace testing and credentialing (e.g., civil-service measures, personnel-selection procedures, licensing tests) and educational testing and assessment (e.g., admissions testing, disability accommodations, performance assessment). In industrial testing, the focus has been primarily on fairness toward underrepresented groups for tests used in personnel selection and promotion. A few cases involving personnel testing have reached the U.S. Supreme Court (e.g., *Griggs v. Duke Power Co.*, 1971; *Guardians Association of New York City v. Civil Service Commission*, 1983), and important regulations set by federal agencies such as the Equal Employment Opportunity Commission (1978) further enforce civil rights in personnel selection. Cases in education have also often focused on whether psychological testing has produced systematic ethnic group and gender differences (e.g., *Debra P. v. Turlington*, 1981; *Larry P. v. Riles*, 1984; *PASE v. Hannon*, 1980).

These cases highlight the challenges of creating unbiased assessment instruments. Even psychometrically strong tools can lead to systematic disparate group outcomes depending on legally protected characteristics such as race and gender (Bersoff, 1981; Fincher, 1973). These problems are echoed in the current controversy about racial bias in machine learning and algorithmic approaches for predicting risk and recidivism (e.g., Berk, 2012; Skeem & Lowenkamp, 2016).

Conclusion

We investigated the scientific quality of assessment methods used by psychologists in legal cases and the extent to which courts are attuned to such quality. We found significant weaknesses in both the methods that psychologists use to address legal issues and the way courts assess those methods. We hope these findings will encourage (a) psychological scientists to improve the state of knowledge in the field, public access to information about psychological tests to facilitate critical review, and the availability of high-quality, nonproprietary measures; (b) mental health practitioners to be more discerning in the choices they make with regard to assessment tools used in forensic cases; (c) attorneys to challenge and judges to scrutinize psychological-assessment experts more frequently (using resources such as those described in this article); and (d) members of the public to be skeptical, to ask questions, and to take advantage of the adversarial process. Combined, these efforts should go a long way toward producing the highest quality of practice from psychological experts involved in legal cases.

Acknowledgments

Thanks to the Association for Psychological Science and the School of Social & Behavioral Sciences at Arizona State University (ASU) for funding a working conference for this project in Phoenix, Arizona, in January 2019, and to Gloria Sawrey for helping coordinate the meeting. Thanks to Scott O. Lilienfeld (Emory University) and Thomas Grisso (Emeritus, University of Massachusetts Medical School) for their advice about this project, and to Janet Carlson (Buros Center for Testing) for providing useful information about the operational functioning of the Buros Center. Thanks to Christopher King (Montclair State University), who compiled the list of

364 tools from the 22 surveys in the literature; to Zachary Graham for writing a program to automatically search through various texts for the names and acronyms of our list of tools; and to Emily Line for assistance with some of the figures. Carina Philipp, Hannah Goddard, and Ashley Jones helped significantly with the initial round of coding that inspired this project. Many ASU students participated in the "codeathons" to code the psychological assessment data for this article: Brianna Bailey, Bethany Baker, Rex Balanquit, Samantha Bean, Li-Hsin Chen, Veronica Cota, Jacey Cruz, Emily Denne, Renee El-krab, Emily Fatula, Annanicole Fine, Carly Giffin, Emily Line, Nicole Lobo, Laura Malouf, Elizabeth Mathers, Kristen McCowan, Robin Milligan, Olivia Miske, Daisy Ornelas, Jake Plantz, Selena Quiroz, Stephanie Rincon, Samantha Roberts, Kaitlyn Schodt, Hayley Seely, Karima Shehadeah, Stephanie Thibault, Annelisse Velazquez, and Liu Xingyu. And several law-student research assistants helped with coding the legal data: Kevin Bohm, Dora Duru, Kasey Galantich, and Sarah Pook. Portions of these results were presented at the 2018 annual conference of the American Psychology-Law Society (AP-LS) in Memphis, Tennessee, and the 2019 annual convention of the American Psychological Association (APA) in Chicago, Illinois. These results will be presented at the 2020 annual meeting of the American Association for the Advancement of Science (AAAS) in Seattle, Washington, and at the 2020 annual AP-LS conference in New Orleans, Louisiana.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. We did not code for "error rate," the fourth *Daubert* factor, for practical reasons. All of the tools we investigate are testable and thus can, in theory, generate error rates about something. But whether the error rates are legally relevant varies significantly depending on context, such as the type of case, type of referral question, and facts of the case, even within a particular subject area such as insanity (in terms of type and degree of impairment that must be shown) or risk assessment (in terms of probability and type of risk that must be shown).

2. Archer et al. (2006) surveyed 152 doctoral-level psychologists who were members of the American Psychological Association Division 41 (American Psychology-Law Society) and/or diplomates of the American Academy of Forensic Psychology. Elhai et al. (2005) surveyed 277 members of the International Society for Traumatic Stress Studies, 75.3% of whom were doctoral-level clinicians and 23% of whom were masters'-level clinicians; the sample had an average of 11.7 years (SD = 9.3) of experience in the field. LaDuke et al. (2018) surveyed 502 doctoral-level psychologists active in the practice of neuropsychology and members of the National Academy of Neuropsychology and/or International Neuropsychological Society. A little more than half of the LaDuke et al. sample (n = 255, 50.8%) reported engaging in forensic practice, 30% of whom (n = 77) were board certified by a neuropsychology-related board. LaDuke et al. selected only the board-certified forensic psychologists to respond to items about test use to inform general acceptance in the field. Lally (2003) surveyed 64 doctoral-level psychologists, all of whom were board-certified by the American Board of Forensic Psychology. McLaughlin and Kan (2014) surveyed 102 doctoral-level forensic evaluators in professional psychological practice, 16.7% of whom were board-certified by the American Board of Forensic Psychology, who had on average 14.24 years (SD = 11.60) of experience in the field. Neal and Grisso (2014) surveyed 434 experts, 91% of whom were doctoral-level clinicians with an average of more than 16 years of experience, 16.4% of whom were board-certified. Rabin et al. (2005) surveyed 747 doctoral-level psychologists affiliated with Division 40 of the American Psychology Association, the National Academy of Neuropsychology, or the International Neuropsychological Society, 22% of whom were board certified in neuropsychology. Forensic practice was common in the Rabin et al. sample; 32% of the participants were involved in conducting neuropsychological assessments for the court. Ryba et al. (2003b) surveyed 82 doctoral-level clinicians with expertise in juvenile competency evaluations for the courts. Slick et al. (2004) surveyed 24 doctoral-level neuropsychologists (54% board-certified) with expertise in handling financial compensation claims or personal injury litigation cases. To be identified as an expert by Slick et al., neuropsychologists had to have published at least 2 peer-reviewed articles on methods for detecting malingering or suboptimal performance in the 5-year period preceding the survey, and they had to have seen a litigation or compensation-seeking case in the 12 months preceding the survey. It is possible that some of the participants in these studies participated in more than one of the surveys, and thus the total sample of mental health experts could be somewhat lower than the total of 2,384 we cited in the text.

3. The *MMY* publishes a list of tests that the *MMY* requested for review but not receive and thus did not review. This list is published as an index in every edition of the *MMY*.

4. Although we present the methods and results later in this article, some of the results are reported in Table 3 for efficiency purposes so that we did not have to duplicate a table full of these 30 tools with new columns again later in the article.

5. In several cases, the issues arose in Social Security cases or sentencing hearings. The rules of evidence do not apply in these settings. This point could help explain the failure of lawyers to challenge the admissibility of the tests and the judges' failure to exclude them. Nonetheless, even when a particular rule of evidence does not apply, an expert's use of invalid methods might be expected to rise to the surface of a court's opinion, either as a matter of due process or what weight to accord the evidence. Hence, even absent a formal *Daubert*-like challenge, we should expect to see discussion of the scientific merits of the tests used by experts.

6. Most of these discussions were not about a challenge to the admissibility of testimony but rather mere a description of the adversarial process at work (as Table 5 shows, only two categorized challenges involved a question of whether a rational jury could agree with the expert's diagnostic conclusion).

7. For this purpose, we combined the two types of validity challenges.

8. The number of challenges (24) exceeds the number of cases in which challenges occurred (19) because lawyers advanced multiple challenges in some cases. For the same reason,

challenges on specific grounds (e.g., fit or validity) exceed the number of cases in which a tool was successfully challenged. Thus, the percentage of challenges succeeded is higher when considering their success at the challenge level (i.e., 10 of 24 challenges succeeded, 42%) compared with the case level (6 of 19 challenged cases were successful, 32%).

9. The example used here is based generally on the facts of *Savage v. State* (2017).

10. See suggestions for attorneys and judges in the previous section.

References

Abel, G. G., Becker, J. V., & Cunningham-Rathner, J. (1984). Complications, consent, and cognitions in sex between children and adults. *International Journal of Law and Psychiatry*, *7*, 89–103.

Ackerman, M. J. (2005). The Ackerman-Schoendorf Scales for Parent Evaluation of Custody (ASPECT): A review of research and update. *Journal of Child Custody*, *2*, 179–193.

Ackerman, M. J., & Ackerman, M. C. (1997a). Child custody evaluation practices: A 1996 survey of psychologists. *Family Law Quarterly*, *30*, 525–586.

Ackerman, M. J., & Ackerman, M. C. (1997b). Child evaluation practices: A survey of experienced professionals (revised). *Professional Psychology: Research and Practice*, *28*, 137–145.

Ackerman, M. J., Ackerman, M. C., Steffen, L. J., & Kelley-Poulos, S. (2004). Psychologists' practices compared to the expectations of family law judges and attorneys in child custody cases. *Journal of Child Custody*, *1*, 41–60.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382. doi:10.1177/0011000005285875

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*, 32–50.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association Press.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist*, *68*, 7–19. doi:10.1037/a0029889

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Retrieved from https://www.apa.org/ethics/code/

American Psychological Association. (2019). *FAQ: Finding information about psychological tests*. Available from https://www.apa.org/science/programs/testing/find-tests

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment*, *87*, 84–94. doi:10.1207/s15327752jpa8701_07

Bechara, A. (2016). *Iowa Gambling Task, Version 2. Professional manual*. Lutz, FL: Psychological Assessment Resources.

Bell, S., Sah, S., Albright, T. D., Gates, S. J., Denton, M. B., & Casadevall, A. (2018). A call for more science in forensic science. *Proceedings of the National Academy of Sciences, USA*, *115*, 4541–4544. doi:10.1073/pnas.1712161115

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring and interpretation*. Minneapolis: University of Minnesota Press.

Berk, R. A. (2012.). *Criminal justice forecasts of risk: A machine learning approach*. New York, NY: Springer.

Bernstein, D. E., & Lasker, E. G. (2015). Defending Daubert. It's time to amend Federal Rule of Evidence 702. *William & Mary Law Review*, *57*, 1–48.

Bersoff, D. N. (1981). Testing and the law. *American Psychologist*, *36*, 1047–1056.

Boccaccini, M. T., & Brodsky, S. L. (1999). Diagnostic test usage by forensic psychologists in emotional injury cases. *Professional Psychology: Research and Practice*, *30*, 253–259.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *451*, 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061

Borum, R., & Grisso, T. (1995). Psychological test use in criminal forensic evaluations. *Professional Psychology: Research and Practice*, *26*, 465–473.

Bow, J. N., & Quinnell, F. A. (2001). Psychologists' current practices and procedures in child custody evaluations: Five years after American Psychological Association Guidelines. *Professional Psychology: Research and Practice*, *32*, 261–268.

Briere, J. (2011). *Trauma Symptom Inventory-2*. Odessa, FL: Psychological Assessment Resources.

Buros, O. K. (Ed.). (1938). *Mental measurements yearbook*. New Brunswick, NJ: Rutgers University Press.

Burt, M. R. (1980). Cultural myths and supports for rape. *Journal of Personality and Social Psychology*, *38*, 217–230.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration and scoring* (Rev. ed.). Minneapolis: University of Minnesota Press.

Camara, W., Nathan, J., & Puente, A. (1998). Psychological test usage in professional psychology: Report to the APA practice and science directorates. Washington, DC: American Psychological Association.

Cannon v. Commissioner of Social Security, WL 6919311 (D. Alaska 2018)

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Cizek, G. J., Koons, H. K., & Rosenberg, S. L. (2012). Finding validity evidence: An analysis using *Mental Measurements*

Yearbook. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp 119–138). Washington, DC: American Psychological Association.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319. doi:10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*. *31*(12), 1412–1427. doi:10.1037/pas0000626

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., & Meehl, P. H. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Dahir, V. B., Richardson, J. T., Ginsburg, G. P., Gatowski, S. I., Dobbin, S. A., & Merlino, M. L. (2005). Judicial application of Daubert to psychological syndrome and profile evidence: A research note. *Psychology, Public Policy, and Law*, *11*, 62–82. doi:10.1037/1076-8971.11.1.62

Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). 509 U.S. 579, 113 S. Ct. 2786, 125 L. Ed. 2d 469.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Debra P. v. Turlington, 644 F.2d 397 (5th Cir. 1981)

Director of Public Prosecutions for Western Australia v. Mangolamara. (2007). W.A.S.*C*. 71.

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence—User guide*. Burnaby, British Columbia, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Edens, J. F., & Boccaccini, M. T. (Eds.). (2017). Field reliability and validity of forensic psychological assessment instruments and procedures [Special issue]. *Psychological Assessment*, 29(6).

Eichenberger v. Shulkin, No. 16-0252, 2017 WL 2457095 (Vet. App. June 7, 2017).

Elhai, J. D., Gray, M. J., Kashdan, T. B., & Franklin, L. C. (2005). Which instruments are most commonly used to assess traumatic event exposure and posttraumatic effects? A survey of traumatic stress professionals. *Journal of Traumatic Stress*, *18*, 541–545. doi:10.1002/jts.20062

Embretson, S. E. (1995). The new rules of measurement. *Psychological Assessment*, *8*, 341–349.

Epps, J. A., & Todorow, K. (2019). Refryed forensics: Screening expert testimony in criminal cases through Frye plus reliability. *Seton Hall Law Review*, *48*, 1161–1198.

Equal Employment Opportunity Commission. (1978). Office of Personnel Management, Civil Service Commission, Department of Labor, & Department of Justice. Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register*, *43*, 38390–38315.

Ewert v. Canada, 2018 SCC 30 (2018, June 13).

Exner, J. E., Jr., & Erdberg, P. (2005). *The Rorschach: A comprehensive system*. Hoboken, NJ: John Wiley & Sons.

Faigman, D. L., Cheng, E. K., Mnookin, J., Murphy, E. E., Sanders, J., & Slobogin, C. (2018). *Modern scientific evidence: The law and science of expert testimony*, 2018-2019 ed. Eagan, MN: Thomson West.

Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review*, *81*, 417–480.

Faigman, D. L., Slobogin, C., & Monahan, J. (2016). Gatekeeping science: Using the structure of scientific research to distinguish between admissibility and weight in the expert testimony. *Northwestern University Law Review*, *110*, 859–904.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science*, *241*, 31–35.

Fed. R. Evid. 401. (2019). Test for relevant evidence. Retrieved from https://www.rulesofevidence.org/article-iv/rule-401/

Fed. R. Evid. 702. (2019). Testimony by expert witnesses. Retrieved from https://www.rulesofevidence.org/article-iv/rule-702/

Fincher, C. (1973). Personnel testing and public policy. *American Psychologist*, *28*, 489–497. doi:10.1037/h0035195

Flake, J. K., & Fried, E. I. (2019, March 15). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv*. doi:10.31234/osf.io/hs7wm

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. doi:10.1177/1948550617693063

Fradella, H. F., Fogarty, A., & O'Neill, L. (2003). The impact of Daubert on the admissibility of behavioral science testimony. *Pepperdine Law Review*, *30*, 403–444.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. doi:10.1016/j.jad.2016.10.019

Frye v. U.S. (1923). 293 F. 1013. (D.C. Cir. 1923).

Furnham, A. (2018). The great divide: Academic versus practitioner criteria for psychometric test choice. *Journal of Personality Assessment*, *100*, 498–506. doi:10.1080/00223891.2018.1488134

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology*, *2*, 149–162.

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-Daubert world. *Law and Human Behavior*, *25*, 433–458. doi:10.1023/A:1012899030937

General Electric Co. v. Joiner, 522 U.S. 136 (1997).

Giannelli, P. C. (1981). The admissibility of novel scientific evidence: Frye v. U.S., a half-century later. *Columbia Law Review*, *80*, 1197–1250.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).
Grisso, T. (2003). *Evaluating competencies: Forensic assessments and instruments* (2nd ed.). New York, NY: Kluwer Academic.
Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills*, *44*, 367–373.
Groscup, J. L., Penrod, S. D., Studebaker, C. A., Huss, M. T., & O'Neil, K. M. (2002). The effects of Daubert on the admissibility of expert testimony in state and federal criminal cases. *Psychology, Public Policy, and Law*, *8*, 339–372. doi:10.1037//1076-8971.8.4.339
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19. doi:10.1037/1040-3590.12.1.19
Guardians Association of New York City v. Civil Service Commission. (1983). 463 U.*S.* 582.
Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: McGraw-Hill.
Hall v. Florida, 572 U.S.; 134 S. Ct. 1986 (2014).
Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., . . . Nettles, D. S. (2011). The PhenX toolkit: Get the most from your measures. *American Journal of Epidemiology*, *174*, 253–260.
Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119–136.
Hare, R. D. (1991). *Hare Psychopathy Checklist-Revised: 2nd Edition*. Toronto, ON, Canada: Multi-Health System Assessments.
Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. doi:10.1080/19312450709336664
Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtis, G. (1993). *Wisconsin Card Sorting Test: Revised and Expanded*. Odessa, FL: Psychological Assessment Resources.
Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior*, *16*, 257–272.
Heilbrun, K. (1995). Child custody evaluation: Critically assessing mental health experts and psychological tests. *Family Law Quarterly*, *29*, 63–78.
Heilbrun, K., Grisso, T., & Goldstein, A. (2008). *Foundations of forensic mental health assessment*. New York, NY: Oxford University Press.
Hopey v. Spear, WL 4446452 (C.D. Ill. 2016)
Hurskin v. Commissioner of Social Security, 2016 WL 825538.
Hussey, I., & Hughes, S. (2019, June 4). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *PsyArXiv*. doi:10.31234/osf.io/7rbfp
In re Commitment of Gomez, 534 S.W.3d 917 (Tex. App. 2017)
In re Commitment of Hood, WL 4247961 (Tex. App. 2016)
In re Commitment of Sternadel, WL 1802151 (Tex. App. 2018)
In re Dayana J. No. H14CP14011094A, 2016 WL 4497613 (Conn. Super. *Ct.* July 18, 2016).
In re Detention of Wygle, 910 N.W.2d 599 (Iowa 2018)
In the Matter of Kristek, 383 P.3d 183 (Kan. Ct. App., 2016).
In the Matter of Garcia, 414 P.3d 752 (Kan. Ct. App. 2018).
Innocence Project. (n.d.) *Overturning wrongful convictions involving misapplied forensics*. Retrieved from https://www.innocenceproject.org/overturning-wrongful-convictions-involving-flawed-forensics/
J.K.J. v. Polk County, WL 280827 (W.D. Wis. 2017)
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000
Kaufman, A. S., & Kaufman, N. L. (2014) *Kaufman Test of Educational Achievement, Third Edition* (KTEA 3). Minneapolis, MN: Pearson Assessments.
Keilin, W. G., & Bloom, L. J. (1986). Child custody evaluation practices: A survey of experienced professionals. *Professional Psychology: Research and Practice*, *17*, 338–346.
King, C., Wade, N., & Tilson, J. (2017, March). *Case law references as a big-picture snapshot of psychological test use in forensic mental health assessments*. Paper presented at the Annual Conference of the American Psychology-Law Society, Seattle, WA.
Knoff, H. M., & Prout, H. T. (1985). *Kinetic Drawing System for Family and School: A Handbook*. Los Angeles, CA: Western Psychological Services.
Kumho Tire Co. v. Carmichael. 526 U.S. 137 (1999).
LaDuke, C., Barr, W., Brodale, D. L., & Rabin, L. A. (2018). Toward generally accepted forensic assessment practices among clinical neuropsychologists: A survey of professional practice and common test use. *The Clinical Neuropsychologist*, *32*, 145–164. doi:10.1080/13854046.2017.1346711
LaFortune, K. A., & Carpenter, B. N. (1998). Custody evaluations: A survey of mental health professionals. *Behavioral Sciences and the Law*, *16*, 207–224.
Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice*, *34*, 491–498. doi:10.1037/0735-7028.34.5.491
Lambert, N. M. (1991). The crisis in measurement literacy in psychology and education. *Educational Psychologist*, *26*, 23–25.
Larry P. v. Riles, 343 F. Supp. 1306 (N.D. Cal. 1979), aff'd in part and rev'd in part, 792 F. 2d 969 (9th Cir. 1984).
Lees-Haley, P. R. (1992). Psychodiagnostic test usage by forensic psychologists. *American Journal of Forensic Psychology*, *10*, 25–30.
Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology*, *11*, 45–51.
Lefkowitz v. Ackerman, 2017 WL 4237068 (S.D. Ohio, 2017).
Lerner, B. (1971). The Supreme Court and the APA, AERA, and NCME test standards. *American Psychologist*, *33*, 915–919.
Lin, L., Christidis, P., & Stamm, K. (2017, September). Datapoint: A look at psychologists' specialty areas: News from APA's Center for Workforce Studies. *American Psychological Association Monitor on Psychology*, *48*, 15.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Lord, F. N., & Novick, M. (1968). *Statistical theories of mental tests*. New York, NY: Addison-Wesley.

Martin, M. A., Allan, A., & Allan, M. M. (2001). The use of psychological tests by Australian psychologists who do assessments for the courts. *Australian Journal of Psychology*, *53*, 77–82.

McLaughlin, J. L., & Kan, L. Y. (2014). Test usage in four common types of forensic mental health assessment. *Professional Psychology: Research and Practice*, *45*, 128–135. doi:10.1037/a0036318

Meier, S. T. (1993). Revitalizing the measurement curriculum: Four approaches for emphasis in graduate education. *American Psychologist*, *48*, 886–891.

Melton, G., Petrila, J., Poythress, N., Slobogin, C., Otto, R., Mossman, D., & Condie, L. (2017). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (4th ed.). New York, NY: Guilford Press.

Merenda, P. F. (1996). Note on the continuing decline of doctoral training in measurement. *Psychological Reports*, *78*, 321–322.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037//0003-066X.56.2.128

Miller, F. G., & Lazowski, L. E. (2016). *The Adult SASSI-4 Manual*. Springville, IN: SASSI Institute.

Millon, T., Grossman, S., & Millon, C. (2015). *Millon Clinical Multiaxial Inventory-IV*. Minneapolis, MN: Pearson Assessments.

Moore v. Texas. (2017). 137 S. Ct. 1039.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Murray, H. A. (1943). *Thematic apperception test*. Odessa, FL: Psychological Assessment Resources.

Naar, R. (1961). Testing in juvenile courts: A survey. *Journal of Clinical Psychology*, *17*, 210.

National Research Council. (2009). Strengthening forensic science in the United States: A path forward. Washington, DC: The National Academies Press. doi:10.17226/12589

Neal, T. M. S. (2018). Forensic psychology and correctional psychology: Distinct but related subfields of psychological science and practice. *American Psychologist*, *73*, 651–662. doi:10.1037/amp0000227

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior*, *41*, 1406–1421. doi:10.1177/0093854814548449

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

O'Brien, T. I. (2018). Beyond reliable: Challenging and deciding expert admissibility in U.S. civil courts. *Law, Probability, and Risk*, *17*, 29–45. doi:10.1093/lpr/mgx010

Otto, R. K., & Heilbrun, K. (2002). The practice of forensic psychology: A look toward the future in light of the past. *American Psychologist*, *57*, 5–18. doi:10.1037/0003-066X.57.1.5

PASE v. Hannon. 506 F. Supp. 831 (N. D. III. 1980).

People v. Fortin, 218 Cal.Rptr.3d 867 (Cal. Ct. App, 2nd Dist. 2017)

People v. Howard, WL 1023990 (Cal. Ct. App. 2018)

People v. Jing Hua Wu, WL 616744 (Cal. Ct. App, 2016)

Pinkerman, J. E., Haynes, J. P., & Keiser, T. (1993). Characteristics of psychological practice in juvenile court clinics. *American Journal of Forensic Psychology*, *11*, 3–12.

Plake, B. S., Conoley, J. C., Kramer, J. J., & Murphy, L. U. (1991). The Buros Institute of mental measurements: Commitment to the tradition of excellence. *Journal of Counseling & Development*, *69*, 449–455.

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Quinnell, F. A., & Bow, J. N. (2001). Psychological tests used in child custody evaluations. *Behavioral Sciences and the Law*, *19*, 491–501.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*, 33–65. doi:10.1016/j.acn.2004.02.005

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reaes v. City of Bridgeport, WL 553380 (D.Conn. 2017)

Rogers, R. (1984). *Rogers' Criminal Assessment Scale (R-CRAS) and test manual*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., & Cavanaugh, J. L. (1984). Usefulness of the Rorschach: A survey of forensic psychiatrists. *The Journal of Psychiatry & Law*, *11*, 55–67.

Rogers, R., Sewell, K. W., & Gillard, N. (2010). *Structured Interview of Reported Symptoms-2 (SIRS-2) and Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Rotter, J. B., Lah, M. I., & Rafferty, J. E. (1992) *Rotter Incomplete Sentences Blank manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Ryba, N. L., Cooper, V. G., & Zapf, P. A. (2003a). Assessment of maturity in juvenile competency to stand trial evaluations: A survey of practitioners. *Journal of Forensic Psychology Practice*, *3*, 23–45.

Ryba, N. L., Cooper, V. G., & Zapf, P. A. (2003b). Juvenile competence to stand trial evaluations: A survey of current practices and test usage among psychologists. *Professional Psychology: Research and Practice*, *34*, 499–507. doi:10.1037/0735-7028.34.5.499

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*, 892–895. doi:10.1126/science.1111565

Savage v. State, 166 A.3d 183, 201 (Md. 2017)

Schmand, B., Lindeboom, J., Merten, T., & Millis, S. R. (2005). *Amsterdam Short-Term Memory Test: Manual*. Leiden, The Netherlands: PITS.

Sheehan, D. V., & Lecrubier, Y. (1997). *Mini-International Neuropsychiatric Interview (MINI)*. Lyon, France: Mapi Research Trust.

Shuman, D. (2001). Expertise in law, medicine and health care. *Journal of Health Policy & Law*, *26*, 267–293.

Silverton, L., & Gruber, C. (1998). *Malingering Probability Scale (MPS)*. Los Angeles, CA: Western Psychological Services.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*, 680–712.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. London, England: Palgrave Macmillan.

Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: A survey of experts' practices. *Archives of Clinical Neuropsychology*, *19*, 465–473. doi:10.1016/j.acn.2003.04.001

Slosson, R. L., Nicholson, C. L., & Hibpshman, T. H. (1991). *Slosson Intelligence Test, Revised (SIT-R3)*. Austin, TX: Slosson Education Publications.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

State v. Gary K., 46 N.Y.S.3d 477 (N.Y. Sup. Ct. 2016)

State v. Gordon, 919 N.W.2d 635 (Iowa Ct. App.), vacated on other grounds, 921 N.W.2d 19 (Iowa S. Ct. 2018)

Strauss, E., Sherman, E., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). Oxford, England: Oxford University Press.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25. doi:10.1146/annurev.clinpsy.032408.153639

Tardif v. City of New York, 344 F.Supp.3d 579, 600-601 (S.D.N.Y. 2018)

Tollison, C. D., & Langley, J. C. (1995). *Pain Patient Profile manual*. Minneapolis, MN: National Computer Systems, Inc.

United States v. Gamble, No. 2:07CR219-MHT, 2018 WL 3812241 (M.D. Ala. Aug. 10, 2018)

United States v. Jones, WL 1115778 (S.D.N.Y. 2018)

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *Journal of Personality Assessment*, *100*, 53–67. doi:10.1080/00223891.2017.1296455

Warren, W. L. (1994) *Revised Hamilton Rating Scale for Depression*. Los Angeles, CA: Western Psychological Services. https://www.wpspublish.com

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of Psychology. Volume 10: Assessment Psychology* (2nd ed., pp. 50–81). Hoboken, NJ: John Wiley & Sons.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2009). *Advanced Clinical Solutions for WAIS-IV and WMS-IV*. San Antonio, TX: The Psychological Corporation.

Widows, M. R., & Smith, G. P. (2005). *Structured Inventory of Malingered Symptomatology professional manual*. Odessa, FL: Psychological Assessment Resources.

Wisconsin v. Loomis, 881 N.W.2d 749 (2016).