Commentary: Evidence-based Assessment is not Evidence–based Medicine—Commentary on Evidence-based Assessment of Cognitive Functioning in Pediatric Psychology

Eric Youngstrom
Departments of Psychology and Psychiatry, University of North Carolina, Chapel Hill

The authors are commended on a well-done review of instruments to assess cognitive functioning in pediatric psychology. The review follows the framework of a psychological assessment review, and it provides a concise compendium with good coverage of available tests and a compilation of their psychometric “vital statistics.” The article also goes beyond the tropes of routine review by identifying some themes and issues such as the potential value of having disorder specific (or other subgroup specific) norms.

The choice of the words “evidence based” in the title invokes many associations, several perhaps unintended by the authors. For me, the phrase cued expectations developed from reading and teaching in an “Evidence-based Medicine” (EBM) framework (Guyatt & Rennie, 2002; Straus, Richardson, Glasziou, & Haynes, 2005). In this instance, my assumption was mistaken. However, as I read through the review, the mistake seems fortunate: it made me compare the methods we use to evaluate and describe evidence within the field of psychological assessment to the language and perspective that is evolving within EBM. Illuminating the discrepancies reveals many gaps that could be filled in ways that advance clinical practice as well as research.

The structure of this commentary is as follows: first, I will address the numerous strengths of the review. There is much that is praiseworthy. Second, I will touch on the themes of short forms, retest intervals, and multigroup norms. These are important issues, and I believe that they have even more clinical relevance than indicated in the review. It is also likely that my recommendations may be different than what the authors would have provided if given the opportunity to expand their discussion of these issues. Third, I will briefly introduce EBM and sketch out some of the differences I see between EBA—the current approach within psychology—and EBM. Finally, I will close with suggestions for next steps in both clinical practice and research to assimilate some of the best features of both approaches.

Strengths

A challenge that confronts any would-be reviewer is deciding what to include within the scope of the review, and how to choose what falls out of bounds. If the project were a clinical trial, investigators would specify “inclusion and exclusion criteria.” Meta-analysts have adopted similar methods, carefully delineating the search terms used and databases queried. In the present review, the authors opted for a different yet still empirical approach: they surveyed test users directly, and asked them what instruments they used. The review then concentrated on the tests that were most widely employed in practice at the time of the survey. This search strategy focuses attention on the current standard of practice. It guarantees that the critiques will be meaningful to a large portion of practitioners, and will cover familiar instruments. This approach to defining inclusion criteria is also somewhat conservative, especially compared to alternatives such as preferentially reviewing the newest tests or the most aggressively marketed tests. Thus, the chosen selection process avoids common sources of bias that influence clinical practice in other areas, such as when marketing and free samples appear to alter medication prescription decisions (Applbaum, 2006).

A second major strength is that the review explicitly referred to principles and issues drawn from the young but growing literature on “evidence-based assessment” (EBA) in psychology (Hunsley & Mash, 2005; Mash & Hunsley, 2005). This conceptual framework elevates the review from being a mere catalog of options or an encyclopedia of psychometric features.
A third major strength is that the review concentrated on the core activities typically performed by pediatric psychologists. The assessment tools were evaluated against the needs created by the role of pediatric psychology in multidisciplinary treatment contexts. This was possibly the most innovative components of the review. It greatly enhances the clinical relevance of the article.

All three of these strengths—an empirical basis for test selection, a clear linkage to a model of EBA, and the effort to connect to the routine clinical activities of pediatric psychology—set this review apart from the typical.

Three Emergent Themes: Some Elaboration on Short Forms, Retest Intervals, and Norms

The review raised three themes that merit more expansion, although where I take these threads may be different than where the review authors would have followed the twine.

Short Forms

Short forms seemed to be the most important omission from the review. There are several well-validated, psychometrically sound brief tests of cognitive functioning (Glutting, Adams, & Sheslow, 2000; The Psychological Corporation, 1999). These often consist of two to four subtests and as a result they are much less expensive and burdensome than full batteries. Within the context of the roles and activities of a pediatric psychologist, these brief tests often will assess the key construct with the same degree of precision, but with lower costs. For example, the standard error of the measure for the full scale score from the Wechsler Abbreviated Scales of Intelligence (WASI) is 3.1 points (The Psychological Corporation, 1999), whereas the full scale derived from the 10 subtest administration of the Wechsler Intelligence Scales for Children-Fourth Edition (WISC-IV) is 2.7 points (Wechsler, 2003). The WASI delivers similar precision to the WISC-IV, but with only four subtests instead of ten core or fifteen possible subtests. On that basis, short tests would probably be preferred to the extended battery versions that are included in the review. A second, evidence-based argument for preferring the brief tests is that their more humble claims about factor structure and measuring stable traits are much more likely to be statistically valid than the factor scores or subtest scores from longer batteries (Frazier & Youngstrom, 2007; Macmann & Barnett, 1997). The short forms do one or two things very well, versus attempting to assess many constructs in a mediocre fashion (Glutting, Watkins, & Youngstrom, 2003).

Retest Intervals

The review did a good job of tracking down and reporting the retest stabilities of different tests. Retest reliability data are more expensive to gather than internal consistency, so they are much less commonly reported. There are also big variations across studies in terms of the length of the retest interval (which is often poorly documented), and also in the conditions of observation. Some retest stabilities involve follow-up with untreated cohorts, and others examine stability in “treatment as usual” or in experimental treatment conditions (as is the case when correlations are reported for clinical trials). From a clinician’s perspective, the choice between a 1 week, 2 week, or 2-month test interval is a fairly moot point: it is rare for any to be reported, and it is even more unusual for the published retest interval to coincide with the time point at which a clinician wants to evaluate and make a decision about an individual child.

From a clinical viewpoint, the crucial parameters are the sensitivity of the test to real change, and the reliability with which it can rank the degree of change under conditions such as those involved in the treatment. This suggests the need for another exponential number of studies (1 week, 2 week, 6 week, 1 year stabilities—crossed with factors such as people receiving chemotherapy, people receiving ventilation, people receiving counseling for trauma, people receiving special educational services, etc.). In reality, only a tiny portion of the implied list will ever be actualized, given the rarity with which stability studies are published at all.

There are two pragmatic solutions to the problem. One is to use internal consistency estimates in place of retest stabilities. These are clearly satisfying, and not conceptually preferable; but in practice it is often the only option available. The second, complementary approach is to look for factors that affect retest stability, or that alter the sensitivity of the test to change in true score variance. “Reliability generalization” studies offer a meta-analytic methodology for identifying factors associated with changes in reliability (Vacha-Haase, 1998). It is not clear whether reliability generalization studies are published yet on cognitive measures in pediatric samples; but practitioners will want to attend to these as they become available, because they alert us to situations where our instrumentation may provide less accurate information about growth or progress for individual cases.

Separate Group Norms

How many separate group norms are desirable? This debate has occurred in the child psychopathology literature...
Evaluating tests and trying to discourage the use of suboptimal instruments (Buros, 1965).

EBA is also a newcomer. Originating in Canada, EBM has spread rapidly through medicine, giving rise to the Cochrane Collaborative, the CONSORT guidelines for evaluating clinical trials (Moher, Schulz, & Altman, 2001), and the STARD guidelines for evaluating studies of diagnostic tests (Bossuyt et al., 2003). EBM uses a “patient-centered” focus, rather than a test-centered approach. EBM teaches the formulation of answerable questions, and then emphasizes querying the research literature directly to determine what are the best evidence-supported interventions.

Core skills are question formulation, search strategies, understanding how to evaluate the quality of an article or piece of evidence, and knowing how to decide whether a piece of evidence applies to the circumstances of an individual patient. There is now a major infrastructure supporting EBM, including books written about teaching the skills, journals devoted to review articles about evidence, and dedicated databases encoding evidence and “critically appraised topics.” None of this is reflected in the EBA review of cognitive assessment. The two literatures appear to have evolved in parallel, with very little exchange.

EBM presents a “hierarchy of evidence.” At the pinnacle are systematic reviews, which are based on exhaustive reviews of a literature that use explicitly defined search strategies and criteria for evaluating the quality of the research design in each study. Next down is the individual well-designed trial. Near the bottom of the hierarchy are things like “expert opinion” and “standard of practice,” and the EBM literature is replete with examples when both of these sources of information proved to be wrong (Silverman, 1998). Viewed from the EBM vantage, the EBA review of cognitive measures is not “best evidence,” because the inclusion criteria were organized around what is already in use, not what has the best evidence. The search criteria were not exhaustive, and there was not a predetermined set of criteria for grading study design quality (Gray, 2004). This is not a criticism of the particular review: by the same standards, all of the articles in the recent special issues of Psychological Assessment (Hunsley & Mash, 2005) and Journal of Clinical Child and Adolescent Psychology (Mash & Hunsley, 2005) also appear to be on lower rungs of the ladder of evidence.

After shifting between three professional roles as a researcher, a teacher, and a licensed clinician, it seems to me that testing is useful inasmuch as it answers one of...
three “Ps” (Youngstrom, in press): (a) Does it Predict important things, such as long term course? (b) Does it Prescribe a specific treatment, or a change in treatment approach? (c) Does the test inform the Process of working with the patient, either by measuring outcomes, mediators, or intermediate progress? This heuristic has the potential to reduce the length of test batteries and decrease the number of tests given, by limiting the use of tests to situations where they have the potential to directly inform treatment. From the consumer’s perspective, lean batteries reduce the burden of testing. From a public health perspective, lean batteries decrease costs and could enable more effective resource allocation (Kraemer, 1992). From a practitioner’s perspective, “less can be more,” too. Lean batteries that concentrate on constructs directly relevant to treatment reduce the likelihood of false positive results (“Type I errors” in the language of academic psychologists) (Silverstein, 1993), and they also prevent diluting the quality of assessment by mixing valid with less valid sources of information (Gigerenzer, 1996; Kraemer, 1992). Perhaps most rewarding for psychologists, lean batteries free more time and resources for us to engage in other professional activities that compete for our time, such as therapy and consultation.

Next Steps: Integrating the Two Models of “Evidence-based”

Both EBA and EBM have strengths. There are implications and potential benefits to practitioners and researchers who take an eclectic stance and adopt the most helpful aspects of each.

Agenda for the Researcher

Investigators who familiarize themselves with both EBA and EBM will be in a position to start conducting studies and writing papers to fill in the gaps in the literature. Importing ideas from EBM into psychology guarantees that the product will have clinical relevance; exporting from psychology EBA into EBM assures that the product will have psychometric sophistication. Studies evaluating the predictive value of cognitive testing could examine criteria such as response to treatment, long-term educational or functional outcomes (Moffitt, Caspi, Harkness, & Silva, 1993), or health outcomes. Studies investigating the prescriptive value could look at the contribution of cognitive measures to diagnosis, using statistical procedures such as logistic regression to evaluate incremental validity. Cognitive variables may be a particularly rich vein to mine for process variables: general ability or verbal ability may have a big impact on the degree to which children and adolescents can benefit from more cognitive strategies, and working memory may be a limiting reagent in how well people can manage complex treatment regimens without more external scaffolding. Addressing these questions would clearly enhance the relevance of cognitive testing to pediatric practice, and it would also facilitate making much more concrete and practical recommendations—improving upon what has consistently been perceived as a shortcoming of most psychological assessment reports (Sattler, 2001; White, Nielsen, & Prus, 1984).

An exciting aspect of exchanging ideas between EBM and psychology is that many of the new questions can be answered via secondary analyses of existing data. For example, diagnostic likelihood ratios—statistics that allow a practitioner to use research evidence to make statements about risk for individual cases (Jaeschke, Guyatt, & Sackett, 1994a, 1994b)—can be computed from epidemiological data, studies of comparative risk, norms tables for psychological tests (Frazier & Youngstrom, 2006), and sensitivity and specificity estimates, as well as from raw data. Because so many of the ideas could be incorporated by reframing and repackaging existing data, there is an opportunity for rapid change in assessment tools and practice that need not wait for a new generation of data collection to be proposed and completed.

Agenda for the Practitioner

There are several steps practitioners can take right away, too. One is to evaluate each test that they routinely use, and ask what value the test adds in terms of charting important predictions, prescribing changes in choice of intervention, or informing about the process of working with a specific patient. Each “incumbent” measure should be considered for each of the clinical purposes. A major value of the earlier review of cognitive measures is that it identifies the incumbents; the tools reviewed were selected based on their widespread usage. In appraising each instrument through an EBM lens, the place to search is PsycINFO, not the test manual. EBM books suggest specific search terms to use for finding evidence related to clinical questions, such as using “PROGNOSIS AND…” to find studies related to long-term course or prediction, and “SENSITIVITY OR SPECIFICITY AND…” to find evidence pertaining to diagnostic efficiency (Straus et al., 2005). When critically appraising an instrument, the test reviewer should consider not only its adequacy for a specific purpose, but also whether there is a better tool available for the same purpose. Querying the research literature directly can overcome the inertia...
built into current standards of practice and training, helping new and better measures to spread more rapidly. Finally, the appraiser should look for evidence about whether a given test is appropriate for a specific patient (Jaeschke et al., 1994b). Psychometric methods can enrich this tenet of EBM via techniques such as multi-group confirmatory factor analysis or “differential item functioning”—sophisticated quantitative methods that have not yet been brought to bear in the arena of EBM.

Anyone interested in connecting the research evidence more tightly to individual cases can learn a lot quickly by reading about EBM techniques (Straus et al., 2005) and seeing them applied to mental health (Gray, 2004). Students and researchers will find that there is a wealth of opportunity in terms of re-analyzing data and designing new projects that provide answers to meaningful clinical questions. The sparseness of the literature in many areas where evidence is needed means that the proverbial glass is more than half ready for filling. A lot of good would be accomplished by sitting down with the cognitive review in one hand and a copy of Evidence Based Medicine: How to Practice and Teach EBM (Straus et al., 2005) in the other, and then creating a reference sheet of clinically relevant information as well as outlining a program of studies.

Conflicts of interest: None declared.

References


