AJSLP AMERICAN JOURNAL OF SPEECH-LANGUAGE PATHOLOGY

A Comment On Test Validation: The Importance of The Clinical Perspective

Journal:	*American Journal of Speech-Language Pathology*
Manuscript ID	AJSLP-18-0048.R2
Manuscript Type:	Viewpoint
Date Submitted by the Author:	02-Aug-2018
Complete List of Authors:	Daub, Olivia; Western University, Health and Rehabilitation Sciences Skarakis-Doyle, Elizabeth; University of Western Ontario, School of Communication Sciences and Disorders Bagatto, Marlene; Western University, National Centre for Audiology Johnson, Andrew; The University of Western Ontario, School of Health Studies Cardy, Janis; The University of Western Ontario, Communication Sciences and Disorders
Keywords:	Assessment, Diagnostics, Ethics

A Comment On Test Validation: The Importance of the Clinical Perspective

Olivia Daub, Elizabeth Skarakis-Doyle, Marlene P. Bagatto, Andrew M. Johnson and Janis Oram Cardy

The University of Western Ontario

Author Note

Olivia Daub, Graduate Program in Health and Rehabilitation Sciences, The University of Western Ontario; Elizabeth Skarakis-Doyle, School of Communication Sciences and Disorders, The University of Western Ontario; Marlene P. Bagatto, National Centre for Audiology, The University of Western Ontario; Andrew M. Johnson, School of Health Studies, The University of Western Ontario; Janis Oram Cardy, School of Communication Sciences and Disorders, The University of Western Ontario.

Correspondence concerning this article should be addressed to: Olivia Daub, Graduate Program in Health and Rehabilitation Sciences, The University of Western Ontario, Elborn College, London, Ontario, Canada, N6G 1H1. Email: odaub@uwo.ca

Conflict of Interest: The authors do not have any competing interests to declare.

Funding: The authors do not have any funding related to this project to declare.

Abstract

Purpose: The misuse of standardized assessments has been a long standing concern in speech-language pathology, and has been traditionally viewed as an issue of clinician competency and training. The purpose of this paper is to consider the contribution of communication breakdowns between test developers and the end users to this issue.

Method: We considered the misuse of standardized assessments through the lens of the two-communities theory, in which standardized tests are viewed as a product developed in one community (researchers/test-developers) to be used by another community (front-line clinicians). Under this view, optimal test development involves a conversation to which both parties bring unique expertise and perspectives.

Results: Consideration of the interpretations that standardized tests are typically validated to support revealed a mismatch between these and the interpretations and decisions that speech-language pathologists typically need to make. Test development using classical test theory, which underpins many of the tests in our field, contributes to this mismatch. Application of item response theory could better equip clinicians with the psychometric evidence to support the interpretations they desire, but is not commonly found in the standardized tests used by speech-language pathologists.

Conclusions: Advocacy and insistence on the consideration of clinical perspectives and decision-making in the test validation process is a necessary part of our role. In improving the nature of

the statistical evidence reported in standardized assessments, we can ensure these tools are appropriate to fulfill our professional obligations in a clinically feasible way.

A Comment on Test Validation: The Importance of the Clinical Perspective

> If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing the rationale and collecting new evidence, if necessary – *Standards of Psychological and Educational Testing* (AERA, APA & NCME, 2014)

Assessment is a core foundation in definitions of the speech-language pathologist's scope of practice (American Association of Speech and Hearing, 2016; Speech-Language & Audiology Canada, 2016). As a part of the assessment process, standardized testing informs us about whether an individual is performing above or below age expectations. Despite their ubiquity, misuse of the results of standardized assessments has been a long standing concern in our field. For over two decades, calls for increasing clinicians' psychometric knowledge have permeated our field with limited impact. In 2003, Kerr, Guildford and Kay-Raining Bird noted a bleak trend: misuses documented by McCauley and Swisher in 1984 (e.g., using test items to select treatment goals, use of age-equivalents to summarize test results) continued to be common. The onus, traditionally, has been placed on clinicians or clinical training programs to increase psychometric competency. Ensuring adequate understanding of the tests we are administering is inarguably important (and a matter of professional ethics; see Palmer, 2009, for a discussion) but is clearly not sufficient. We argue that the misuse of standardized assessments is not only an issue of professional competency, but additionally, one of communication. Reducing the misuse of standardized assessments relies on a two pronged approach: increased clinical competency and advocating for our clinical perspective in the test validation process.

The gap between research findings and clinical practice has been routinely documented and is not unique to speech-language pathology (Graham et al., 2006), resulting in the development of fields of study (such as knowledge translation and implementation science) dedicated to understading these gaps and the ways in which they can be mitigated. The two-communities theory (Caplan, 1979) describes knowledge-users (speech-language pathologists) and researchers (test-developers) as inhabiting different professional communities of practice with distinct professional jargon, values, resources, and beliefs. Bridging the *knowledge-to-action* gap rests on increasing communication between these two professional communities. Most current models of knowledge translation describe understanding the clinical context as of paramount importance to successful knowledge implementation and sustained knowledge use over time (Dobrow, Goel & Upshur, 2004; Graham et al., 2006).

Standardized Test Misuse: The Two-Communities Theory

Viewed through the lens of the two-communities theory (Caplan, 1979), standardized tests are a product developed in one community (researchers/test-developers) to be used by another community (front-line clinicians). The community of test developers, researchers, and psychometricians brings important knowledge regarding the psychometrically appropriate ways to measure speech and language, the type of statistical evidence that is needed to support our interpretations, and the limitations of their analyses. As the end users of the assessment tool, front-line clinicians have equally important insights into the decisions that will be made based on assessment results, the information we need to enrich our interpretations, and the interpretations we are required to make to fulfill program requirements. Optimal test development, therefore, is a conversation where both parties bring their unique expertise and perspectives. Are misuses of standardized assessment, then, solely a failure of one community to develop necessary competencies? Or are these misuses exacerbated by breakdowns in communication between clinicians and test-developers, particularly with regards to the types of decisions that stakeholders are required to make? The goal of the present paper is to highlight the value of clinicians' perspectives and to empower clinicians to initiate conversations with other stakeholders (i.e., researchers and test-developers).

Standardized assessment results are designed to determine when an individual performs significantly below their peers. However, as demands for public program accountability and demonstration of intervention effectiveness increase, results from standardized assessments are being requested at program and government levels for reasons such as evaluating a client's change over time in response to intervention. As an illustrative example of the issue, the Joint Committee of Infant Hearing recommends that all children who are deaf/hard of hearing receive standardized, norm-referenced assessment of speech and language outcomes on a semi-annual basis up to 3 years of age and annually thereafter (Muse et al., 2013), which is re-iterated in international consensus documents (Moeller, Carr, Seaver, Stredler-Brown & Holzinger, 2013). The results of these assessments are intended to be used by clinicians to identify whether the child is progressing towards age-appropriate language, whether the child has made significant progress over time, and whether changing the intervention plan is appropriate. These recommendations, while necessary to demonstrate program effectiveness, require more interpretation of standardized assessment results than the tests are traditionally validated to support and, indeed, require clinicians to make interpretations that are traditionally described as inappropritate for those tests (e.g., McCauley & Swisher, 1984).

Similar mismatches between test use and program requirements have been documented in the state-mandated application of cut-off scores. Spaulding, Szulga, and Figueroa (2012) documented that 8 of 45 (16%) state departments of education applied mandated severity cut-off scores to determine a child's eligibility for special education services. These cut-offs, however, were neither consistent across states nor consistent with the appropriate diagnostic accuracy cut-offs reported in standardized tests' examiner's manuals. Advocating for change in government policy is laudable, but pending such change, clinicians find themseives in a no-win situation: meet requirements needed of them by governing bodies or avoid statistical misuses. Indeed, when asked to give reasons for inappropriate uses of scores such as using age-equivalent scores to summarize test results, Kerr, Guldford and Kay-Raining Bird's (2003) respondants gave reasonable explanations grounded in clinical reality such as communicating with parents, securing funding, or using them when norms don't apply. Similarly, 86% of respondants accurately identified two problems with using individual items on a standardized tests to set treatment goals, but 55% felt the practice was an *efficient* way to identify goals. Clearly, misuse stems not only from a lack of knowledge but is also related, in part, to trying to meet a variety of needs with limited time and resources, and with time intensive tools (standardized tests) that may not only be mandatory to administer, but also limited in the scope of information they are capable of providing.

Statistical Justification for Misuses: Item Response Theory and its Implications

The description of misuses by McCauley and Swisher (1984) highlighted that some interpretations cannot be made when using tests that are designed according to specific psychometric theories and that use a particular set of statistical analyses. These misuses are not due to intransient properties of tests, but rather, are due to the nature of statistical evidence that is commonly reported in the test manuals. For example, McCauley and Swisher argued that using standardized tests to measure change is inappropriate because such tests measure a large set of relatively stable skills and are not sensitive enough to provide detailed information regarding a child's ability, or to detect small changes over time. The fault with using tests to measure progress lies not with the desire to do so *per se* but with an incompatibility between this desire and the statistical evidence or psychometric theory used to guide development of the tests. When tests are designed using classical test theory (CTT), this is indeed true. CTT assumes that all questions on a test are equally good measures of a single, unchanging skill. When standardized tests are evaluated according to CTT, this limits their interpretation in a number of ways and can thereby restrict their clinical utility.

Clinically, we can intuit that the assumptions underlying CTT about item equivalence are not true in all cases. Sometimes, questions may be harder than they should be, may require skills to answer them that aren't (intentionally) being measured (e.g., working memory), or may simply be poorly written. Because of a lack of empirical data to support such intuitions, we are traditionally required to ignore them. There is no reason, beyond psychometric tradition, that this should be the case. Statistical analyses do exist that can allow clinicians to gather much more information from a single test item than that with which we are currently being provided, and are in fact well established within the psychometric literature. Consider Item Response Theory (IRT; see Baylor et al., 2011 for a comprehensive tutorial). Contrary to the assumption of item equivalence underlying CTT, IRT analyses are guided by the assumption that a client's performance on a single item can be influenced by four parameters: (a) the client's true ability, (b) an item's difficulty, (c) an item's discriminability (sensitivity to differences between levels of difficulty), (d) and randomness (guessing). With enough data, these four factors can be statistically teased apart, yielding a wealth of information with numerous potential clinical applications.

When tests are developed using IRT, the item parameters can be used to identify (a) items that are easier or harder than others, (b) items that are more, or less, related to the skill of interest (supporting clinical intuition), and (c) items that are redundant with other items. Through this knowledge, prospective studies correlating performance on individual items, or pre-intervention ability, to therapeutic outcome could support clinicians in determining candidacy for intervention based on test performance. Item parameters can also be compared across clinical populations to identify items that are easier or harder for different groups. Using IRT parameters, test developers can then use logistic regression to identify items to which individuals with various disorders respond differently, providing information to support differential diagnosis even in situations where the overall number of items answered correctly is the same across individuals. For instance, research evaluating the language outcomes of children who are deaf/hard-of-hearing (CD/HH) receiving early intervention repeatedly documents that, as a group, children perform within normal limits on standardized assessments (e.g., Tomblin et al., 2015). This finding can mean one of two things: (a) CD/HH have language abilities commensurate with their same-aged peers or (b) the norm-referenced tests used to measure language are not sensitive to the linguistic differences between CD/HH and children with typical hearing. IRT-based analyses can be helpful when the total number of correctly answered questions isn't sensitive to subtle differences, that is, by identifying individual items that point to differences between groups. For instance, despite the fact that CD/HH are documented to perform within normal limits on omnibus measures of language, they are still known to be at risk for impairments in specific domains such as articulation and morphology, and in specific structures within these domains (Moeller, Tomblin, Yoshinaga-Itano, Connor & Jerger, 2007). In cases where total scores are not sensitive, IRT analyses have the potential to identify individual items *within the whole test* that are (a) sensitive to differences between clinical populations and typical populations and (b) sensitive to differences within clinical populations. Further, IRT can be used to identify both the whole test's and individual items' direct relation to underlying ability in a single skill. Finally, IRT parameters can be used to develop shorter (i.e., less time consuming) tests without compromising informativeness.

An additional important clinical application of IRT relates to ability scores. Test information curves can identify levels of ability in a skill where the overall test is maximally informative, but individual items can also be used to quantify ability. Because ability estimates (also known as theta scores, growth scale values, progress values, or W scores) directly estimate ability and control for the other three parameters (difficulty, discriminability, and guessing), they support uses of a test that are otherwise considered to be misuses. For example, age-equivalent scores have been described by clinicians to be clinically helpful in summarizing test results to parents and teachers (Kerr et al., 2003), however, their interpretation and calculation is statistically problematic. Age-equivalents statistically "represent the mean or median score derived for a normative sample for a particular age group" (Maloney & Larrivee, 2007, p.p 86) – that is, the age at which a child's score is considered average. Like standard scores, age-equivalents are assigned based on comparisons of an individual to a group of peers. Age-equivalents do not imply, for example, that a 6-year-old child with an age-equivalent score of 3 years uses and understands the same language as a 3-year-old child. Rather, age-equivalents imply that the child correctly responded to the same number of questions to which a typical 3-year-old in the norming sample would respond. Unlike age-equivalents, ability scores enable the interpretation of *how much* ability a client has in a specific skill (loosely defined) based on the pattern of their responses to individual items. Ability scores more directly capture what age-equivalents attempt to by virtue of their underlying relation to ability in a skill.

With sufficient evaluation and correlation of ability scores to other measures of language, a norm-referenced test could theoretically be validated to provide a summary statistic that more closely aligns with a child's stage or profile in language development than the age-equivalent score. Clinically, this statistic could be transformed to be reported using terminology similar to an age-equivalent but in a statistically appropriate way. This statistic would provide clinicians with a *psychometrically appropriate* way to communicate test results in ways they have identified as important (i.e., to parents and teachers). Similarly, ability scores can be used to document whether or not an individual acquired more/less of that skill over time. Rather than interpreting a client's performance only in relation to their peers, IRT analyses enable interpretation of a client's performance relative to a skill, as well as to themselves. Here again, IRT matches statistical evidence with the types of decisions clinicians are already making. Rather than attempt to limit clinical interpretations to suit statistical evidence, our field is better supported by designing statistical evidence to fit clinical uses. When taken together, traditional psychometrics (e.g., reliability, normative scores) and IRT analyses enrich clinical interpretation and the value of administering a single test.

Daub, Baggato, Johnson and Oram Cardy (2017) illustrated the utility of growth scale values in measuring change over time using data from a province-wide database. When measured using standard scores alone, CD/HH did not demonstrate change in language ability relative to their same-aged peers after they were fitted with hearing aids. This lack of change might be misinterpreted as no improvement, which would stem from relying on types of scores that are not sensitive enough to tell the full story. When the same children's progress was evaluated using growth scale values, significant improvement on the expressive communication and auditory communication scales of the *Preschool Language Scale, 4th edition* (PLS-4; Zimmerman, Steiner & Pond, 2002) was observed. These differences have important implications: misinterpreting results as no growth could, theoretically, be used as justification for reducing services or de-funding programs. Jointly considering changes in children's *relative standing* (standard scores) and *ability* (growth scale values) demonstrated that children in this database not only improved in their spoken language comprehension and use, but did so at a rate sufficient to maintain their standing relative to same-aged peers who mostly were not hard of hearing (>99% in the PLS-4 sample), a very positive story. Measurement errors can have costly and potentially devastating consequences for clients: denying, delaying, or discontinuing services to those who need it; providing services to those who don't; and misallocating resources.

Although analyses such as IRT are well-established in the psychometric literature and require sample sizes often collected in traditional norming samples, they are not yet commonly reported in examiner's manuals for tests used by speech-language pathologists. Although a review of all commercially available standardized tests is beyond the scope of the present work, the authors explored top publishers and retailers in the field of speech-language pathology to identify the prevalence of IRT-based scores in commercially available tests. The websites of Brookes Publishing, Linguisystems, Pro-Ed, Pearson Assessment, and Super Duper Publications were examined. No tests of adult language were identified that either included IRT-based analyses or reported growth scale values (or an alternately named equivalent). Seven tests of child speech and language (all published by Pearson Education Inc from 2004 on; see Table 1) reported growth scale values and one test, the *Test of Integrated Language & Literacy Skills* (TILLS; Nelson, Helm-Estabrooks, & Hotz, 2016) used IRT-based analyses, but did not provide growth scale values (or equivalently derived scores).

It is not a new concern that test examiner manuals do not provide all sources of statistical evidence that would allow us to make the most of our assessment results. McCauley and Swisher (1984) noted that z-scores were not frequently reported in standardized assessment manuals. Ten years later, Plante and Vance (1994) noted that very few preschool standardized assessments contained a sufficient level of detail in reporting their psychometric properties, although they did provide more detail than in the tools evaluated by McCauley and Swisher. Friberg (2010) observed a trend of improvements in the examiner's manuals for school-aged language assessments, in terms of their frequency of reporting the validity evidence for which previous work had advocated. Historically, advocating for more statistical detail from test developers has resulted in seeing improvements in the level of detail provided in examiner manuals.

Closing this knowledge gap within standardized assessment is an ethical obligation to our clients (Palmer, 2009), as they are entitled to the best available assessment protocols. Currently, assessment tools do not exist to support all of the decisions we are required to make within our profession such as whether or not a client has made significant progress, or whether or not they are progressing appropriately towards goals. The responsibility, therefore, lies with us to communicate with test developers on an ongoing basis about additional interpretations we need to make within our practice. A caveat, however, is that advocacy cannot occur in the absence of clinical competence. We, as clinicians, are not justified in calling for changes that we do not understand how to use, or how to use appropriately.

Moving Forward: Increasing Clinical Competency

With respect to psychometric competency, it is our role as clinicians to be able to identify *when* an interpretation is statistically supported and when it is a misuse. When encountering examiner's manuals that do not provide statistical evidence for an interpretation we may wish to make, we must ask ourselves: is there evidence that an interpretation is inappropriate to make or is there simply no evidence at all? In cases where the evidence suggests that our interpretation is inappropriate, then the test should not be used in this way. For instance, using individual items to set therapeutic goals is a psychometric misuse because tests have not been designed, and evidence has not been collected, to demonstrate that individual items are sufficient to capture broad areas of skills or are associated with improved therapeutic outcomes when used this way. In this case, the misuse is the result of a lack of evidence. Consider, however, using a -2*SD* (standard score = 70) cut-off to rule out language disorder using the Total Language Score on the *Preschool Language 5th edition* (PLS-5; Zimmerman, Steiner & Pond, 2011). First, the PLS-5 examiner's manual only provides sensitivity and specificity values using a -1*SD* cut-off (which are 0.83 and 0.80, respectively, both meeting the *acceptable* accuracy level of .80 proposed by Plante & Vance, 1994). Therefore, SLPs lack some of the necessary evidence to determine whether the PLS-5 has adequate diagnostic accuracy at a -2*SD* cut-off. The PLS-5 examiner's manual does, however, provide information on both the positive predictive power (PPP; the percentage of children identified as having a language disorder who are accuractely classified) and negative predictive power (NPP; the percentage of children identified as *not* having a language disorder who are accurately classified) of this cut-off in samples with different disorder base rates. When used in settings where children are very likely to have a language disorder (i.e., clinics where 70-90% of children being assessed truly have a language disorder), the PPP of a -2*SD* cut-off is quite high: SLPs can be between 97-99% certain that children receiving a standard score of 70 or lower on the PLS-5 truly have a language disorder. However, at this same base rate range of 70-90%, the NPP of a -2*SD* cut-off is quite low, ranging from .16 to .43. This indicates that between 57-84% of children classified as *not* having a language disorder due to receiving a standard score above 70 will be misclassified. In this case, statistical evidence clearly demonstrates that applying a -2*SD* cut-off for the purposes of *ruling out* a language disorder is not well-supported in similar clinical settings. Therefore, using the PLS-5 to rule out language disorder in this type of clinical scenario is a misuse, but not because of an absence of evidence: the evidence has been collected, and instead suggests that the PLS-5 is not sufficiently accurate for this purpose. Statistical evidence does suggest that a -2*SD* cut-off on the PLS-5 Total Language Score has strong diagnostic utility in *ruling in* language disorder in these settings. However, the absence of sensitivity/specificity information for the -2*SD* cut-off leaves open the possibility that it is nonetheless not clinically useful. If the sensitivity of the -2*SD* cut-off is in fact low (say, for example, .58), this would mean that SLPs would only detect 58% of children who have a language disorder. In this scenario, the high PPP values indicate that SLPs could be highly confident whenever they have classified a child as having a language disorder using a -2*SD* cut-off, but the low sensitivity value would mean that this would happen for only 58% of the children who truly have a language disorder – 42% of them would be missed (see Lange & Lippa, 2017, for a helpful discussion of the importance of joint consideration of sensitivity/specificty and PPP/NPP in selecting cut-off scores and evaluating the clinical utility of diagnostic tests).

As clinicians, we need to know how we intend to use a test and what statistical information we require to justify its use. In order to bring about changes to standardized tests, we must understand psychometric best practices and the most appropriate ways to use and interpret the types of psychometric data reported in examiner's manuals. There is evidence, empirical and anecodotal, to suggest that clinical knowledge surrounding psychometrics could be strengthened in our profession. A survey of Canadian speech-language pathologists documented that only 17% (of 143 clinicians) felt "completely confident" with their psychometric knowledge where 66% were "somewhat confident", and 17% reported that they were "not at all confident" (Kerr et al., 2003). Psychometric knowledge, in this study, was broadly defined as having the knowledge to "evaluate tests adequately" (Kerr et al., 2003, p. 20). Further consider that IRT analyses are relatively new to our field – it is unlikely that clinicians in this study were considering their ability to evaluate IRT based analyses when responding to the survey. That the majority of clinicians reported being only "somewhat confident" in their ability to evaluate tests *adequately*, it is unsurprising that our field continues to see gaps in best assessement practices. For instance, a survey of American speech-language pathologists by Betz, Eickhoff and Sullivan (2013) documented that only a few tests tended to be frequently used, and that test selection was correlated with publication year rather than metrics of psychometric quality such as reliability, criterion validity, or diagnostic accuracy.

Clearly, our profession needs more support to promote psychometric competency if we are to expect appropriate uptake of newer statistical analyses such as IRT. This is not to dismiss the laudable efforts of researchers within our profession who have worked to tackle psychometric issues in clinically accessible ways. There exists a large body of literature, particularly within the area of child language, dedicated to exploring issues such as diagnostic accuracy (e.g., Pena, Spaulding & Plante, 2006; Plante & Vance, 1994), application of cut-off scores (Spaulding et al., 2012), and outlining evidence-based practice (including for assessment; Dollaghan, 2004). However, our profession lacks access to comprehensive education surrounding psychometrics. Ideally, such an educational resource would (a) be developed by psychometric leaders, (b) be consistent across service regions, (c) offer tangible recommendations for test selection and interpretation, and (d) support clinicians when they are required to deviate from psychometric best practice. Numerous possible solutions to this problem exists (e.g., establishing corpuses that compare and contrast the uses of different tests, psychometric webinars and tutorials, clinical practice guidelines and practice statements), but they will not be successful if the clinician's voice is absent. Clinicians are in the best position to evaluate their own understanding of psychometrics and determine what materials are accessible, feasible, and manageable given the context of their clinical practice. We argue that misuses of test stem from problems in communication, and that the solution relies on communication from *both sides* of the knowledge-to-action gap. Researcher initiated efforts, such as publications in peer-reviewed journals, over the past three decades have not been sufficient to close this gap.

Moving Forward: Advocacy

With knowledge can come advocacy. As clinicians, we have the ability to change the way standardized assessments are reported. Historically, our field has seen major gains in the reporting of psychometric detail through calls to action (as discussed above), but we must continue this push as the demands for assessment use, and the nature of psychometric best practices, change. At its simplest level, we have financial leverage in choosing which standardized tests we purchase. However, we also have ongoing opportunities to communicate with test developers via direct correspondances, at national conference booths, or through test-developer intiated calls for feedback (e.g., in Februray 2018, Pearson Education Inc. published an online survey requesting clinician feedback on the PLS-5). Sound knowledge of psychometrics, both new and old, supports the thoughtful response to invitations such as these. For instance, clinicians in regions with mandated cut-off scores might consider responding to a survey by outlining the cut-off requirements they are obligated to fulfill, and a test-developer may respond by designing the test to be either maximially (or at least appropriately) diagnostically accurate at that mandated cut-off score.

In cases where the evidence has not been provided, this is an opportunity to communicate with test developers to continue the test validation process. Consider the recommendations put forth by the Joint Committee of Infant Hearing. With a clearly defined call for a specific frequency of assessment, tests that are designed to be used for CD/HH ought to provide evidence that they are appropriate to meet this clinical need. These recommendations can serve as concrete evidence to a test-developer that it is financially in their best interest to report on analyses that support this test use, or develop new tests that can. These unified calls for annual or semi-annual assessment are a wonderful example of an impetus that test developers can use to continue the iterative validation process and appraise their tests' appropriateness for assessment at these intervals. In bringing our voices to the test-development conversation, we have the potential to dramatically shape the nature of future standardized assessment tools and facilitate our own clinical interpretations with tools tailored to support us and the clients we serve.

Conclusions

Improving evidence-based practice in assessment is a necessary goal. However, calls to improve psychometric knowledge amongst speech-language pathologists do not acknowledge that clinicians are, often, required to make decisions about a client that standardized tests do not commonly provide statistical evidence to support. Inarguably, there is room for improvement in regards to psychometric competency within our profession, but clinicians must also recognize and insist that the assessments they use provide them with the most statistical information possible to support their interpretation. Standardized assessments are costly in terms of price, time to administer, and time spent analyzing and interpreting results. Maximizing the clinical utility of our assessments is necessary to improve our assessment practices, but doing so requires that we advocate for ourselves, on behalf of our clients, and communicate with test-developers.

We conclude with a comment to researchers and test developers. The purpose of the present paper has been to highlight the value of the clinical perspective in test development and to encourage clinicians to insist their voices are present in the conversation. It is equally important that researchers and test developers actively seek to understand the clinical perspective. Three decades worth of research has routinely documented that the status quo for reporting statistical results in examiner's manuals has been insufficient for clinicians and, more importantly, for their clients. Active efforts on the part of both communities to engage in the conversation about test validation has the potential to substantially improve the quality and value of standardized assessments for the people with communication disorders we all aim to serve.

References

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

American Speech-Language-Hearing Association. (2016). Scope of Practice in Speech-Language Pathology [Scope of Practice]. Available from www.asha.org/policy.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*, *20*(3), 243–259. https://doi.org/10.1044/1058-0360(2011/10-0079)

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*, 133-146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Caplan, N. (1979). The two-communities theory and knowledge utilization. *American Behavioral Scientist*, *22*(3), 459–470. https://doi.org/10.1177/000276427902200308

Daub, O., Bagatto, M. P., Johnson, A. M., & Oram Cardy, J. (2017). Language outcomes in children who are deaf/hard of hearing: The role of language ability before hearing aid intervention. *Journal of Speech, Language and Hearing Research, 60,* 3310-3320. https://doi: 10.1044/2017_JSLHR-L-16-0222

Dobrow, M. J., Goel, V., & Upshur, R. E. G. (2004). Evidence-based health policy: Context and utilisation. *Social Science and Medicine*, *58*(1), 207–217. https://doi.org/10.1016/S0277-

9536(03)00166-7

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37*(5), 391-400. https://doi.org/10.1016/j.jcomdis.2004.04.002

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, 4th ed.* Bloomington, MN: Pearson Education Inc.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92. https://doi.org/10.1177/0265659009349972

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation, 3rd ed.* Bloomington, MN: Pearson Education Inc.

Graham, I. D., Logan, J., Harrison, M. B., Straus, S. E., Tetroe, J., Caswell, W., & Robinson, N. (2006). Lost in knowledge translation: Time for a map? *The Journal of Continuing Education in the Health Professions*, *26*(1), 13–24. https://doi.org/10.1002/chp.47

Graham, I., & Tetroe, J. M. (2009). Getting evidence into policy and practice: Perspective of a health research funder. *Journal of the Canadian Academy of Child and Adolsecent Psychiatry*, *18*(1), 46-50. Retrieved from http://www.cacap-acpea.org/en/cacap/Journal_p828.html

Kaufman, A. S., & Kaufman, L. N. (2014). *Kaufman Test of Educational Achievement, 3rd edition.* Bloomington, MN: Pearson Education Inc.

Kerr, A., Guildford, S., & Kay-Raining Bird, E. (2003). Standardized language test use: A Canadian survey. *Journal of Speech-Language Pathology and Audiology*, *27*(1). Retrieved from https://cjslpa.ca/archive.php

Kothari, A., & Wathen, C. N. (2013). A critical second look at integrated knowledge translation. *Health Policy*, *109*(2), 187–191.https:// doi.org/10.1016/j.healthpol.2012.11.004

Lange, R. T., & Lippa, S. M. (2017). Sensitivity and specificity should never be interpreted in isolation without consideration of other clinical utility metrics. *The Clinical Neuropsychologist, 31*(6-7), 1015-1028. https://doi.org/10.1080/13854046.2017.1335438

Maloney, E. S., & Larrivee, L.S. (2007). Limitations of age-equivalent scores in reporting the results of norm-referenced tests. *Contemporary Issues in Communication Science and Disorders, 54,* 66-93. Retrieved from https://www.asha.org/uploadedfiles/asha/publications/cicsd/2007flimitationsofageequivalentscores.pdf

McCauley, R. R., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*, 338-348.

Moeller, M. P., Carr, G., Seaver, L., Stredler-Brown, A., & Holzinger, D. (2013). Best practices in family-centered early intervention for children who are deaf or hard of hearing: An international consensus statement. *Journal of Deaf Studies and Deaf Education*, *18*(4), 429–445. https://doi.org/10.1093/deafed/ent034

Moeller, M. P., Tomblin, J. B., Yoshinaga-Itano, C., Connor, C. M., & Jerger, S. (2007). Current state of knowledge: language and literacy of children with hearing impairment. *Ear and Hearing*, *28*(6), 740–753. https://doi.org/10.1097/AUD.0b013e318157f07f

Muse, C., Harrison, J., Yoshinaga-Itano, C., Grimes, A., Brookhouser, P. E., Epstein, S., … Martin, B. (2013). Supplement to the JCIH 2007 position statement: principles and guidelines for early intervention after confirmation that a child is deaf or hard of hearing. *Pediatrics*, *131*(4), e1324-49. https://doi.org/10.1542/peds.2013-0008

Nelson, N. W., Helm-Estabrooks, N., & Hotz, G. (2016). *Test of Integrated Language and Literacy Skills*. Baltimore, MD: Brookes Publishing Co.

Palmer, C. V. (2009). Best practice: It's a matter of ethics. *Audiology Today, 5*, 31–35. Retrieved from https://www.audiology.org/publications-resources/audiology-today/archives

Pena, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology*, *15*, 247-254. https://doi.org/10.1044/1058-0360(2006/023)

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *23*, 15–24. https://doi.org/0161-1461-94-2501-0015$01.00/0

Semel, E., Wiig, E., & Secord, W. A. (2004). *Clinical Evaluation of Language Fundamentals – Preschool, 2nd ed.*. Bloomington, MN: Pearson Education Inc.

Spaulding, T. J., Szulga, M.S., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools, 43*, 176-190. https://doi.org/10.1044/0161-1461(2011/10-0103)

Speech-Language & Audiology Canada. (2016). Scope of Practice for Speech-Language Pathology. Available from www.sac-oac.ca.

Tomblin, J. B., Harrison, M., Ambrose, S. E., Walker, E. A., Oleson, J. J., & Moeller, M. P. (2015). Language outcomes in young children with mild to severe hearing loss. *Ear and Hearing, 36*(Suppl 1), 76S–91S. https://doi.org/10.1097/AUD.0000000000000219.

Wiig, E. H., Semel, E., Secord, W. A. (2013). *Clinical Evaluation of Language Fundaments, 5th edition.* Bloomington, MN: Pearson Education Inc.

Wiig, E. H., Semel, E., Secord, W. A. (2014). *Clinical Evaluation of Language Fundaments, 5th edition – Metalinguistics.* Bloomington, MN: Pearson Education Inc.

Williams, K. T. (2007). *Expressive Vocabulary Test, 2nd edition.* Bloomington, MN: Pearson Education Inc.

Woodcock, R.W. (2011). *Woodcock Reading Mastery Tests, 3rd edition.* Bloomington, MN: Pearson Education Inc.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale, 4th edition.* San Antonio, TX: The Psychological Corporation.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scale, 5th edition.* San Antonio, TX: The Psychological Corporation.

Table 1.

Test Name	**Publication Year**
Clinical Evaluation of Language Fundamentals, Preschool, 2nd edition	2004
Peabody Picture Vocabulary Test, 4th Edition	2007
Expressive Vocabulary Test, 2nd edition	2007
Preschool Language Scale, 5th edition	2011
Woodcock Reading Mastery Tests, 3rd edition	2011
Clinical Evaluation of Language Fundamentals, 5th edition	2013
Clinical Evaluation of Language Fundaments, 5th edition - Metalinguistics	2014
Kaufman Test of Educational Achievement, 3rd edition	2014
Goldman Fristoe Test of Articulation, 3rd edition	2015

Standardized tests of speech or language that include IRT-based ability scores