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

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

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

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

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

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

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

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 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.

References

American Educational Research Association, American Psychological Association, National

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.□□tG;H)jz@H6(jzg□H)j/66zgA□wG

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

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;;/66zg cw6zgowG;H)jz J7;;/6.6zgtwGLH(zgtwGLHLN22(zgpwGNH@jzgsw

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
E7x□□A□□□A□□□□□yk□□AA□□□□□Ay□□□A□□□□□yk7RK□A□□□□□yk□□A□□□A□□□□□	

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

nition