Interpretation and Use of Standardized Language Assessments for Diverse School-Age Individuals

Teresa Girolamo, **Samantha Ghali**, **Ivan Campos**, **Andrea Ford**
University of Connecticut, University of Kansas, CAHELP, University of Minnesota

Abstract

Purpose: The ultimate aim of an assessment is to help examiners make valid conclusions about an individual's skill given their performance on a particular measure. Yet, assessing the language abilities of culturally and linguistically diverse individuals requires researchers and practitioners to carefully consider the appropriateness of traditional parameters of test psychometrics (e.g., reliability, or consistency of assessments as measurement) *plus* the intersectional identities that inform the generalizability of these parameters. The purpose of this clinical focus article is to provide clinicians and researchers with resources to interpret and use common standardized language assessments in English for culturally and linguistically diverse school-age youth. We present theories from psychometrics, legal studies, and education relevant to language assessment of diverse individuals, review standardized language assessments in English, and provide theory-to-practice applications of language assessment scenarios.

Conclusions: Implementing intersectional approaches in working with diverse children and using assessment scores as just one piece of evidence amid a broader evidence base will contribute to a more accurate evaluation of culturally and linguistically diverse children's language abilities. A comprehensive approach involving multiple stakeholders across the field of communication sciences and disorders may support achieving such implementation.

Keywords

assessment; cultural and linguistic diversity; interpretation and use

Assessing the language abilities of culturally and linguistically diverse (CLD) individuals with suspected and diagnosed language needs requires clinicians and researchers to adopt an approach responsive to the multi-faceted, intersectional identities of these individuals (Castilla-Earls et al., 2020). These intersectional identities may include, but are not limited to, an individual's culture, language, race, ethnicity, gender, sexuality, and other

Correspondence concerning this article should be addressed to Teresa Girolamo, Cognitive Neuroscience of Communication, 406 Babbidge Road, U-1020, University of Connecticut, Storrs, CT, 06269. Phone: 860-486-0195. teresa.girolamo@uconn.edu.
Teresa Girolamo, Cognitive Neuroscience of Communication, Storrs, CT, University of Connecticut. Samantha Ghali, Child Language Doctoral Program, Lawrence, KS, University of Kansas. Ivan Campos, California Association of Health and Education Linked Professions (CAHELP), Apple Valley, CA. Andrea Ford, Educational Psychology, Minneapolis, MN, University of Minnesota.

Conflict of Interest

The authors have no financial disclosures. Nonfinancial disclosures are as follows. Ivan Campos serves on ASHA's Multicultural Issues Board, Samantha Ghali serves on ASHA's International Issues Board, and Teresa Girolamo serves on CAPCSD's Diversity, Equity, and Inclusion Committee.

sociodemographic characteristics (Crenshaw, 1989, 1991; Gillborn, 2012; Hernandez-Sáca et al., 2018). Appreciating multiple identities in this way shifts away from a “box-like” approach to diversity – where individuals are diverse or not – and into a space that more appropriately recognizes the complexity of each individual.

In adopting this approach, examiners must *proactively* consider how the language needs of CLD individuals together with these other identities can interact with one another during an assessment. On one hand, race and disability are social constructs that primarily entail how others react to individual differences, rather than individual differences themselves (Annamma et al., 2013). Further, examiners must recognize that standardized, norm-referenced test scores and assessment performance constitute only part of the evidence base (Daub et al., 2021). As an expert, it is the examiner’s responsibility to identify which additional pieces of information (i.e., evidence) are necessary to inform conclusions about an individual’s language abilities, and in turn, the decisions made (Kane, 2001, 2006, 2016). This perspective is consistent with the American Speech-Language-Hearing Association (ASHA) Code of Ethics (2016) and Individuals with Disabilities Education Act (2004), which mandate CSD professionals and school-based practitioners to *not* discriminate in their professional activities based on individual differences, including race, culture, and disability status.

It is widely recognized that standardized, norm-referenced assessments are insufficient for use with CLD populations, particularly when we consider intersectional identities (e.g., Castilla-Earls et al., 2020). Yet clinicians report that standardized assessments are often included in school eligibility policy (Selin et al., 2019), with little attention to the psychometric properties during test selection (Betz et al., 2013). To circumvent this issue, examiners need explicit training and information on how to understand psychometric properties of these assessments in relation to intersectional identities to ensure that they develop valid conclusions about language ability. Prior work has focused on the robustness of standardized language tests following traditional metrics of psychometric validity (e.g., reliability, diagnostic accuracy; Betz et al., 2013; Castilla-Earls et al., 2020; Daub et al., 2021; Friberg, 2010, McCauley & Swisher, 1984; Plante & Vance, 1994). To our knowledge, there has yet to be a focus on how to account for intersecting identities in Black, Indigenous, and People of Color (BIPOC) in making valid conclusions about language ability. To address this knowledge gap, this clinical focus article describes theory- and data-based observations on language assessment for diverse school-age individuals by: (a) introducing theories from multiple disciplines relevant to language assessment of diverse individuals; (b) reviewing standardized language assessments in English; and (c) providing theory-to-practice applications of language assessment for diverse children.

Theoretical Considerations in Language Assessment of Diverse Individuals

To organize these theoretical considerations, we first provide a proposed pathway from assessment performance to interpretation and use in Figure 1, with recognition of the need to attend to multiple influences (e.g., individual identities, context, and validity evidence). This pathway is informed by unified validity from psychometrics (Kane, 2001), intersectionality theory from legal studies (Crenshaw, 1989, 1981) and critical race theory (Gillbert, 2015), and DisCrit theory from education and disability studies (Annamma et al., 2013), all of which help to explain the interaction between examinees' assessment performance and examiners' interpretation. In the following sections, we will highlight particular elements of the figure to promote understanding of the pathway and its connection to each theory. Briefly, however, the pathway involves four major steps that occur sequentially: (a) gathering assessment performance, (b) using assessment performance to make an interpretation of an individual's language ability based on that performance, (c) synthesizing multiple sources of assessment performance to make conclusions about an individual's overall language ability, and (d) using conclusions to inform the specific decisions that need to be made (e.g., eligibility/diagnosis, services, or placement, or grouping individuals with particular characteristics).

The pathway in Figure 1 begins with an individual's assessment performance on a singular measure (i.e., Step A). This performance is couched within and connected to two factors that we must recognize as influences: (a) the attributes that examinees bring with them into assessment, such as the construct an assessment measures (i.e., language ability), and (b) the context of the assessment (e.g., tasks and situation; Bachman, 2005). Examinee attributes include not only language ability but also other dimensions of identity. These factors influence how examiners make conclusions about language ability and eventual decisions related to perceptions of ability (Kane, 2001).

Validity—In our approach and proposed pathway, we adopt a concept of unified validity, which posits that validity is a singular construct and involves the quality of *inferences* an examiner makes, rather than the quality of a given assessment (Messick, 1989, 1995; Kane, 2006, 2016); see Figure 1. Thus, examiners amass different pieces of evidence (i.e., Step A-1 and Step A-2 in Figure 1), such as performance on a particular language measure (e.g., Clinical Evaluation of Language Fundamentals-5th Edition [CELF-5]; Wiig et al., 2013), and use their best professional judgment to make conclusions about an individual's language ability (i.e., Step C in Figure 1). For each language measure, then, there are also different considerations in validity; see Validity Evidence for Assessment Interpretations between Step A and Step B in Figure 1. For example, these considerations include but are not limited to the: (a) content (i.e., relevance and representativeness of the test items), (b) generalizability and boundaries (i.e., the degree to which interpretations can go beyond the norming sample), (c) external associations (i.e., the degree to which there are associations with measures of the same or different constructs), and (d) diagnostic accuracy (i.e., degree to which it accurately discriminates between disorder and typical development; Eusebi, 2013; Grimm & Widaman, 2012; Messick, 1995; Purpura et al., 2015). These considerations, or sources of evidence for validity, are part of the overall, interconnected validity versus being independent types of validity that operate in isolation (Messick, 1995; Purpura et al., 2015). As Daub and colleagues (2021) noted, this concept and approach to validity entails critical attention to both how an examiner will interpret and use a test.

Given this premise, examiners assessing individuals with language needs must consider how information they gathered about the individual's intersectional identities may align with (or not) the intended interpretations and eventual use of language assessment performances (e.g., eligibility, diagnosis, or grouping of individuals within an investigation; Messick, 1989). In Figure 1, the bidirectional arrows between an individual's identity and Step C and Step D illustrate this notion. At a broad level, this could mean requiring a higher benchmark of reliability for educational decision-making ($r \geq .90$) than for lower-stakes processes like educational screening ($r \geq .80$; Salvia et al., 2016). In brief, reliability refers to the degree of consistency of an assessment as a measuring instrument when following the same administration procedures and scoring rules; thus, a reliability of .90 or above indicates that the ordering of all examinees' scores on a test would nearly perfectly correspond to the hypothetical ordering of all examinees' scores if examinees took an equivalent, hypothetical form of the test (American Educational Research Association, 1999). At a detailed level, decision-making is a complex process, with necessary attention to several parameters. Examiners should consider how relevant their interpretation of a language assessment score is to decision-making, how useful their interpretation is for making the decision, the consequences of assessment use and subsequent decision-making, and whether an assessment suffices for decision-making purposes (Bachman, 2005). For example, an examiner may suspect that a child has language impairment yet their score on a measure of overall language ability is an 86 (i.e., within the typical range). Under these circumstances, an examiner might reference specific indices, such as a confidence interval, which provides a range of estimates for an unknown parameter (in this case, a "true score" of language ability), and realize that the possible range of "true scores" does include scores in the language impairment range (Selin et al., 2019).

Intersectionality and DisCrit—In deciding how to interpret and use language assessment performance, examiners must consider the *whole* individual. Following intersectionality theory, Black, Indigenous, and People of Color (BIPOC) may have multiple intersecting identities that are each tied to experiences of marginalization and give rise to multiple marginalization (Crenshaw, 1989, 1991). Multiple marginalization is not additive but an examination of how identities grapple with one another (Bauer et al., 2021). Subsequent work posited that while race is central to intersectionality, other dimensions which are highly relevant to one's identity and how an individual is situated in society (e.g., disability), must be deeply considered (e.g., Gillborn, 2015; Hernández-Saca et al., 2018). Moreover, following DisCrit theory, race and disability are social constructs that exist largely in relation to the perceptions of others and the categorization of individuals as "Other" (i.e., deviant from the norm; Annamma et al., 2013, 2017). As constructs that entail *responses* to individual differences (versus individual differences themselves), disability and race reinforce one another and can exacerbate bias (Annamma et al., 2013, 2017).

For example, the first author is a Korean American immigrant who was a late talker and received multiple evaluations in childhood. Having a protracted period of language development, being an immigrant, and being Asian each informed how clinicians and researchers in speech-language pathology interpreted her assessment performance. While many expected, as expressed to her parents, Asians to "be smart," having a language delay and perceptions about the author's cultural and linguistic background resulted in highly discordant reports that recognized her language needs yet provided no clear diagnosis or pathway for receipt of services. Had examiners considered their own biases as a piece of evidence together with assessment performance and parent report as other pieces of evidence, their interpretation of assessment performance may have led to different inferences about validity – and outcomes for the author. In this way, others' perceptions about race and disability in language assessment can give rise to nuanced bias that adversely impacts diverse individuals.

At a systemic level, intersectionality and DisCrit have real-world implications for assessment of culturally and linguistically diverse individuals. One claim is that minorities are underrepresented in special education, as special education teachers nationwide were less likely to report Hispanic, Black, and "other" minorities (i.e., Asian and Native American) as having a diagnosis of speech or language impairments or four other diagnoses than white children (Morgan et al., 2015). Yet this approach erases intersectional variability (Skiba et al., 2016). Nationwide, Black students are only overrepresented in low-status disability categories, such as intellectual disability, emotional disturbance, and learning disability (Robinson & Norton, 2021; Skiba et al., 2016; Skrtic et al., 2021). At a state level, Black students are underrepresented in speech or language impairment (Robinson & Norton, 2021). In turn, while Asian & Pacific Islander students are underrepresented and Native American children are overrepresented in special education nationwide, collapsing them into one group masks these differences (Skiba et al., 2016). Further, when considering 11 different groups of Asian & Pacific Islander students versus one group, eight were underrepresented for speech or language impairment (Cooc, 2019). Last, representation of Hispanic or Latinx children varies by location, highlighting the importance of race and disability as social constructs (Skiba et al., 2016).

Researchers play a role in mitigating discrepancies, regardless of whether they utilize standardized language assessments in their studies, because study findings contribute to the evidence base. For example, in the case of Autism Spectrum Disorder, Hispanic/Latinx and female individuals are each more likely to receive a delayed diagnosis or under-diagnosis (Loomes et al., 2017; Maenner et al., 2021). One factor in this inequity is that BIPOC and female individuals – and especially those with co-occurring diagnoses such as intellectual disability – are underrepresented in autism research (Durkin et al., 2015; Russell et al., 2019). Consequently, autistic individuals of marginalized backgrounds are less likely to be part of the scientific literature used to develop assessments and diagnostic criteria (Buchanan & Wiklund, 2020; Giwa Onaiwu, 2020). At the same time, research studies may fail to acknowledge lack of inclusive samples as a limitation to generalizability of the findings (Russell et al., 2019), thus discounting the importance of including *all* variability in science and reinforcing norms built upon only a subset of the population (Annamma et al., 2017). Thus, inequity in language assessment interpretation and use involves not just *who* researchers include in studies, but also *how* they characterize participants and findings.

To integrate intersectionality and DisCrit in assessment, examiners must appreciate that all of a diverse individual's identities interact with one another and inform assessment performance and interpretation (Annamma et al., 2013). In Figure 1, we illustrate this notion by the arrows from an individual's identity to both performance and interpretation of each assessment (i.e., Step A and Step B), but also the combination of evidence for making conclusions about an individual's overall language ability (i.e., Step C). In the case of BIPOC autistic youth with language impairment, assessment performance may be impacted due to specific sociocultural norms pertaining to assessment. For instance, if testing takes place in an environment where the examinee visibly differs from the examiner, the examinee may have to adapt the sociocultural norms of the testing environment, such as how to engage in social interactions, which could vary from their own. Thus, in addition to the cognitive load arising from assessment itself, the examinee might also have an additional cognitive load – that of toggling between two sets of norms could be unduly increased (Girolamo et al., 2020). If an examiner fails to consider these factors as they seek to build valid inferences based on the assessment performance, they risk perpetuating harm to diverse individuals. An additional consideration is how standardized language assessments conceptualize diverse examinees.

Intersectional Approaches to Norming in Standardized Language Assessments

Assessment manuals offer rich information for evaluating how assessment performance may contribute to examiners' overall evidence base for making valid conclusions about an individual's language ability. Traditional parameters of validity when making conclusions about language ability from assessments (e.g., internal consistency, reliability) do not necessarily consider intersectional identities (Denman et al., 2017). Following intersectionality and DisCrit, an intersectional approach to test standardization would evaluate all intersections of identities *and* report out findings on all intersections (Bauer et al., 2021). To date, a systematic review of 707 articles across disciplines where intersectionality is more prevalent (education, epidemiology, political science, psychology, sociology) revealed that such methods are few (Bauer et al., 2021). It is unknown whether test norming uses intersectionality. As CSD professionals, individual examiners *can* evaluate to what extent the diverse individuals they assess are represented in test norming and whether this information supports inferences about validity.

Procedure

Having served as school-based practitioners and conducted language research with individuals from birth through high school, the authors selected assessments to better understand intersectional representation in assessment development. This activity was not a review but an example of how to use assessments as one piece of evidence for making inferences about validity. Thus, the authors evaluated 13 common standardized language assessments in English that were published in 2010 or later; see Table 1. The motivation for selecting more recently published assessments was, firstly, that the U.S. population data informing norming sample composition would more closely approximate current population demographics. Second, clinicians report using language assessments more if they are more recent (Betz et al., 2013). The motivation for selecting assessments in English was to reflect that about 92% CSD professionals in ASHA do not identify as bilingual service providers (ASHA, 2020). All assessments were measures that are commonly available in settings across research and practice (Betz et al., 2013).

The authors coded each assessment for six criteria: (a) age range of target examinees; (b) domains of language assessed; (c) reliability coefficients of composite or index scores;

(d) sample size, age range, and the U.S. population year used as a benchmark for the total norming sample; (e) availability of scoring rules for speakers of variants of English other than General American English, and; (f) clinical groups included in validity studies. Criteria (a) through (d) provided general information about assessments. Criteria (e) and (f) provided information on whether standardized language assessments utilized intersectional approaches to norming. For criterion (f), the authors considered the sample size, age range, and selection criteria of several clinical groups: Autism Spectrum Disorder (ASD), language disorder, intellectual disability, learning disability, hearing impairment, and speech sound disorder. For selection criteria, only specific information (i.e., cutoff scores, or the scores used to differentiate individuals with or without some characteristic, such as using −1.5 *SD* to differentiate individuals with and without language impairment) and not descriptive text (e.g., "borderline language impairment") was included, as descriptive text does not provide a metric that can be compared to other assessments.

General Information About Assessment Norming

Age Range, Language Domains, and Reliability—Given the scope of the learning exercise, all 13 assessments were for school-age individuals, eight of which were also for younger individuals (i.e., examinees within the birth to 3 range) and two of which were for older individuals across adulthood (i.e., criterion [a]); see Table 1. Six of 13 assessments addressed global language ability, and seven addressed specific language domains: oral and written language, narrative production and comprehension, social language, articulation, and expressive and receptive vocabulary (i.e., criterion [b]). As for criterion (c), 11 of 13 assessments reported sufficient reliability for making educational decisions ($r \geq .90$), and all had sufficient reliability for educational screening ($r \geq .80$; Salvia et al., 2016).

Total Norming Sample Demographics—Assessment total norming samples were typically representative of the U.S. population at the time when test norming took place in terms of geographic region, gender when defined as female and male, race, and parent socioeconomic status or education level (i.e., criterion [d]). Seven of 13 assessments used population data from 2013 or later, four of 13 used data from 2010, and two of 13 used data from before 2010; see Table 1.

The norming sample demographics are an area worthy of critical evaluation. On one hand, assessments using earlier population data as a benchmark may have norming samples that would not be representative of the U.S. population today. Preliminary 2020 Census findings indicate the U.S. population had significant differences in proportions of racial and ethnic groups from the 2010 Census; however, the Census in 2020 developed separate items on race and ethnicity, while the 2010 Census did not (United States Census Bureau, 2021, August 12). Hence, in considering whether test norming samples had sufficient representation of diverse examinees, examiners must also consider whether test norming data differ from current population demographics. If an examiner feels test norming samples materially differ from the current population, then examiners should justify in their decision to use test scores or not why they believe language assessment scores are or are not applicable. Furthermore, utilizing a binary of female and male does not reflect the real-world gender diversity that is imperative for examiners to support (ASHA, n.d., 2016). In all, these are two ways in which examinees can have intersectional identities that examiners need to consider in their evaluation of language assessment validity, interpretation, and use.

Intersectional Assessment Information

Scoring Rules for Multiple Variants of English—Across 13 assessments, eight had scoring rules for multiple variants of English (i.e., criterion [e]); see Table 1. The five assessments that did not have scoring rules for multiple variants of English also did not mention general considerations when assessing speakers of multiple variants of English in their manuals. The absence of this information may limit the validity of performance on a given language measure. Not having explicit scoring rules can mean that interpretation of assessment performance (i.e., scoring) can vary from one examiner to the next, thus potentially lowering the consistency of measurement. Clearly, examiners should not consider standardized language assessments to be broadly applicable to speakers of multiple speaker communities, nor should they utilize a standardized score from one assessment in place of holistic evaluation (Castilla-Earls et al., 2020). Nevertheless, determining whether test norming and standardized administration instructions reflect the linguistic identities of examinees is important for deciding how to characterize and use assessment performance.

Clinical Groups and Selection Criteria—While all 13 assessments included specific clinical populations in validity studies, some were more common than others (i.e., criterion [f]); see Table 1. Common clinical groups were those where evaluation of language abilities is important for diagnosis or differential diagnosis, such as language disorders (12 of 13), ASD (11 of 13), and learning disability (8 of 13). Less common clinical groups were hearing impairment (4 of 13), intellectual disability (6 of 13), speech sound disorder (5 of 13), social pragmatic communication disorder (1 of 13), developmental delay (4 of 13), and other health impairments (4 of 13).

As Table 2 shows, language assessments differed in their selection criteria and the information provided in the manual about the following clinical groups: ASD, language disorder, intellectual disability, learning disability, hearing impairment, and speech sound disorder. Some assessments used a sample of various clinical groups or referenced including clinical groups in development (see Table 1) yet did not provide specific details on each group (see Table 2). Further, some assessments that did not include a specific clinical population cited previous norming studies that showed good differentiating ability (e.g., Goldman-Fristoe Articulation Test; Pearson, 2015). Other assessments used developmental delay instead of specific diagnoses, as some children may receive services without a diagnosis in response to a general concern (Zimmerman et al., 2011). Overall, as the clinical groups that follow illustrate, examiners are limited in the inferences they can build about language assessment performance of an examinee relative to test norming if assessment manuals do not provide specific information.

Autism Spectrum Disorder: Ten of 11 assessments including individuals with a diagnosis of ASD in a validity study provided information on sample size (range: 20–125) and age range; see Table 2. Yet six of 10 assessments that included sample size and age range did not provide information on specific selection criteria in their manuals. Of the four assessments that provided information on selection criteria, three included individuals who performed ≥ −1.5 *SD* on an overall language assessment. As per the current diagnostic criteria for ASD, these individuals would qualify for co-occurring language impairment (American Psychiatric Association [APA], 2013; Tomblin et al., 1996). Failing to include a more heterogeneous group limits the ability to compare the performance of individuals with a formal diagnosis of ASD – but who may not score in the language impairment range – to the norming sample. In turn, one assessment included individuals with a NVIQ > 60. Given that a NVIQ < 70 indicates intellectual disability (APA, 2013), it is impossible to tell what proportion in the ASD group had co-occurring intellectual disability. At a broader level, six of 10 assessments included formal diagnostic information. All six included individuals with a formal diagnosis of ASD and co-occurring diagnoses, such as language impairment or intellectual disability. Four assessments did not provide information on whether individuals in the ASD group had co-occurring diagnoses.

Language Disorder: All twelve assessments including individuals with a language disorder in a validity study provided information on sample size (range: 25–248); see Table 2. Most assessments (10 of 12) included information on age range but did not report specific selection criteria (eight of 12). The four assessments that provided selection criteria used a cutoff of ≥ −1.5 *SD* on an overall language assessment. A majority (7 of 12) of assessments did not report whether the language disorder diagnosis was mutually exclusive with other diagnoses, four of 12 had a mutually exclusive diagnosis (i.e., specific language impairment), and one of 12 did not require the diagnosis to be mutually exclusive. While evaluating whether a test can differentiate individuals who vary by one identity only (i.e., language impairment) is relevant for psychometric development, it may limit real-world interpretation and use of assessment performance.

Intellectual Disability: The five assessments including individuals with intellectual disability in validity studies featured information on sample size (range: 14–54), but only three assessments provided information on age range; see Table 2. No assessment provided specific information on selection criteria, and only one assessment provided information on whether the diagnosis was mutually exclusive with other diagnoses. More information is necessary to understand the validity of the assessment for examinees with specific identities. On one hand, IQ is not a covariate of language ability in individuals with neurodevelopmental disorders (Dennis et al., 2009). At the same time, autistic individuals may only have co-occurring language impairment if intellectual disability does not better explain the nature of their language difficulties (APA, 2013); thus, assessing NVIQ is relevant for framing assessment performance.

Learning Disability: All eight assessments that included individuals with a learning disability in validity studies provided information on sample size (range: 15–162); see Table 2. Nearly all assessments (7 of 8) also included information on age range. In contrast, most assessments did not include information on selection criteria (6 of 8) or whether the diagnosis was mutually exclusive with other diagnoses (7 of 8). Two assessments used cutoffs of IQ ≥ reading or writing ability by 1 *SD* or for there to be a discrepancy between reading or writing ability (< 85) and scores in some other area (> 90).

Hearing Impairment: The four assessments with a hearing impairment validity study included information on sample size (range: 23–70). Three of four assessments also included information on age range. No assessment provided specific selection criteria in terms of quantitative cutoffs or whether hearing impairment was mutually exclusive with other diagnoses.

Speech Sound Disorder: Five of five assessments including individuals with a speech sound disorder in a validity study provided information on sample size (range: 19–90) and age range. Only one of five assessments provided information on selection criteria, using a cutoff score of ≥ −1.5 *SD*. Two of five did not require speech sound disorder to be mutually exclusive with other diagnoses, and three did not report whether speech sound disorder was mutually exclusive.

In all, the variability of the clinical groups in validity studies underlines the importance of caution in assuming that standardized language assessment performance is applicable, interpretable, and usable for diverse examinees. Even if assessments have high reliability and total norming samples that may seem representative, clinical group standardization samples may not facilitate the interpretation and use of assessment performance for diverse examinees with intersecting identities – especially if they are BIPOC. The take-home point is that examiners must consider test manual information as simply one piece of evidence among many.

Theory to Practice: Validity in Language Assessment of a Diverse Individual

To illustrate the applicability of intersectionality and DisCrit in building an evidence base for making assumptions about validity, we present one hypothetical profile of a diverse school-age individual with language needs. In addition to using more recently published tests, as per Betz and colleagues (2013), the authors used their experience as practitioners serving school-aged individuals and as researchers with experience in creating such profiles (Girolamo et al., 2022; Selin et al., 2019) to develop this example. Moreover, we use singular “they”, which some have suggested is an incorrect singular pronoun form that ought to be “she” or “he.” Use of singular “they” is consistent with style guidelines from the American Psychological Association (2020), which ASHA publications adhere to in use of bias-free language (ASHA, 2022), as well as guidance from the field of CSD (ASHA, n.d.; Shotwell & Sheng, 2021).

Example

A Black examinee is 15 years old with a diagnosis of Autism Spectrum Disorder (ASD). Their nonverbal intelligence quotient (NVIQ) is below 70. When they complete the Clinical Evaluation of Language Fundamentals-5th Ed. (Wiig et al., 2013) at school with a speech-language pathologist, their core language standard score is 40, their receptive language index standard score is 50, and their expressive language index standard score is 45. Thus, as per the manual, their receptive-expressive language difference scores are not significant. Throughout the assessment session, the examinee is engaged in assessment but talks about an unrelated topic toward the end of assessment: their special interest of doing laundry. They mostly speak in full sentences, some of which have two or more phrases (i.e., complex syntax): “After clothes come out of the wash, you have to put them in the dryer, low heat.” In addition, the examinee takes part in multiple conversational turns when the examiner probes for more information. The examiner learns that the examinee seems to come from a close-knit family and enjoys doing laundry for their parents, older sibling, and an intergenerational extended family member.

In this scenario, the first piece of evidence is performance on the CELF-5; see Step A-1 in Figure 1. The examiner must consider the *context* in which assessment takes place, or in a formal assessment session in school using General American English. Next, the examiner must consider the examinee’s identity and evidence in making interpretations about assessment validity (i.e., between Step A-1 and Step B-1). In terms of abilities, the examinee is autistic with intellectual disability but perhaps not language impairment, as their cognitive difficulties may better explain their language difficulties (APA, 2013). The examiner must also consider other aspects of identity that inform assessment, such as how the examinee and their family view their own cultural identity and diagnoses like ASD, as well as the examinee’s language background. For example, an examinee might be the child of immigrants, identify as Haitian, and see being autistic as central to their identity. They may also speak just one variant of one language (i.e., General American English) or multiple variants of English or multiple languages. All these identities can interact in nuanced ways and inform how they perceive – and perform on – standardized language assessment (Annamma et al., 2013; Crenshaw, 1991). It is the examiner’s job to proactively identify and understand these nuances.

As for validity evidence, the CELF-5 included a validity study on autistic individuals (Wiig et al., 2013); see Table 1. However, it is unclear whether the CELF-5 norming sample included overlapping ASD plus intellectual disability plus language impairment groups; the broad generalization is that the ASD sample performed lower than the “typically developing” sample (i.e., non-autistic individuals with no known diagnoses or delays; Wiig et al., 2013); see Table 2. Further, the sample of ASD participants in the validity study (*n* = 69) may not be large enough to interpret the results relative to other autistic individuals. Considering that language in autism is highly heterogeneous (Magiati et al., 2014), interpreting assessment performance would require test development recognizing the full variability of the autistic population or providing more specific information on the subset of autistic individuals in norming. In turn, having insufficient information about whether individuals like this examinee were represented in the validity study and total norming sample limits the interpretations an examiner can make about an individual’s language ability based on this assessment performance; see Step B-1.

A second piece of evidence is the examinee sharing information about their special interest with the examiner (i.e., Step A-2). Unlike the first piece of evidence, assessment took place informally, in that the examiner probed for more information in a conversation that was *not* part of a standardized language assessment. In addition to the dimensions of the examinee’s identity above, here, considering the identity of the examinee in terms of their special interests is relevant and has implications for accumulating validity evidence (i.e., between Step A-2 and Step B-2). Given their ability in speaking about and engaging in social communication about their special interest, the CELF-5 may not assess the *construct* of overall language in the way it should for this examinee. Similarly, the *content* of the CELF-5 may not be meaningful to assessing overall language in this examinee. These are key considerations in an examiner arriving at an interpretation of an individual's language ability (i.e., Step B-2). Critically, these considerations should be reflected in the examiner's characterization of this examinee, whether in a clinical or research report.

Ultimately, evidence from both assessments – as well as other pieces of evidence (e.g., parent report; Castilla-Earls et al., 2020) – should inform an examiner's conclusions about their overall language ability; see Step C. In deciding how to translate *conclusions* about language ability into *decisions* (i.e., Step D), an examiner must again consider how an individual's identity relates to each component. For example, an examiner could self-reflect on whether the evidence they have collected about an individual's identity *supports* their conclusions about language ability; if not, perhaps collecting more is merited. Similarly, an examiner could ask whether the way they plan to *use* their conclusions to make decisions involving the examinee aligns to how the examinee sees their own intersecting identities. The take-home point is that in order to make valid conclusions about language ability, CSD professionals must proactively learn about an examinee's intersecting identities in building an evidence trail. If CSD professionals fail to do so, they will *not* make appropriate decisions, which has implications for perpetuating – and exacerbating – harm to culturally and linguistically diverse individuals with language needs (Annamma et al., 2013).

Summary

This case example demonstrates the role of the examiner in supporting the valid interpretation and use of assessment performance. Although beyond the scope of this report, we highlight a few other strategies that can inform use of standardized language assessments with culturally and linguistically diverse individuals (Evard and Sabard, 1979). One strategy involves developing new assessments (Evard & Sabard, 1979), such as the Diagnostic Evaluation of Language Variation (Seymour et al., 2018) or the Bilingual English-Spanish Assessment (Peña et al., 2018). Examiners can also break standardization to modify test items or responses on existing assessments (in addition to test manuals, see Castilla-Earls et al., 2020, for a review). A third strategy is to develop new norms for existing assessments (Evard & Sabard, 1979), such as by validating assessments for dual language learners (e.g., the Diagnostic Receptive and Expressive Assessment of Mandarin; Liu et al., 2016). Overall, in each scenario, examiners must still operate under Figure 1 to build an evidence base for validity.

Pathways Forward

Given the ways in which school-age individuals can have multiple identities and the ways in which race and disability can reinforce one another, one question is how to support fair and equitable assessment – and thus building inferences valid inferences about an individual's language ability – at the systems level. To recognize intersecting identities in language assessment of diverse individuals as the *norm*, which may include testing for each intersection in test norming as per Bauer and colleagues (2021), we propose a middle-out advocacy approach. This model includes: (a) CSD professionals and self-advocates with language needs; (b) examinees, and (c) test developers and organizations (Janda & Parag, 2013). In this approach, stakeholders in the middle can exert influence sideways, upstream, and downstream, while those at the bottom can exert influence upstream, and those at the top can exert influence downstream (Janda & Parag, 2013).

CSD Professionals as Middle Stakeholders

Clinicians, researchers and self-advocates with communication needs can exert influence in several ways. First, they may provide feedback *upstream* to test developers, organizations, and workplaces. As the case examples illustrated, if the identities of examinees, clients, or a given individual do not seem to be well represented in standardized language assessment norming, CSD professionals can advocate for more inclusive norming practices to test developers. They can also advocate for appropriate test use by clearly stating evidence and their interpretation of assessment performance in clinical and research reports, which in turn could help support more inclusive organizational policies.

Second, CSD professionals – and ideally self-advocates – can exert influence *downstream* and advocate for the ideas of diverse clients and examinees. The use of “ideally” denotes that self-advocates with communication needs – especially those who are culturally and linguistically diverse – may have less social power, and thus, others with more social power may not choose to hear their voices (Annamma et al., 2013). As in the first case example, if the examinee or caregiver expresses information about the examinee’s strengths not captured in standardized language assessments, the examiner can elect to bring that information to their workplace, whether a school or research study, and advocate for including that information in their interpretation and use of assessment performance.

Third, clinicians, researchers, and self-advocates can exert influence *sideways* by engaging with one another about assessment. Formal channels include ASHA boards, committees, and special interest groups. Informal channels include social media, forums and email distribution lists. Topics related to fair and equitable assessment could include application of theories in assessment, self-reflection on biases which can work against fair and equitable assessment, and advocacy for more inclusive assessment methods and instruments that reflect the strengths of diverse individuals.

Lower and Upper Stakeholders

Unlike CSD professionals, examinees – if they are not involved with practice or research beyond the assessment session and if they are not self-advocates – can exert influence only *upstream* by providing feedback to examiners. This is not to say that examinees are voiceless, but rather that examiners choose how to respond to their feedback. Further, unless examinees take on a role of self-advocate, they are most likely not a stakeholder in the middle.

At the top are organizations, academic programs, and test developers. Organizations, such as ASHA, set policies for its members regarding fair and equitable assessment, (e.g., ASHA’s Code of Ethics; ASHA, 2016). In turn, while academic programs must follow accreditation standards, they have leeway in deciding how to prepare the next generation of clinicians and researchers to serve multicultural populations. Finally, test developers determine test norming procedures and what information about test norming is in assessment manuals (e.g., reporting partial or all information on each intersection in norming; Bauer et al., 2021). Daub and colleagues (2021) have advocated for organizations like ASHA to adopt the validity framework of the American Educational Research Association (2014). We specifically advocate for the integration of DisCrit and intersectionality as lenses through which examiners establish a cumulative evidence base, with downstream effects for making inferences about validity and use of assessments. In all, stakeholders at each level must be a part of the pathway forward to intersectional assessment.

Conclusion

In reviewing unified validity, intersectionality, and DisCrit, and how they apply to assessment contexts, the takeaway of this clinical focus article is that examiners must utilize a nuanced perspective on diversity. Examiners must proactively consider race and disability, as well as other identities, as constructs in assessment to fulfill their ethical obligation to be nondiscriminatory professionals (Annamma et al., 2013; ASHA, 2016). This involves the recognition that technical knowledge of assessments and knowledge of validity juxtaposed against DisCrit and intersectionality are skillsets that can enhance the evidence trail to support valid conclusions about language ability. A broader need is that all professionals in the field of CSD must practice lifelong cultural humility. While one cannot be expected to be an expert in all aspects of diversity, appreciating the many ways in which diversity may show up in assessment, knowing where to access resources to learn more, and exerting influence when possible to advocate for more inclusive and ethical outcomes, is a workable solution that supports practicing at the top of the license (McNeilly, 2018).

Acknowledgements

This work was supported by T32DC000052 (Director: Mabel L. Rice), T32DC017703 (Co-Directors: Inge-Marie Eigsti and Emily Myers), and R324A170032 (PI: LeAnne Johnson).

References

American Educational Research Association. (1999). Standards for educational and psychological testing. Author.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). Author.

American Speech-Language-Hearing Association. (n.d.). Supporting and working with transgender and gender-diverse individuals. https://www.asha.org/practice/multicultural/supporting-and-working-with-transgender-and-gender-diverse-individuals/

American Speech-Language-Hearing Association. (2016). Code of Ethics. https://www.asha.org/policy/

American Speech-Language-Hearing Association. (2020). Bilingual service providers, year-end 2020. https://www.asha.org/research/memberdata/

American Speech-Language-Hearing Association. (2022). Guidelines for reporting your research. https://academy.pubs.asha.org/asha-journals-author-resource-center/manuscript-preparation/guidelines-for-reporting-your-research/

Annamma SA, Connor D, & Ferri B (2013). Dis/ability critical race studies (DisCrit): Theorizing at the intersections of race and dis/ability. Race Ethnicity and Education, 16(1), 1–31. https://doi.org/10.1080/13613324.2012.730511

Annamma SA, Jackson DD, & Morrison D (2017). Conceptualizing color-evasiveness: Using dis/ability critical race theory to expand a color-blind racial ideology in education and society. Race Ethnicity and Education, 20(2), 147–162. https://doi.org/10.1080/13613324.2016.1248837

Bachman LF (2005). Building and supporting a case for test use. Language Assessment Quarterly: An International Journal, 2(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

Bauer GR, Churchill SM, Mahendran M, Walwyn C, Lizotte D, & Villa-Rueda AA (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. SSM - Population Health,, 100798. https://doi.org/10.1016/j.ssmph.2021.100798

Betz SK, Eickhoff JR, & Sullivan SF (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. Language, Speech, and Hearing Services in Schools, 44(2), 133–146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Bowers L, Huisingh R, & LoGiudice CM (2016). Social Language Development Test-Elementary: Normative Update. PRO-ED.

Bowers L, Huisingh R, & LoGiudice CM (2017). Social Language Development Test-Adolescent: Normative Update. PRO-ED.

Buchanan NT, & Wiklund LO (2020). Why clinical science must change or die: Integrating intersectionality and social justice. Women & Therapy, 43(3–4), 309–329. https://doi.org/10.1080/02703149.2020.1729470

Carrow-Woolkfolk E (2011). Oral and Written Language Scales-2nd Ed. Pearson.

Castilla-Earls A, Bedore L, Rojas R, Fabiano-Smith L, Pruitt-Lord S, Restrepo MA, & Peña E (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. American Journal of Speech-Language Pathology, 29(3), 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Cooc N (2019). Disparities in the Enrollment and Timing of Special Education for Asian American and Pacific Islander Students. The Journal of Special Education, 53(3), 177–190. https://doi.org/10.1177/0022466919839029

Crenshaw K (1989). Demarginalising the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 140, 25–42. https://doi.org/10.4324/9781315582924-10

Crenshaw K (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review, 43(6), 1241–1499.

Daub O, Cunningham BJ, Bagatto MP, Johnson AM, Kwok EY, Smyth RE, & Cardy JO (2021). Adopting a conceptual validity framework for testing in speech-language pathology. American Journal of Speech-Language Pathology, 1–15. https://doi.org/10.1044/2021_AJSLP-20-00032

Dawson JI, Stout CE, & Eyer JA (2003). Structured Photographic Expressive Language Test (3rd ed.). Janelle Publications.

Denman D, Speyer R, Munro N, Pearce WM, Chen Y-W, & Cordier R (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. Frontiers in Psychology, 8, 1515. https://doi.org/10.3389/fpsyg.2017.01515

Dennis M, Francis DJ, Cirino PT, Schachar R, Barnes MA, & Fletcher JM (2009). Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. Journal of the International Neuropsychological Society, 15(3), 331–343. https://doi.org/10.1017/S1355617709090481

Dunn DM (2018). Peabody Picture Vocabulary Test (5th ed.). Pearson.

Durkin MS, Elsabbagh M, Barbaro J, Gladstone M, Happe F, Hoekstra RA, Lee L–C, Rattazzi A, Stapel-Wax J, Stone WL, Tager-Flusberg H, Thurm A, Tomlinson M, & Shih A (2015). Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. Autism Research, 8(5), 473–476. https://doi.org/10.1002/aur.1575

Eusebi P (2013). Diagnostic accuracy measures. Cerebrovascular Diseases, 36(4), 267–272. https://doi.org/10.1159/000353863

Fenson L, Marchman VA, Thal DJ, Dale PS, Reznick JS, & Bates E (2006). MacArthur-Bates Communicative Development Inventories (2nd ed.). Brookes Publishing.

Friberg JC (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

Gillam RB, & Pearson NA (2017). Test of Narrative Language (2nd ed.). Pro-Ed.

Gillborn D (2015). Intersectionality, critical race theory, and the primacy of racism: Race, class, gender, and disability in education. Qualitative Inquiry, 21(3), 277–287. https://doi.org/10.1177/1077800414557827

Girolamo T, Rice ML, Selin CM, & Wang CJ (2022). Teacher educational decision-making for children with specific language impairment. American Journal of Speech-Language Pathology, 31(3), 1221–1243. https://doi.org/10.1044/2021_ajslp-20-00366.

Girolamo TM, Rice ML, & Warren SF (2020). Assessment of Language Abilities in Minority Adolescents and Young Adults With Autism Spectrum Disorder and Extensive Special Education Needs: A Pilot Study. American Journal of Speech-Language Pathology, 29(2), 804–818. https://doi.org/10.1044/2020_AJSLP-19-00036

Grimm KJ, & Widaman KF (2012). Construct validity. In Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, & Sher KJ (Eds.), APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics (pp. 621–642). American Psychological Association. https://doi.org/10.1037/13619-033

Giwa Onaiwu M (2020). “They don’t know, don’t show, or don’t care”: Autism’s white privilege problem. Autism in Adulthood, 2(4), 270–272. https://doi.org/10.1089/aut.2020.0077

Goldman R, & Fristoe M (2015). Goldman-Fristoe Test of Articulation (3rd ed.). Pearson.

Hernández-Saca DI, Gutmann Kahn L, & Cannon MA (2018). Intersectionality dis/ability research: How dis/ability research in education engages intersectionality to uncover the multidimensional construction of dis/abled experiences. Review of Research in Education, 42(1), 286–311. https://doi.org/10.3102/0091732X18762439

Hresko WP, Reid DK, & Hammill DD (2018). Test of Early Language Development (4th ed.). Pro-Ed.

Hubley AM, & Zumbo BD (2011). Validity and the consequences of test interpretation and use. Social Indicators Research, 103(2), 219–230. https://doi.org/10.1007/s11205-011-9843-4

Individuals with Disabilities Education Act, U.S.C. § 1400 (2004).

Janda KB, & Parag Y (2013). A middle-out approach for improving energy performance in buildings. Building Research & Information, 41(1), 39–50. https://doi.org/10.1080/09613218.2013.743396

Kane MT (2001). Current concerns in validity theory. Journal of Educational Measurement, 38(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane MT (2006). Validation. In Brennan RL (Ed.), Educational measurement (4th ed., pp. 17–64). American Council of Education and Praeger Series on Higher Education.

Kane MT (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice, 23(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Liu X, De Villiers J, Lee W, Ning C, Rolfhus E, Hutchings T, Jiang F, & Zhang Y (2016). New language outcome measures for Mandarin speaking children with hearing loss. Journal of Otology, 11(1), 24–32. https://doi.org/10.1016/j.joto.2016.04.001

Loomes R, Hull L, & Mandy WPL (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. Journal of the American Academy of Child & Adolescent Psychiatry, 56(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, Furnier SM, Hallas L, Hall-Lande J, Hudson A, Hughes MM, Patrick M, Pierce K, Poynter JN, Salinas A, Shenouda J, Vehorn A, Warren Z, Constantino JN, DiRienzo M, Fitzgerald RT, Grzybowski A, Spivey MH, Pettygrove S, Zahorodny W, Ali A, Andrews JG, Baroud T, Gutierrez J, Hewitt A, Lee L-C, Lopez M, Mancilla KC, McArthur D, Schwenk YD, Washington A, Williams S, & Cogswell ME (2021). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2018. MMWR Surveillance Summaries, 70(11), 1. https://doi.org/10.15585/mmwr.ss7011a1

Magiati I, Tay XW, & Howlin P (2014). Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood. Clinical Psychology Review, 34(1), 73–86. https://doi.org/10.1016/j.cpr.2013.11.002

McCauley RJ, & Swisher L (1984). Psychometric review of language and articulation tests for preschool children. Journal of Speech and Hearing Disorders, 49(1), 34–42. https://doi.org/10.1044/jshd.4901.34

McNeilly L (2018). Why we need to practice at the top of the license: To demonstrate our true value and effectiveness, we need to maximize time spent delivering services we are uniquely qualified to provide. The ASHA Leader, 23(2), 10–11. https://doi.org/10.1044/leader.FMP.23022018.10

Messick S (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5–11. https://doi.org/10.2307/1175249

Messick S (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

Morgan PL, Farkas G, Hillemeier MM, Mattison R, Maczuga S, Li H, & Cook M (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. Educational Researcher, 44(5), 278–292. https://doi.org/10.3102/0013189X15591157

Nelson N, Howes B, & Andersson MA (2016). Test of Integrated Language and Literacy Skills. Brookes Publishing.

Newcomer PL, & Hammill DD (2019). Test of Language Development: Primary (5th ed.). Pro-Ed.

Peabody JW, Luck J, Glassman P, Dresselhaus TR, & Lee M (2000). Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. Jama, 283(13), 1715–1722. https://doi.org/10.1001/jama.283.13.1715

Peña ED, Gutiérrez-Clellen VF, Iglesias A, Goldstein BA, & Bedore LM (2018). Bilingual English-Spanish Assessment. Brookes Publishing.

Plante E, & Vance R (1994). Selection of preschool language tests: A data-based approach. Language, Speech, and Hearing Services in Schools, 25(1), 15–24. https://doi.org/10.1044/0161-1461.2501.15

Purpura JE, Brown JD, & Schoonen R (2015). Improving the validity of quantitative measures in applied linguistics research. Language Learning, 65(S1), 37–75. https://doi.org/10.1111/lang.12112

Robinson GC, & Norton PC (2019). A decade of disproportionality: A state-level analysis of African American students enrolled in the primary disability category of speech or language impairment. Language, Speech, and Hearing Services in Schools, 50(2), 267–282. https://doi.org/10.1044/2018_LSHSS-17-0149

Russell G, Mandy W, Elliott D, White R, Pittwood T, & Ford T (2019). Selection bias on intellectual ability in autism research: A cross-sectional review and meta-analysis. Molecular Autism, 10(1), 1–10. https://doi.org/10.1186/s13229-019-0260-x

Salvia J, Ysseldyke J, & Witmer S (2016). Assessment in special and inclusive education (13th ed.). Cengage Learning.

Selin CM, Rice ML, Girolamo T, & Wang CJ (2019). Speech-language pathologists' clinical decision making for children with specific language impairment. Language, Speech, and Hearing Services in Schools, 50(2), 283–307. https://doi.org/10.1044/2018_LSHSS-18-0017

Seymour HN, Roeper TW, & de Villiers J (2018). Diagnostic Evaluation of Language Variation-Norm Referenced. Ventris Learning.

Shotwell S, & Sheng L (2021). Making a Case for Studying Gender-Neutral Pronouns in Speech-Language Pathology. Language, Speech, and Hearing Services in Schools, 52(4), 1141–1145. https://doi.org/10.1044/2021_LSHSS-21-00021

Skiba RJ, Artiles AJ, Kozleski EB, Losen DJ, & Harry EG (2016). Risks and consequences of oversimplifying educational inequities: A response to Morgan et al. (2015). Educational Researcher, 45(3), 221–225. https://doi.org/10.3102/0013189X16644606

Skrtic TM, Saatcioglu A, & Nichols A (2021). Disability as status competition: The role of race in classifying children. Socius, 7. https://doi.org/10.1177/23780231211024398

Tomblin JB, Records NL, & Zhang X (1996). A system for the diagnosis of specific language impairment in kindergarten children. Journal of Speech, Language, and Hearing Research, 39(6), 1284–1294. https://doi.org/10.1044/jshr.3906.1284

United States Census Bureau. (2021, August 12). 2020 Census statistics highlight local population changes and nation’s racial and ethnic diversity. https://www.census.gov/newsroom/press-releases/2021/population-changes-nations-diversity.html

Voress JK, Maddox T, & Hammill DD (2012). Developmental Assessment of Young Children (2nd ed.). Pro-Ed.

Wiig EH, Semel EM, & Secord W (2013). Clinical Evaluation of Language Fundamentals (5th ed.) Pearson.

Williams KT (2018). Expressive Vocabulary Test (3rd ed.). Pearson.

Zimmerman IL, Steiner VG, & Pond RE (2011). Preschool Language Scales (5th ed.). Pearson.

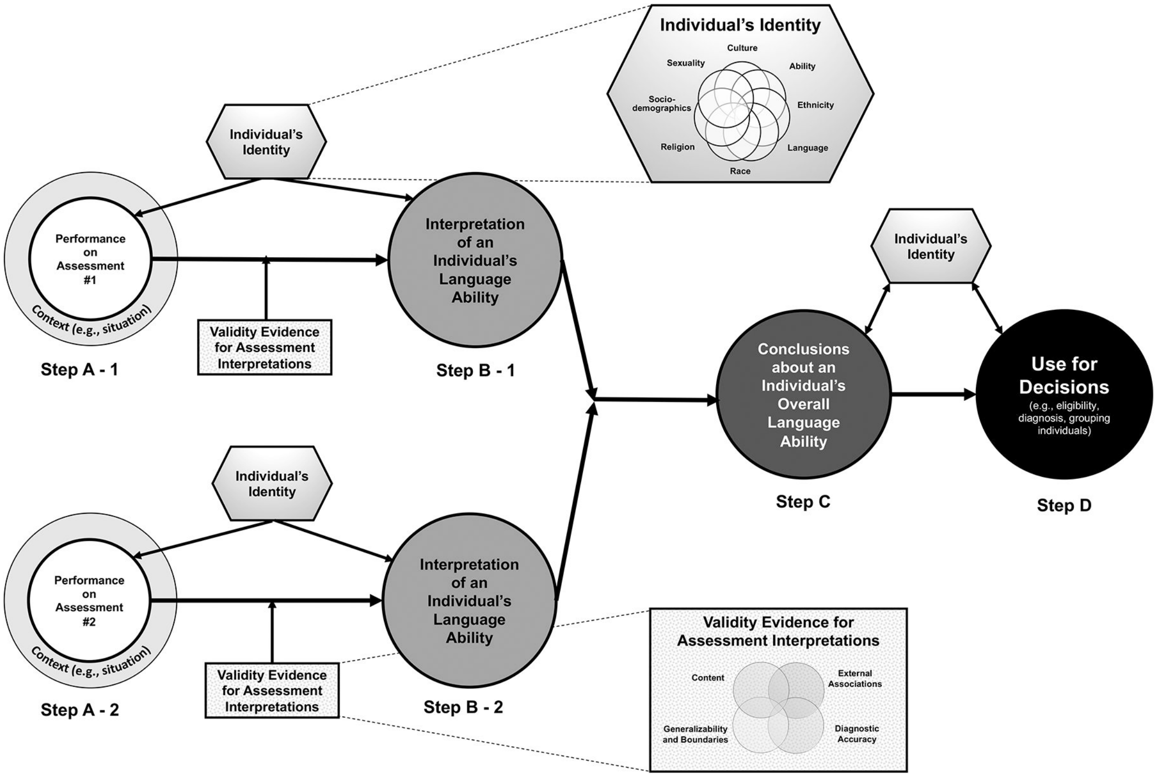
Individual's Identity
Culture
Sexuality
Ability
Socio-demographics
Ethnicity
Religion
Language
Race
Individual's Identity
Performance on Assessment #1
Context (e.g., situation)
Validity Evidence for Assessment Interpretations
Interpretation of an Individual's Language Ability
Step A - 1
Step B - 1
Individual's Identity
Conclusions about an Individual's Overall Language Ability
Use for Decisions
(e.g., eligibility, diagnosis, grouping individuals)
Step C
Step D
Individual's Identity
Performance on Assessment #2
Context (e.g., situation)
Validity Evidence for Assessment Interpretations
Interpretation of an Individual's Language Ability
Step A - 2
Step B - 2
Validity Evidence for Assessment Interpretations
Content
External Associations
Generalizability and Boundaries
Diagnostic Accuracy

Figure 1.

Proposed Pathway from Assessment Performance to Interpretation and Use of Performance When Drawing Conclusions about an Individual's Language Ability and Making Subsequent Decisions

Note. For the sake of space, we only present this pathway when two assessments (A & B) are conducted. We recognize, however, that there may be more depending on the scope of the assessment. In this case, we encourage examiners to consider the first part of the process (i.e., performance on assessment to interpretation of an individual's language ability) as repeating for each assessment included.

Table 1

An Overview of Select Language Assessments for School-Age Children

Test	Ages	Domains	reliability	Total Norming Sample			Scoring Rules for 1+ Variants of English	Special Populations Included in Validity Studies								
				n	Ages	Year		ASD	HI	LD	ID	SLD	SSD	SPCD	DD	OHI
CASL-2	3–21	receptive & expressive	≥0.95	2,394	3;0–21;11	2010	Yes	x	x	x	x	x		x	x	
CELF-5	3–21	receptive & expressive	0.96	>3000	5;0–21;11	2010	Yes	x		x	x	x				
DAYC-2	0–5	receptive & expressive	.89–.98	1,832	0–5.91	2010	No	x		x						x
EVT-3	2;6–90+	expressive vocabulary	0.97	2,720	2;6–90+	2017	Yes	x	x	x		x				
GFTA-3	2–21	articulation	.94–.97	1,500	2;0–21;11	2013	Yes						x			
OWLS-II	3–12	oral & written language	0.85	2,123	3;0–21;11	2009	Yes	x		x	x	x	x		x	x
PLS-5	0–7	receptive & expressive	0.91	1,400	0–7;11	2008	Yes	x		x	x	x			x	x
PPVT-5	2;6–90+	receptive vocabulary	0.97	2,720	2;6–90+	2017	Yes	x	x	x		x				
SLDT-E:NU	6–11	social language	0.94	1,002	6;0–11;11	2015	No	x		x						x
TELD-4	3–7	receptive & expressive	.95–.97	1,074	3;0–7;11	2015–6	No	x		x		x	x		x	
TILLS	6–18	oral & written language	0.92	1,262	6;0–18;11	2010	No	x	x	x	x					
TNL-2	4–15	narrative production & comprehension	0.9	1,310	4;0–15;11	2015–6	No	x		x		x	x			
TOLD-P5	4–8	receptive & expressive	0.8	1,007	4;0–8;11	2017	Yes	x		x	x	x	x			

Note. Ages = [years;months]. ASD = Autism Spectrum Disorder. HI = hearing impairment. LD = language disorder. ID = intellectual disability. SLD = specific learning disability. SSD = speech sound disorder. SPCD = social pragmatic communication disorder. DD = developmental delay. OHI = other health impairment. CASL-2 = Comprehensive Assessment of Spoken Language-2nd Ed. (Carrow-Woolfolk, 2017). CELF-5 = Clinical Evaluation of Language Fundamentals-5th Ed. (Wiig et al., 2013). DAYC-2 = Developmental Assessment of Young Children-2nd Ed. (Voress et al., 2012). EVT-3 = Expressive Vocabulary Test-3rd Ed. (Williams, 2018). GFTA-3 = Goldman-Fristoe Test of Articulation-3rd Ed. (Goldman & Fristoe, 2015). OWLS-II = Oral and Written Language Scales-2nd Ed. (Carrow-Woodfolk, 2011). PLS-5 = Preschool Language Scales-5th Ed. (Zimmerman et al., 2011). PPVT-5 = Peabody Picture Vocabulary Test-5th Ed. (Dunn, 2018). SLDT-E:NU = Social Language Development Test-Elementary: Normative Update (Bowers et al., 2017). TELD-4 = Test of Early Language Development-4th Ed. (Hresko et al., 2018). TILLS = Test of Integrated Language and Literacy Skills (Nelson et al., 2016). TNL-2 = Test of Narrative Language-2nd Ed. (Gillam & Pearson, 2017). TOLD-P5 = Test of Language Development-Primary, 5th Ed. (Newcomer & Hammill, 2019).

Table 2

Characteristics of Select Assessment Development Clinical Groups

Test	ASD				Language Disorder				Intellectual Disability				Learning Disability				Hearing Impairment				Speech Sound Disorder			
	n	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.
CASL-2	49	NR	NR	NR	72	NR	NR	NR	36	NR	NR	NR	43	NR	NR	NR	23	NR	NR	NR	-	-	-	-
CELF-5	69	5–21	NVIQ > 60	No	67	5–15	≥ −1.5 *SD*	No	54	NR	NR	NR	66	8–21	NR	-	-	-	-	-	-	-	-	-
DAYC-2[*]	NR	NR	NR	No	NR	NR	NR	NR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EVT-3	118	3–18	≥ −1.5 *SD*	No	220	3–18	≥ −1.5 *SD*	Yes	-	-	-	-	162	7–18	IQ ≥ 1 *SD* reading/ writing; other area > 90 & reading/ writing < 85	NR	70	3–18	NR	NR	-	-	-	-
GFTA-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54	4–8	≥ −1.5 *SD*	No
OWLS-II	24	3–21	NR	NR	60	3–21	NR	NR	30	3–21	NR	NR	20	3–21	NR	NR	-	-	-	-	90	3–21	NR	NR
PLS-5[*]	5[**]	3–7	≥ −1.5 *SD*	No	79	3–7	≥ −1.5 *SD*	NR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PPVT-5	118	3–18	≥ −1.5 *SD*	No	220	3–18	≥ −1.5 *SD*	Yes	-	-	-	-	162	7–18	IQ ≥ 1 *SD* reading/ writing; other area > 90 & reading/ writing < 85	NR	70	3–18	NR	NR	-	-	-	-
SLDT-E:NU[*]	125	6–11	NR	NR	227	6–11	NR	Yes	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TELD-4	19	3–7	-	-	78	3–7	NR	NR	-	-	-	-	24	3–7	NR	NR	-	-	-	-	74	3–7	NR	NR
TILLS	79	6–18	NR	NR	248	6–18	NR	NR	14	8–17	NR		-	-	-	-	40	6–15	NR	NR	-	-	-	-
TNL-2	-	-	-	-	25	7–14	NR	NR	-	-	-	-	15	8–15	NR	NR	-	-	-	-	19	5–13	NR	NR

Test	ASD				Language Disorder				Intellectual Disability				Learning Disability				Hearing Impairment				Speech Sound Disorder			
	n	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.	*n*	age	criteria	exc.
TOLD-P5	29	4–8	NR	No	173	4–8	NR	Yes	30	4–8	NR	No	41	4–8	NR	No	-	-	-	-	32	4–8	NR	No

Note. Exc. = mutually exclusive with other diagnoses. NR = not reported. - = population not included in development. CASL-2 = Comprehensive Assessment of Spoken Language-2nd Ed. (Carrow-Woolfolk, 2017). CELF-5 = Clinical Evaluation of Language Fundamentals-5th Ed. (Wiig et al., 2013). DAYC-2 = Developmental Assessment of Young Children-2nd Ed. (Voress et al., 2012). EVT-3 = Expressive Vocabulary Test-3rd Ed. (Williams, 2018). GFTA-3 = Goldman-Fristoe Test of Articulation-3rd Ed. (Goldman & Fristoe, 2015). MB-CDI-2 = MacArthur-Bates Communicative Development Inventories-2nd Ed. (Fenson et al., 2006). OWLS-II = Oral and Written Language Scales-2nd Ed. (Carrow-Woodfolk, 2011). PLS-5 = Preschool Language Scales-5th Ed. (Zimmerman et al., 2011). PPVT-5 = Peabody Picture Vocabulary Test-5th Ed. (Dunn, 2018). RESCA-E = Receptive, Expressive and Social Communication Assessment-Elementary (Hamaguchi & Ross-Swain, 2015). SLDT-E:NU = Social Language Development Test-Elementary: Normative Update (Bowers et al., 2017). SLDT-A:NU = Social Language Development Test-Adolescent: Normative Update (Bowers et al., 2017). SPELT-3 = Structured Photographic Expressive Language Test-3rd Ed. (Dawson et al., 2003). TELD-4 = Test of Early Language Development-4th Ed. (Hresko et al., 2018). TILLS = Test of Integrated Language and Literacy Skills (Nelson et al., 2016). TNL-2 = Test of Narrative Language-2nd Ed. (Gillam & Pearson, 2017). TOLD-P5 = Test of Language Development-Primary, 5th Ed. (Newcomer & Hammill, 2019).

*Assessments included broad clinical samples including children with developmental delays who may not have received diagnoses yet.

**PLS-5 referenced PLS-4 ASD sample (*n* = 88), as item sets were similar (Zimmerman et al., 2011).