Scientific Research Publishing

Psychology, 2019, 10, 181-198
http://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

L2/Ln Sign Language Tests and Assessment Procedures and Evaluation

Rosemary Liñan Landa[1,2*], M. Diane Clark[1]

[1]Department of Deaf Studies and Deaf Education, Lamar University, Beaumont, USA
[2]University of Texas at Rio Grande Valley, Edinburg, USA
Email: *rosemary.landa@utrgv.edu

How to cite this paper: Landa, R. L., & Clark, M. D. (2019). L2/Ln Sign Language Tests and Assessment Procedures and Evaluation. *Psychology, 10,* 181-198.
https://doi.org/10.4236/psych.2019.102015

Received: December 21, 2018
Accepted: January 30, 2019
Published: February 2, 2019

Abstract

The purpose of this review is to investigate sign language assessments in use as a tool for evaluating the linguistic abilities for job related requirements for bilingual professionals. Each corresponding sign language assessment will be discussed in terms of its original purpose, content, how it is used today, followed by a discussion of its psychometric properties via strengths and weaknesses. In some cases, when available, a discussion of the background in regards to test development will be given. Psychometrically sound assessments will allow a more consistent evaluation of the professionals who support the Deaf community. In terms of research, having normed measures will lead to an increase in research to improve pedagogical practices and the delivery of educational services.

Keywords

Sign Language Assessment, L2, Validity, Reliability

1. Introduction

There are different categories of sign language assessments, each unique in its purpose. For the most part, sign language assessments have been used when working with Deaf and hard of hearing (DHH) children. These assessments have primarily been used for educational purposes, measuring the progression of language acquisition across different age groups (Haug, 2005). In addition, there are also those who wish to assess ASL linguistic abilities for purposes of linguistic research; for example research investigating if there are differences between native signers and non-native signers on a specific linguistic feature (Haug, 2005). Finally, there are measures to evaluate sign language fluency, typically for purposes of job related requirements.

Here the focus is solely on investigating sign language assessments in use as a tool for evaluating the linguistic abilities for job related requirements for bilingual professionals. These professionals include any personnel positions requiring sign language proficiency, including teachers in bilingual programs, social service providers, and interpreters. These positions require sign language proficiency for communication with deaf students and deaf clients as they learn in school or interact with hearing people.

Overall, these assessments are described as designed to assess sign communication ability and include evaluating expressive and receptive sign language proficiency (Jacobowitz, 2007). Due to the variety of origins for developing each sign language assessment, their designs differ. Therefore, each corresponding sign language assessment will be discussed in terms of its original purpose, content, how it is used today, followed by a discussion of its strengths and weaknesses. In some cases, when available, a discussion of the background in regards to test development will be given prior to discussing the assessments strengths and weaknesses.

2. Theoretical Perspectives

Language assessments in general are designed to examine the linguistic abilities of test takers (Haug, 2005). Assessments in use among bilingual professionals tend to be criterion referenced assessments, which use Subject Matter Experts (SMEs) during test development to set-up the criterions for scoring. It is important to note that these criteria are not always standardized. With regards to sign language assessments, SME's consist of native users of sign language, such as members of the Deaf community, and/or those who possess near native fluency.

Psychometric properties, such as validity and reliability, provide evidence of the strength and/or weakness of each assessment (Rust & Golombok, 2009). Validity provides evidence that the measure accurately reflects the concept for which the measure was designed (Haug, 2005) and includes four types. Content validity determines if the assessment is accurately measuring the underlying construct of interest. Construct validity (or concurrent validity) is determined by the degree of correlation between a proposed assessment and another measure, formerly standardized, purporting to measure the same construct. Face validity evaluates if the test taker believes that the assessment measures the construct it claims to measure. Predictive validity is based on the ability of a measure to accurately predict the construct the assessment was designed to measure.

Then reliability relates to the degree of consistency of a measure and can be measured in several ways. Test-retest reliability is when the same assessment is repeatedly given to the same individual. Parallel forms can also be developed to show that scores on one form have high correlations with scores on the other form. Finally, split half reliability can be used to determine if an individual's scores on the even numbers of the measure are highly correlated with their

scores on the odd numbers.

Inter-rater reliability is a non-statistical way to determine reliability (Rust & Golombok, 2009). Raters must be trained so that they evaluate responses using the same criteria; they need to be periodically retrained on these scoring criteria to maintain high levels of reliability. Sign language assessments frequently attempt to establish inter-rater reliability, although it is difficult to obtain due to its reliance on subjective agreements among raters. Many, if not most, sign language assessments have difficulty training raters to obtain effective inter-rater reliability.

Finally, another important consideration is the assessment's usability. Usability refers to the procedures in the administering, scoring, and notifying test-takers of their results. In other words, how easy is it to use a particular assessment in relation to the time it takes to complete the process from scheduling a test date to receiving scores (Haug, 2005).

When evaluating the strengths and weaknesses of each of the following sign language assessments the preceding psychometric properties will be noted. Strengths will include those relating to the assessments validity, reliability, and ease of usability. Weaknesses will consist of those not addressed in the preceding strengths listed (Haug, 2005).

3. History of Language Assessment Practices

The Theory of Communicative Competence (Canale & Swain, 1980) brought about the American Council on the Teaching of Foreign Language's (ACTFL) adoption of the Oral Language Proficiency Interview (OLPI) assessment (He & Young, 1998), which is a discourse approach in which the assessment of communication skills occurs during an interview. Here, the test-taker is interviewed by a proficient user of the target language to determine both their receptive and expressive abilities (He & Young, 1998). In order to do well on the OLPI, a person will need a large enough lexicon to carry on a conversation, know how to repair conversational misunderstandings, have the ability to express themselves in appropriate syntactic structures, and be prepared to manage turns or shifts in topics of conversation (He & Young, 1998).

This type of assessment has been adapted to evaluate sign language proficiency levels. Some of the tests described below tend to be used for a more general demonstration of sign language fluency, while others are used for certifications required for specific professions. However, overall these tests are all used for bilingual professionals to demonstrate their sign language fluency.

3.1. Sign Language Proficiency Interview: American Sign Language (SLPI: ASL)

The SLPI: ASL Rating Scale (Caccamise & Newell, 2015a), formerly referred to as the Sign Communication Proficiency Interview (SCPI), similar to the OLPI, assesses communication skills. Candidate performance is compared to a stan-

dard scale of ASL proficiencies in an attempt to categorize the performance. The SLPI was created by the National Technical Institute for the Deaf, Rochester Institute of Technology (NTID) in New York and although the assessment was developed in America, it is currently used in other countries such as Canada, Kenya, and South Africa (Caccamise, Newell, & Mitchell-Caccamise, 1983; Maller, Singleton, Supalla, & Wix, 1999). Interviews are approximately 20 - 30 minutes in length and content is highly dependent on the background of each candidate. Interview questions range from basic factual questions to those posing hypothetical scenarios asking how one would respond or react (Caccamise et al., 1983). The interviews are video recorded to be rated separately by three highly proficient ASL raters followed by a discussion to agree on an Official Rating after the three individual Final Ratings are collected. Ratings consist of five ordinal levels: Superior, Advanced, Intermediate, Survival, and Novice, although now a plus may be earned at each of those levels (e.g. Survival Plus). Also, a No Functional Skills rating could be given if no ASL proficiency exists (Morere & Allen, 2012; "Rochester Institute of Technology ASL Training and Evaluation," n.d.). The range of levels exist to accommodate candidates from various proficiencies from native-like to no proficiency exists. Functional descriptions of what constitutes a rating in each of the categories are provided (Caccamise et al., 1983) and these were updated by Caccamise and Newell in 2015. Each of the descriptions address the relative linguistic competence, sign cohesion, depth of topics discussed, and adaptability displayed by the candidate to navigate and continue the conversation (Caccamise & Newell, 2015a).

During the 1980's there was little knowledge about the linguistic structures of ASL. In fact, the SLPI was originally used to rate a person's use of Contact signing and/or any signing style on its continuum (Lucas & Valli, 1992). Throughout the years, the SLPI has now become a measure of ASL proficiency. Currently, and in regards to contact signing, it is mentioned that since ASL and English have been in contact for over a century, a degree of lexical borrowing is anticipated and raters are trained to use their expertise in determining acceptable versus unacceptable linguistic borrowing.

Much of the strength of the SLPI comes from its face validity. The SLPI is a criterion referenced test created by SMEs who are highly proficient ASL users; possessing native or near-native ASL proficiency (Caccamise et al., 1983; Caccamise & Newell, 2015a). Interviewers are also qualified by native or near native ASL proficiency themselves, they are also said to have the experience and knowledge of sign varieties by differing regional locations (Morere & Allen, 2012). Caccamise and Newell (2010) provided procedures for scheduling and interviewing so that the SLPI-ASL would be administered consistently across interviewers, contributing to its reliability.

Recently, Caccamise (2008) has updated the instructions for monitoring raters' consistency for both inter-rater reliability and validity. He states that this monitoring will create fairness to interviewees taking the SLPI and will assist in

planning and providing SLPI Team Members with appropriate in-service training. An SLPI Coordinator has been charged with monitoring inconsistency across raters and is supposed to provide one-on-one training for a Team Member who is providing inconsistent ratings. Caccamise (2008) also recommends re-calibration of the SLPI Team Members every two to three years to maintain reliability. The structure of the teams has varied discussed next.

Caccamise, Newell, Lang, and Metz (1999; as cited in Caccamise, 2008) investigated consistency among the Three Rater Team Procedure. Here Final Ratings were found to be within one official rating of each other 72.2% of the time on the first Final Rating and this inter-rater reliability increased to 97.6% after a second round of Final Ratings. NTID elected to move to a one-member Rating Procedure between 2001 and 2003, which was found to be unreliable. Therefore, they returned to the SLPI Three Rater Team Procedure in the 2003-2004 academic year. Caccamise and Newell have evaluated the reliability and validity of the SLPI and now SLPI-ASL since the 1980s. They have created workshops, training materials, and rating forms to stabilize and validate the assessment. However, this measure has never been published in a peer review format. Also, the appearance of reliability seems to occur in the consensus of scores among the three raters although reaching a consensus is not sufficient in establishing true inter-rater reliability. However, reliability is questioned by the fact that across interviews a person could do better or worse given the topic being discussed in each particular interview (Caccamise et al., 1983).

Additionally, the SLPI demonstrates usability by encouraging use of the assessment by other institutions. There is also a document encouraging institutions that will use the assessment to monitor the consistency of scoring among rater teams to establish inter-rater reliability. This document along with instructions for test-takers and one with tips for taking the SLPI is also provided (Caccamise et al., 1983; Caccamise & Newell, 2015b).

3.2. Sign Language of the Netherlands (NGT) Functional Assessment Instrument (NFA)

An adaption of the SLPI, the NFA is an assessment modified by the Hogeschool Utrecht, University of Applied Sciences in the Institute for Sign Language & Deaf Studies which houses Bachelor degrees for those wishing to either teach in Deaf Education or become interpreters (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). Since the NFA has been adapted from the SLPI, it is similar in many ways. The assessment consists of an interview and is interactive. The interviewer attempts to elicit an expressive sample of NGT use, to be analyzed at a later time. However, the assessment differs in the inherent grammatical differences between both sign languages (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). Also, only two raters are used and raters submit their results, independently, to a third person or supervisor to analyze and assign a score or language level. If there is some disagreement between the two raters the supervisor will arrange for a meeting between the three of them. Should the

disagreement continue, the supervisor then becomes the third rater and decides (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). The NFA also differs in that raters are given a rater sheet and on each, there are language level descriptors where raters have a space to note examples of participants demonstration of each descriptor. In order to pass each level, candidates must demonstrate all competencies in each level (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). Finally, the number of levels differ as well, the NFA grants 6 levels which coincide with the Common European Framework of Reference for Languages (CEFR). Therefore, the NFA has adapted the mapped learning outcomes to each level of proficiency (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014).

Strengths and weaknesses are similar to those found for the SLPI. Overall, an attempt at reducing subjectivity has been made by the additional use of the rater sheet ("European Centre for Modern Languages of the Council of Europe," n.d.). However, inter-rater reliability is still a weakness as raters subjectively score candidates. Finally, no psychometric testing has been completed.

3.3. American Sign Language Proficiency Interview (ASLPI)

The American Sign Language Proficiency Interview (ASLPI) has now become the most widespread ASL assessment in the United States (Morere & Allen, 2012). Similar to the SLPI: ASL, the ASLPI consists of a 20 - 25 minute interview. There are four parts to the interview a warm-up, level check, probing, and finally a cool down ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.). Probing is when the interviewer tries out a number of topics in the conversation that assists in identifying a test-takers' possessive communicative competence. Interviewers are said to seek the candidate's maximum potential for demonstrating their linguistic abilities ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.).

Interviews are video recorded and then scored separately by three raters. Raters are highly proficient users of ASL and are tasked with evaluating the relative grammar, vocabulary, accent or production, fluency, and comprehension of each candidate ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.). Raters take notes while evaluating each candidate and then each rater comes to an overall subjective score. The overall score indicates the candidate's level of accuracy, consistency, complexity, and flexibility of linguistic abilities ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.). Notes in hand, along with an independently determined score, the three raters meet to come to a consensus. Scores range from a 0 - 5 and in between each number some may earn an additional +, for example 0 or 0+ ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.). It takes six weeks from the time one has tested to receive results. These are provided via e-mail to expedite the turnaround ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.).

Regarding the validity of the ASLPI, face validity is high as the interviews are

done with highly proficient ASL users, with one improvement over the SLPI in that the ASLPI rates candidates on specific linguistic features of ASL ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.). On the other hand, content validity has not been established for this measure. Therefore, validity is still in question in that rating practices are primarily subjective. Raters are to evaluate a candidate's overall level of proficiency and while functional descriptions are available for each level of proficiency, those could be clarified further by adding specific illustrative behaviors raters are seeking at each level of performance (Morere & Allen, 2012). Differences in raters understanding of what constitutes each rating level could result in different scores when testing and retesting. While raters are trained, it is uncertain how often refresher trainings occur. Therefore, shifts in understanding of the scoring determinants could mean shifts in scoring over the years (Morere & Allen, 2012).

Raters evaluate test-takers independently and then meet, to come to a consensus score. When raters differ in assigning scores there are three methods used in coming to a consensus. One is by averaging the three individual scores. This system of resolving different scores assigned by raters is a weakness to the assessment as an ordinal number (as opposed to an interval or ratio) cannot be averaged when there are no decimal points in between 0 - 5 (Morere & Allen, 2012). Another method for coming to a consensus is the negotiation among raters; it could be said that the person with the strongest personality chooses the score. Finally, there is an option to have the interview rated by another set of raters although when re-administering the test there is no guarantee the same lack of agreement in scores will not happen once more (Morere & Allen, 2012). The process is established to focus on inter-rater reliability but no published results are available to determine the effectiveness of these strategies.

Usability is a strength of the ASLPI in that it can be administered nationwide at proctor sites. There are site requirements, protocols, and procedures to be adhered to such as making sure the candidate takes the test in an undisturbed area, the video is sufficiently clear, and that the proctor can provide assurance that the test integrity had been preserved ("Gallaudet University ASL Diagnostic and Evaluation Services," n.d.).

3.4. American Sign Language Discrimination Test (ASL-DT)

The ASL Discrimination Test (ASL-DT) was developed by a team at NTID, who recognized the potential to translate the NTID Speech Recognition Test (NSRT) into an ASL based assessment. The ASL-DT assesses candidates ability to discriminate between differences in phonological properties in ASL. ASL phonology consists of the location of a sign, its handshape, orientation, movement, and morphophonology (Bochner et al., 2016). Candidates must discern differences between two sentences, that may or may not present differing phonological characteristics or minimal pairs. The pre-recorded video stimuli, presenting ASL sentences, consist of 48 items (Bochner et al., 2016). Two pairs of ASL sentences

are presented at a time and called, Trial 1 and Trial 2. Each Trial will present a Standard Sentence and a Comparison Sentence. The ASL sentences are three to nine signs in length and are presented by three native signers, two males and one female (Bochner et al., 2016).

There are in fact six types of ASL sentences that may be presented. Five out of the six types of sentences will include minimal pairs by one of the five phonological characteristics in ASL; location, handshape, orientation, movement, and morphophonology. The sixth type of sentences is identical (Bochner et al., 2016). Candidates need to determine if the pair of sentences in Trial 1 and the pair of sentences in Trial 2 are similar or different. Candidates receive a correct score if they answer both Trial 1 and Trial 2 correctly. Based on the number of correct responses, candidates will be categorized as either possessing a proficiency level of High, Intermediate, or Low (Bochner et al., 2016).

The ASL-DT appears to have strong evidence for validity and reliability. Construct validity was determined during test development by comparison with data on native signers. Here, the comparison was with these native signers' phonological processing and then a hierarchy of difficulty for each of the phonological characteristics was established. Based on this hierarchy, test items were proposed, an item analysis was performed and the test was developed (Bochner et al., 2016). Also noteworthy is the manner in which the ASL-DT garnered evidence for the predictive validity of their assessment. Participants (who included faculty, staff, and students of NTID) were predicted to have either a Low, Intermediate, or High level of ASL proficiency based on their experience using ASL. For example, native users of ASL, including Deaf and hearing CODAs (Child of Deaf Adults), were predicted to perform at the High proficiency level, while those at the Intermediate level were students majoring in ASL/English Interpretation. Finally, those predicted to be at the Low proficiency level were in beginning ASL classes (Bochner et al., 2016). The ASL-DT was able to accurately classify the majority of these participants regarding their relative ASL proficiency levels (Bochner et al., 2016).

Also, removing any subjective evaluations, the scoring consists of simply correct or incorrect responses. Further, to remove any correct responses due to chance, two pairs of sentences are presented at a time and candidates must answer both of them correctly in order to receive credit.

The ASL-DT while relatively new is planning on expanding. A possible computer based test (CBT) is in development which will allow the test to be taken in various locations, increasing usability. Also, the pool of test items is currently being increased from 48 to 350. It is purported that once the CBT becomes available it will contain 35 items and take only ten minutes to administer ("Sign Language Assessment," n.d.). A final note mentioned by the ASL-DT developers is that an assumption was made in regards to testing only the receptive abilities in candidates. Assuming similarity to spoken languages, in that receptive abilities correlate with expressive and overall language proficiency, the ASL-DT made the conscious choice to only assess receptive skills (Bochner et al., 2016).

3.5. Aachen Test for Basic German Sign Language Competence (ATG)

The ATG assesses basic communicative competence in German Sign Language (DGS). While the test may be used to assess language competencies in children, for the purposes of this chapter, discussion of how the ATG is used to evaluate the communication competence in adults working in the field of Deaf education and/or services will be the primary focus.

The ATG was developed by Rheinisch-Westfälische Technische Hochschule, a university in Aachen, Germany (Haug, 2005). The assessment is relatively lengthy as it seeks responses to a high number of prompts, which are divided into nine tasks. Participants on average require four hours to complete all nine tasks (Haug, 2005). It should be noted however, that these tasks may be split up and given over a period of two days. Test instructions are standardized and given in DGS. They are said to be adapted to the language level of the participants although written standardized test instructions may also be provided (Haug, 2005). The ATG was developed by pinpointing specific linguistic structures of DGS and each task purposefully targets a set of linguistic abilities. Some of these tasks measure expressive DGS, others, comprehension, and finally, a couple requires alternating between both expressive and receptive responses (Haug, 2005).

Task 1 is an expressive piece which is similar to the SLPI and ASLPI in that candidates are interviewed so that their spontaneous expressions of DGS may be assessed. Interviewers ask candidates about certain everyday topics, which they are asked to expand (Haug, 2005). The spontaneous expression provided is videotaped and analyzed by raters at a later time (Haug, 2005).

Task 2 is an assessment of receptive skills. Participants are shown a video which consists of a picture, set to an unspecified duration, followed by five to eight sign choices. Out of the five to eight choices, there will be one sign that represents part of the meaning and then another sign which represents the meaning in its entirety (Haug, 2005). The pictures consist of a combination of objects, animals, and situations. There are a total of 60 questions (Haug, 2005). After each prompt a participant has to rate on a four point scale whether each sign represents the meaning of the picture. In this way, participants are being tested at the lexical level.

During Task 3 participants must both receptively comprehend a prompt and then express their comprehension through mimicry. First, participants are shown a video with a signed phrase then asked to use a set of dolls to mimic the signed phrases (Haug, 2005). Here, participants must have knowledge of DGS's lexical, morphological, and syntactic structures. Spatial grammar is evaluated through this task.

Tasks 4 and 5 are essentially two parts to the same task. Participants are asked to view 60 different picture cards. After viewing each card they must first describe what they saw, then, provide the lexical name for the picture (Haug, 2005). During their description, the grammatical use of DGS is analyzed. When

participants fail to provide enough description or if they have not completed the second part of naming the picture, they will be given a prompt allowing them to try again (Haug, 2005).

Task 6 is similar to Task 3 in that participants must both receptively comprehend a prompt followed by expressing their comprehension through acting out the phrases with dolls. The prompts consist of eight to 11 phrases, presented one at a time. Each phrase is shown twice and the difference in Task 3 and Task 6 lies in the fact that Task 6 does not explicitly mention all nouns. Some phrases include pronouns instead (Haug, 2005); therefore, participants must rely on their ability to understand the principle of simultaneity and the use of space for pronominalization (Haug, 2005).

Tasks 7 and 8 also are similar and participants are presented with either single words or phrases, which they must repeat. In Task 7, there are 36 single signed utterances divided into three categories: single, compound (functioning as nouns), or predicate words (Haug, 2005). In Task 8, phrases have increasing levels of difficulty. These phrases are of four types: a main clause with spatial markers, a main clause with tense markers, a complex sentence with a conjunction incorporated, or a complex sentence with an explicit conjunction (Haug, 2005). Scoring is not verbatim and credit is given if the meaning is conveyed (Haug, 2005).

Finally, in Task 9 participants are asked to retell six stories; the content includes simple everyday occurrences, with an unexpected event as well as one or more DGS idioms. Participants may have up two attempts to complete this task (Haug, 2005).

Scores on the ATG are a percentage of items that are scored as correct. Scores for a native signer should be within the ninety percentile (Haug, 2005). The assessment appears to have construct validity, in that native signers are the raters. In addition, the assessment used a DGS language scale, currently in development, which was criterion referenced (Haug, 2005). One hundred individuals have taken the assessment. Most participants took Tasks 1, 2, 6, and 9. Pilot results have been used to modify the ATG (Haug, 2005). A weakness of the assessment is while it uses a language scale there are still no norms for the assessment as the language scale is in development. There are also no psychometric testing results available in the literature. Finally, as far as usability, the assessment has yet to be available. The assessment is undergoing development.

3.6. Registry of Interpreters for the Deaf (RID)

The Registry of Interpreters for the Deaf (RID) has ASL assessments in use for evaluating the linguistic abilities of ASL/English bilingual professionals to become certified interpreters. The RID assessment is a bit more complex in that it is a series of knowledge and performance exams. There are two types of tests offered by RID, the Certified Deaf Interpreter (CDI) exam for DHH individuals and the second for hearing candidates, the National Interpreter Certification

(NIC) exam.

A prerequisite to taking either of these exams is that one must have a Bachelor's degree (in any major) or educational equivalent ("RID," n.d.). For those who choose the Educational Equivalency Application (EEA) route, candidates must demonstrate experience in working as an interpreter and/or hours of interpreter related training. The EEA uses a point system to determine the amount of credit one will receive. Each experience credit is equal to one college credit. A total of 120 experience credits will then grant the EEA equivalency of a Bachelor's degree.

Once someone has met the education prerequisite, there are two separate assessments that must be taken prior to obtaining interpreter certification. These include Knowledge Exams as well as Interview and Performance Exams. The Knowledge Exams must be taken and passed prior to qualifying for the Interview and Performance Exam.

Also important to note is that there are two different Knowledge Exams followed by two different Interview and Performance Exams. An important distinction between the CDI Knowledge Exam and the NIC Knowledge Exam is that there is an additional forty hour CDI training requirement that must be satisfied prior to taking the CDI Knowledge Exam ("RID," n.d.). The CDI Knowledge Exam itself consists of 100 multiple choice questions and candidates can choose to take the exam in English or in ASL ("RID," n.d.). In either language candidates have three hours to complete the assessment. The CDI Knowledge Exam questions revolve around knowledge of the RID Code of Professional Conduct, interpreting issues and theory, the ability to mediate between Deaf and hearing consumers, and ASL linguistics ("RID," n.d.). These are similar to the NIC Knowledge Exam, however, CDIs in addition will be asked questions in regards to specialized topics specifically related to a CDI's profession. Examples include questions surrounding the ability to gesture, use props, draw, or any other relevant means of transmitting communication ("RID," n.d.). The exam contains four weighted domains of competence. Thirty seven percent focuses on the knowledge of how to render appropriate services, 33% examines knowledge of professional roles and responsibilities, 20% in preparation for service delivery and finally, 10% for knowledge of post-service procedures ("RID," n.d.). A score of 72 points or more indicates passing.

The NIC Knowledge Exam on the other hand, is administered in English over a computer. There are 150 multiple choice questions and questions again revolve around the knowledge of the RID Code of Professional Conduct, interpreting issues and theory, ability to mediate between Deaf and hearing consumers, as well as ASL linguistics. Test-takers have three hours to complete the exam and while it is stated that the answers are worth one point each and there is no penalty for a wrong answer, the scoring procedures also indicate they take the number of correct answers and convert them to a scale of scores ranging from a 200 - 800 ("RID," n.d.). Scores at or above 500 indicate a passing score.

Once Knowledge Exams are passed, candidates are given a period of five years from passing scores to take the next exam, the Interview and Performance portion. The CDI Interview and Performance Exam uses what they term vignettes, which are presentations of mock settings to assess the candidate's ability to make ethically sound decisions, as well as effectively interpret, given a mock scenario. Candidates responses to the vignettes are video recorded. Each vignette is scored by three raters, sometimes three different raters per vignette ("RID," n.d.). The scores of each scenario are then combined and a candidate must earn a minimum unspecified score.

The NIC Interview and Performance Exam is highly similar to the CDI exams. They also use vignettes, although specific to this exam, and assess the candidate's ability to make ethically sound decisions as well as effectively interpret. There are a total of seven vignettes and candidates responses are also video recorded. Again, each vignette is scored by three raters, sometimes three different raters per vignette. Raters are said to be using vignette specific rubrics identifying critical components of the communication being relayed. Candidates can either receive a Pass, Borderline Pass, Borderline Fail, or Fail ("RID," n.d.). The scores of each scenario are then combined and a candidate must earn a minimum unspecified score. Should a candidate fail, feedback is offered. Feedback includes a candidate's score relative to the passing score as well as rater suggestions for improvement ("RID," n.d.).

RID has recently, in 2016, formed a Center for the Assessment of Sign Language Interpretation (CASLI), LLC to work on test administration and development ("RID," n.d.). In 2018, a Job Task Analysis (JTA) was completed and will impact test administration, development, and/or subsequent revisions of their exams.

The exams appear to have some validity, in that SME's have been involved in the test development, administration, and future revisions. Members of the CASLI Board of Managers and Testing Committee consist of both Deaf and hearing proficient ASL users, many with interpreter credentials themselves who occasionally review the test for accuracy.

A strength of the RID exams are that JTA's were performed by SMEs who developed and published the domains of competencies. Raters also consist of highly proficient ASL users and appear to reach a consensus in scoring agreement. On the NIC exam alone, a 79% rater agreement of scores was found showing an acceptable level of reliability ("Registry of Interpreters for the Deaf", 2013). A point biserial correlation was performed to measure the pass/fail rates across the seven vignettes and a moderate correlation was obtained ("Registry of Interpreters for the Deaf", 2013).

Much of the weakness this exam possesses revolves around the rating system. Raters use rubrics that indicate what constitutes a Pass, Borderline Pass, Borderline Fail, and Fail. Each vignette has its own rubric for scoring. While RID cannot share the rubrics for each of these scenarios, they have provided an example,

which indicates what sort of behaviors they seek. The example rubric indicates a Pass as a good interpretation including substantial details, following a Borderline pass as an acceptable interpretation rendered with some errors concerning the details in relaying the communication, a Borderline Fail as a weak interpretation with too many errors (especially those important to the communication), and a Fail as a poor interpretation filled with too many errors, showing that the message was not accurately conveyed ("RID," n.d.). The ratings are highly dependent on subjective evaluations. Also, there is no mention of any sort of training for raters on scoring procedures.

In addition in regards to usability, the website for obtaining information on what tests are offered, the procedures for taking them, and how to qualify were not in any sort of organized location. It is recommended they consider revising the contents within the website or provide candidates with a handbook of operating procedures. Finally, given the recent transition of testing and test development to CASLI, this exam may be undergoing major revisions especially since a JTA was ordered to be completed by 2018.

3.7. National Authority for the Accreditation of Translators & Interpreters (NAATI)

NAAIT is the accrediting body for Auslan Sign Language interpreters. There are three ways in which an interpreter in Australia can become credentialed. One way is by completing a NAATI approved course for Auslan Sign Language; another, is to have NAATI review your credentials and/or qualifications obtained from elsewhere, and finally, take and pass one of two NAATI assessments (Napier, 2004). Both NAATI assessments measure four things: fluency in Auslan Sign Language, knowledge of the Interpreting Code of Ethics, knowledge of cultural, linguistic and social issues within the Deaf community, as well as knowledge of professional behaviors ("NAATI," n.d.). In order to sit for either of the assessments, one must pre-qualify by holding a Bachelor's degree. Also, each assessment is similar in that they are recorded and scored at a later time (Napier, 2004). The assessments differ by the level of expectations; one may just be starting out as an interpreter and take the paraprofessional interpreter exam or one may sit for the professional interpreter exam.

The purpose of a paraprofessional interpreter exam is to certify individuals with basic interpreting skills for general purposes. These interpreters are not to be used in advanced interpreting situations. General purposes include general dialogue by which master linguistic ability is not required (Bontempo & Levitzke-Gray, 2009). The assessment is 30 to 40 minutes in length and each section contains a varying numbers of questions. In the social and cultural awareness section, four questions are posed. Two questions are posed in English and two questions are posed in Auslan. Participants must answer these questions in the same manner that the questions were posed. This section is worth five points ("NAATI," n.d.). The second section on ethics is worth five points and is han-

dled in the exact same manner as the social and cultural section ("NAATI," n.d.). Finally, the interpreting section consists of two dialogues that participants interpret. Both dialogues are approximately 300 words in length and are divided further into segments of approximately 35 words each. In Dialogue 1 participants are asked to consecutively interpret while Dialogue 2 is a simultaneous interpretation. Each dialogue is worth 45 points ("NAATI," n.d.).

In order to achieve a paraprofessional interpreter certification one must score a minimum of 70 points out of a hundred. More specifically, at least 2.5 points out of five in both the social and cultural section and the ethics section and score a minimum of 29 points per dialogue ("NAATI," n.d.).

The professional interpreter certification requires interpreting in a wide range of contexts, some of which require knowledge of specialized terminology such as medical or legal, all requiring advanced linguistic abilities (Bontempo & Levitzke-Gray, 2009). A paraprofessional interpreter must take the professional assessment within nine years of initial certification or risk losing their certification ("NAATI," n.d.). In order to receive a passing score on the NAATI, participants must score 70 points out of 100.

SMEs are made up of native and non-native signers and have to apply for rater positions (Bontempo & Levitzke-Gray, 2009). Also, there is mention of a third and fourth type of certification although there is no assessment for those mentioned. Further, a Conference Interpreter and Senior Conference Interpreter Certificate may be awarded (Bontempo & Levitzke-Gray, 2009) due to recent modifications implemented by the organization. Modifications in the future may include adding a mandatory training course and/or modifying assessments content and structure. Structural changes may allow for a free and literal simultaneous interpretation. Content wise, NAATI recently sent out a JTA and results are pending analyzation (Bontempo & Levitzke-Gray, 2009). Without further background information on how the assessment was developed and/or how participants are assessed, a discussion on the psychometric properties may not be discussed.

3.8. The Canadian Evaluations System (CES) Certificate of Interpretation (COI)

In order to qualify to take the COI candidates must first become a member of the Association of Visual Language Interpreters of Canada (AVLIC). AVLIC will only accept members who have graduated from an approved IEP ("AVLIC," n.d.). Once a member, there are three phases to the exam. Phase 1 and 3 are actual separate exams, while Phase 2 is called a preparation phase. Currently, Phase 2 is not available but it will consist of two workshops to assist candidates in preparation for Phase 3 ("AVLIC," n.d.). Phase 1 is a 73 multiple choice Written Test of Knowledge (WTK) with questions pertaining to the area of ASL/English interpreting. Phase 3 is a performance exam called the Test of Interpretation (TOI). The purpose of the TOI is to evaluate the ability of the candidate to interpret between ASL and English ("AVLIC," n.d.). Raters score the

assessment and look for message equivalence in the interpretations. The TOI includes an ASL to English section as well as an English to ASL section ("AVLIC," n.d.). Candidates are scored by three Deaf raters and three hearing raters, who are all certified interpreters. These two groups of raters score the measures independently and then come together to develop a consensus ("AVLIC," n.d.).

It is unspecified whether the assessment was criterion referenced, although there may be an assumption made given the assessments use of SMEs during rating. If SMEs assisted in the test development, face validity could be a strength; however, without further information is hard to evaluate. Also, there are concerns for reliability as raters are coming to a consensus which is a subjective score. Without further information on how the assessment was developed and/or how participants are assessed, a discussion on the psychometric properties is limited.

3.9. Council for the Advancement of Communication with Deaf People (CACDP)

Finally, the CACDP is an organization that offers Levels of Certificates in British Sign Language (BSL) and Irish Sign Language (ISL). CACDP offers a Level 1 Award, Level 2 Certificate, Level 3 Certificate, a Level 4 Certificate, and a Level 6 NVQ Certificate in BSL and ISL ("Signature," n.d.).

Each of these levels has been designed to match the United Kingdom's Occupational Language Standards. For example, a Level I Certificate specifies its purpose as teaching one who works with a Deaf peer and would like to have some conversational ability while Level 6 NVQ works as a qualified interpreter ("Signature," n.d.).

At each level a candidate is essentially taking an approved course or set of courses with an assessment at the end to ensure that learning objectives have been met. Exams vary from a conversational interview, a multiple choice exam, a presentation, and/or a written paper, and more ("Signature," n.d.). At each level the assessments differ and the proficiency of BSL or ISL is expected to increase. They also diversify the tracks students can take including interpreter, Deaf blind communication specialist, communication support for Deaf learners, modifier of written English for Deaf people, and lip speakers for the Deaf ("Signature," n.d.).

4. Future Trends

Many places of employment require their personnel to be proficient in sign language, which necessitates an assessment of their sign language skills that is psychometrically sound. Key practices that ensure psychometrically sound properties, such as the validity and reliability of an assessment, within these assessments, are frequently lacking. From the review, it is apparent that all current sign language assessments in use start with face validity as each relies on the use of SMEs. It is after this starting point, however, where designs begin to differ.

Specifically, the NAATI, COI, and CACDP interpreting exams have not published information detailing test development. The criterion used to organize the

assessments is not specified nor is their information on how participants' scores are valid. SMEs are used; however, the tests have no published information on the piloting, revision by item analysis, and/or validity that should have been established.

Where the ASLPI, RID, SLPI: ASL, and NFA assessments divert in terms of its psychometric properties, is in the area of reliability. Each of these exams is heavily reliant on subjective assessments of raters assessing global levels of proficiency. While the NFA tries to introduce a rater sheet to guide raters in their evaluation, the measure is still subjective. Where possible, scores should be objective, rather than subjective, and seek to eliminate rater bias. Without reliance on objective criteria, scores can shift across tests, across raters and/or parallel forms.

Finally, in terms of the ATG and ASL-DT assessments, they have attempted to establish psychometric properties. Beginning with the ATG, it is commendable that they have referenced a DGS language scale, although the norms for the scale have yet to be established. Also, although still in development, the assessment is undergoing piloting for which an item analysis may be performed. The ASL-DT assessment has been piloted, sampled (by a diverse population and across fluency levels), and has undergone item analyses that result in assessments containing construct validity. Additionally, establishing inter-rater reliability by the standardization of the procedures for testing includes how to present, evaluate, and especially important, score the assessments. Further, the ASL-DT has published its reliability correlations and has done an excellent job at attempting to establish predictive validity comparing its assessment with SLPI scores and across levels of ASL fluency based on the experience of the participants (Bochner et al., 2016).

There is much work to be done in the way of ASL assessments establishment of psychometrically sound methods. Those that have had a jump start with this procedure, such as the ATG and ASL-DT, will aid research in that assessment. Further, the ASL-DT is easier to score, requires less time to administer, and obtain results. Of all these ASL assessments, only the ASL-DT was published in a peer review journal. Note that much of the information is from websites and it would benefit the profession to have more of these assessments evaluated by peer review.

Finally, it is worth noting that the NFA assessment was able to match NGT texts to specific stages of student language development. The NFA took advantage of the Common European Framework of Reference for Languages (CEFR) to map student learning outcomes. The NFA was able to analyze three areas: morpheme sign rate, use of space, and use of non-manual markers. All areas seemed to play a role in determining the stages of language learning (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). Familiar vocabulary at which the morpheme was signed at a higher rate indicated students had met that particular language level. In order to meet a certain level students needed to have the appropriate vocabulary and sign morpheme rate for that sign, 90% of the time (Boers-Visker, van den Bogaerde, & van den Broek-Laven, 2014). In other words, if students knew 90% of a text's vocabulary, it was highly likely they

would master that text's language level.

Assessments in use for the purposes of linguistic research will assist in better understanding the stages of language learning. These psychometrically sound assessments will in turn allow more consistent evaluation of the professionals who support the Deaf community. In terms of research, having normed measures will lead to an increase in research to improve pedagogical practices and the delivery of educational services.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Association of Visual Language Interpreters of Canada (AVLIC). (n.d.). *Canadian Evaluation System.* http://www.avlic.ca/ces

Bochner, J. H., Samar, V. J., Hauser, P. C., Garrison, W. M., Searls, J. M., & Sanders, C. A. (2016). Validity of the American Sign Language Discrimination Test. *Language Testing, 33,* 473-495. https://doi.org/10.1177/0265532215590849

Boers-Visker, E., van den Broek-Laven, A., & van den Bogaerde, B. (2014). Determining Aspects of Text Difficulty for the Sign Language of the Netherlands (NGT) Functional Assessment Instrument. *Language Testing and Assessment, 3,* 53-75.

Bontempo, K., & Levitzke-Gray, P. (2009). Interpreting Down under: Sign Language Interpreter Education and Training in Australia. *International Perspectives on Sign Language Interpreter Education*, 149-170.

Caccamise, F. (2008). *SLPI Paper #9: Scoring the Consistency of Your SLPI Team Members' Rating.* https://www.rit.edu/ntid/slpi/sites/rit.edu.ntid.slpi/files/page_file_attachments/SLPIUseReliability.pdf

Caccamise, F., & Newell, W. (2010). *Sign Language Proficiency Interview (SLPI): American Sign Language (SLPI: ASL).* Scheduling and Interview Procedures. http://www.ntid.rit.edu/slpi

Caccamise, F., & Newell, W. (2015a). *Paper 2: What Is the SLPI Rating Scale*? https://www.rit.edu/slpi/faq

Caccamise, F., & Newell, W. (2015b). *SLPI Paper 5: SLPI Training Workshop (WS) & Follow up Services.* http://www.rit.edu/ntid/slpi

Caccamise, F., Newell, W., & Mitchell-Caccamise, M. (1983). Use of the Sign Language Proficiency Interview for Assessing the Sign Communicative Competence of Louisiana School for the Deaf Dormitory Counselor Applicants. *Journal of the Academy of Rehabilitative Audiology*, *16*, 283-304.

Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, *1*, 1. https://doi.org/10.1093/applin/I.1.1

European Centre for Modern Languages of the Council of Europe. (n.d.) *Sign Languages and the Common European Framework of Reference for Languages.* http://www.ecml.at/ECML-Programme/Programme2012-2015/ProSign/Assessment/tabid/1766/Default.aspx

Gallaudet University ASL Diagnostic and Evaluation Services. (n.d.) *ASLPI American Sign Language Proficiency Interview*. http://www.gallaudet.edu/asl-diagnostic-and-evaluation-services/aslpi

Haug, T. (2005). Review of Sign Language Assessment Instruments. *Sign Language & Linguistics, 8,* 61-98. https://doi.org/10.1075/bct.14.04hau

He, A. W., & Young, R. (1998). Language Proficiency Interviews: A Discourse Approach. In R. Young, & A. W. He (Eds.), *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency* (Book 14, pp. 1-24), John Benjamins Publishing Company. https://doi.org/10.1075/sibil.14.02he

Jacobowitz, E. L. (2007). A Look at Teaching Standards in ASL Teacher Preparation Programs. *Sign Language Studies, 8,* 4-41. https://doi.org/10.1353/sls.2007.0030

Lucas, C., & Valli, C. (1992). *Language Contact in the American Deaf Community*. San Diego: Academic Press Inc.

Maller, S., Singleton, J., Supalla, S., & Wix, T. (1999). The Development and Psychometric Properties of the American Sign Language Proficiency Assessment (ASL-PA). *Journal of Deaf Studies and Deaf Education, 4,* 249-269. https://doi.org/10.1093/deafed/4.4.249

Morere, D., & Allen, T. (2012). *Assessing Literacy in Deaf Individuals: Neurocognitive Measurement and Predictors*. New York: Springer. https://doi.org/10.1007/978-1-4614-5269-0

Napier, J. (2004). Sign Language Interpreter Training, Testing, and Accreditation: An International Comparison. *American Annals of the Deaf, 149,* 350-359. https://doi.org/10.1353/aad.2005.0007

National Authority for the Accreditation of Translators & Interpreters (NAATI) (n.d.). *Accreditation in Auslan/English Interpreting*. https://www.naati.com.au/media/1105/auslan_english_interpreting_information_booklet.pdf

Registry of Interpreters for the Deaf (2013). *Assessing the Validity and Reliability of the NIC Examination: A Technical Report*. https://drive.google.com/file/d/0B-_HBAap35D1ZWxJY2JoSDRhTk0/view

Rochester Institute of Technology ASL Training and Evaluation (n.d.). *Sign Language Proficiency Interview Services*. https://www.rit.edu/ntid/aslte2/

Rust, J., & Golombok, S. (2009). *Modern Psychometrics* (3rd ed.). New York: Routledge.

Sign Language Assessment: Tests of L2 Learning (n.d.). *American Sign Language Discrimination Test*. http://www.signlang-assessment.info/index.php/american-sign-language-discrimination-test.html

Signature (n.d.). *Career Paths*. http://www.signature.org.uk/career-paths