The Texas BEI Interpreter Certification Examinations: Exemplary Models of Valid and Reliable Certification Instruments for Interpreters for the Deaf

> *Validity is…the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the…score interpretations.*
>
> *Joint Commission of the AERA, APA & NCME, 1999, p. 9*

1. Introduction

In 2001, the Texas Department of Assistive and Rehabilitative Services, Office for Deaf and Hard of Hearing Services (DARS DHHS) contracted with the University of Arizona's National Center for Interpretation Testing, Research and Policy (UA National Center) to redevelop its licensing examinations for interpreters for the deaf. The goal was to produce tests that are empirically grounded, reliable, fair, and valid, thereby ensuring that the state's interpreters possess the minimum level of proficiency required to effectively discharge their responsibilities. Throughout the test development process, DARS DHHS and the UA National Center adhered to the highest standards of validity advocated by scholars, practitioners, and the organizations who govern sound test development practice. The resulting sets of exams have repeatedly proven to be highly effective in identifying proficient interpreters for the deaf at three credentialing levels, thereby supporting the enduring goal of equal access to communication, services, and education for the deaf community.

2. A Proven Model for Valid Interpreter Certification

The UA National Center is widely recognized for its innovative proficiency test development for spoken language interpreting. In 1979, the director of the UA National Center was the key expert in the development of the Federal Court Interpreter Certification Examination (FCICE), a pioneering effort that defined the field of court interpreting and served as a model for all subsequent high-stakes spoken language interpreter proficiency tests in the United States (González, Vásquez & Mikkelson, 1991).

The UA National Center's examination development process conforms to the universally accepted standards of constructing, administering, and evaluating employment tests as prescribed by the joint commission of the American Education Research Association, the American Psychological Association, and the National Council on Measurement in Education in their *Standards for Educational and Psychological Testing* (1999). DARS DHHS and the UA National Center implemented these comprehensive guidelines by:

1. Conducting a comprehensive job/linguistic analysis of the work performed by interpreters for the deaf.

2. Creating well-defined test specifications.
3. Writing and revising the test with experts in the field.
4. Pilot testing the initial instruments with a representative population.
5. Analyzing the exam data psychometrically for final revision and to establish criterion validity and cut scores.
6. Leading rater training conferences to ensure that test administrations met the exacting reliability requirements of high-stakes examinations.

The UA National Center's empirical model of test development, as well as its adherence to rigorous testing and linguistic guidelines, has repeatedly provided evidence for the validity of the interpreter certification tests it develops. This includes applying the model to establish the parameters of the work done by interpreters for the deaf.

3. Legal Defensibility: The End Product of the UA Interpreter Testing Model

The UA interpreter testing process meets the goals of producing legally defensible exams that validly and reliably select interpreters who are competent to work in a high-stakes setting. In 1980 the FCICE prevailed in a challenge to its validity that was brought in federal court after its initial administration, and it has never been successfully challenged since that time (Seltzer V. Foley, 1980). On the strength of the empirically-based development process employed in the test's construction, the Honorable Judge Milton Pollack ruled that:

> *The said tests…bore a rational and proper relation to skills…required for requisite precision interpretation by bilingual interpreters in courtroom settings…. (p. 608)*

The same legally defensible empirical development, validation, and administration processes employed in the FCICE examinations have been used to create interpreter credentialing tests in several languages at the municipal, state, and government agency level, including the DARS DHHS licensing tests.

4. Empirical Foundation of the Texas BEI Interpreter Certification Examinations

The validation process begins well before the development of testing instruments; an empirical investigation is first made into of the nature of the work done by interpreters for the deaf, as defined by the Knowledge, Skills, Abilities, and Tasks (KSATs) required of proficient interpreters in the course of their work.

Establishing the KSATs requires conducting a Job/Linguistic Analysis of the work done by interpreters for the deaf. The Job/Linguistic Analysis conducted for the BEI used multiple sources of data, which were then compared to each other, allowing for corroboration and cross-referencing of the identified KSATs, thereby producing extremely robust evidence for their valid inclusion in the examinations. The sources of data included:

1. a thorough review and analysis of the previous BEI licensing tests and their supporting documents,
2. an extensive collection and analysis of data from incumbent interpreters in Texas, and
3. substantial input from an Expert Panel, consisting of master interpreters, interpreter trainers, certified deaf interpreters, deaf community members, and academicians.

The established KSATs were then operationalized as testing parameters to be built into the resulting certification process and instruments. In this way, the entire certification process and tests reflect a representative sample of the KSATs that are fundamental to the work of interpreters for the deaf. This, in turn, ensures high construct and content validity and allows candidates' test results to be an accurate, empirically-based measure of their ability to perform the work of an interpreter for the deaf.

4.1 Expert Panel Participation Essential for the Creation of a Valid Testing Process

The Expert Panel consisted of BEI members charged with the responsibility to oversee the certification process to ensure that it would meet the needs of all stakeholders. The Expert Panel included a diverse array of stakeholders, including not only working master interpreters but also interpreter trainers, certified deaf interpreters, deaf community members, and academicians, as well as other subject matter experts. The work of the Expert Panel included:

- Oversight and approval of all project activities **at every stage of development.**
- Providing subject matter expertise to identify essential **observable** and **underlying psycholinguistic/linguistic** KSATs.
- Defining key empirical questions to be addressed through the development process, such as: *How many certification levels are required to serve the needs of the deaf community in Texas?* and, *What is the performance standard required of an interpreter at each level?*

The extensive input of the Expert Panel, and diversity of the stakeholders they represented, helped ensure that the job/linguistic analysis was comprehensive and that the resulting certification process and tests accurately reflected the needs of the community and assessed the requisite proficiency of interpreter candidates fairly, accurately, and consistently. The Expert Panel's work complemented the data gathered through the survey of incumbents and provided exceptionally strong evidence for the validity and reliability of the resulting certification process and tests, which has been confirmed through several testing cycles in Texas.

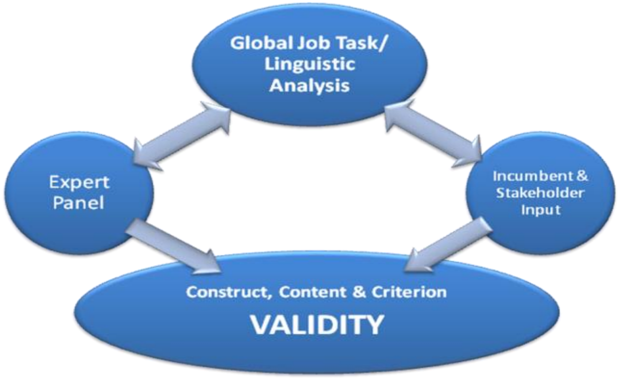
4.2 Survey of Incumbents

A large, representative sample of Texas' incumbent interpreter population was surveyed regarding the nature of the work they undertake. The sample was stratified to ensure that it was representative of all five BEI certification levels that existed at the time. The responses were also analyzed for demographic bias and were found to reflect the make-up of Texas' incumbent interpreter population in terms of gender, ethnicity and geographic spread. An excellent response rate was achieved for each certification level, indicating that the survey response sample is representative of the work done by the larger population of incumbent interpreters.

The extensive survey produced over 300,000 data points which were then subjected to detailed quantitative and qualitative analyses. The survey gathered data on **(1)** the backgrounds of the incumbents; **(2)** the settings in which they work; **(3)** the tasks performed in each setting; and **(4)** the knowledge, skills, and abilities required to complete those tasks. It further included data on the frequency and criticality of the KSATs for each setting.

4.3 Examples of the Interaction of Expert Panel and Incumbent Data

Expert Panel and incumbent data sources worked together to inform the content, structure, and format of the resulting certification process and examinations, providing robust support for their validity. This was further corroborated by the piloting process and additional incumbent feedback gathered during that process, which will be discussed below.



Overview of the Validation Process

The collected data was used to positively identify essential KSATs, to operationalize them as testing parameters, and to guide policy decisions ensuring the validity and fairness of the resulting testing instruments. A few examples can illustrate this process.

1. **Interpreting Modes**
   a. **Sight Translation**—The data clearly indicated that incumbent interpreters regularly engage in sight translation, which, in this case, is the rendering of a written English document into ASL. Because of the criticality of this interpreting mode, and the fact that it requires different KSATs than other interpreting modes to perform competently, the Expert Panel decided that it should be tested and weighted in proportion to its frequency.
   b. **Interactive**—The data also showed that incumbent interpreters regularly engage in interactive interpreting; i.e., interpreting a dialogue between an English speaker and an ASL signer. However, the BEI had removed the interactive section from its original exams, based on their experience and the input of independent subject matter experts. Based on the incumbent data, a further analysis of BEI test results that included an interactive section was performed. It

found that the interactive sections did not contribute additional validity to the inferences that could be drawn from the test results about candidates' proficiency, indicating that the interactive section did not assess KSATs unique to that interpreting mode, but that the required KSATs were effectively assessed in the sign-to-voice and voice-to-sign sections. As a result, the Expert Panel decided to continue excluding this interpreting mode for purposes of fairness, because testing it would add an additional burden for the candidate without yielding important data.

2. **Levels of Certification**—the data indicated that Texas' original system of five levels of certification could be streamlined to three levels. The work of incumbents clearly fell into three distinct groups, both in terms of setting and the level of proficiency required to work in those settings. The Expert Panel decided to reduce the number of certification levels to three in light of this data, eliminating redundancy that did not serve the deaf community and producing a more efficient certification process.
3. **English Proficiency**—it was determined that language proficiency was the most reliable predictor of incumbents' ability to understand English at the required level. As such, the Expert Panel decided to structure the certification process to include a test of English proficiency, which serves both as a screening tool and assesses essential KSATs.

As these examples indicate, there was significant and sustained interplay between the quantitative and qualitative analysis of a broad range of empirical data and the depth of experience and expertise provided by the Expert Panel. Moreover, these are just a few of the literally hundreds of determinations made by the Expert Panel in establishing KSATS, operationalizing them, and establishing sound policy. The result was a certification process and testing instruments firmly rooted in the nature of the work required of interpreters for the deaf that also fairly served candidates, the needs of the state of Texas, and the deaf community.

5. Final Certification Process for BEI English/ASL interpreters

The certification process that the Expert Panel decided upon consists of a two-stage process: (1) a written Test of English Proficiency followed by (2) an interpreting performance examination. Each certification level, Basic, Advanced, and Master, has its own performance test. Each of these tests is specifically calibrated to assess the specific, empirically determined KSATs at the appropriate level of complexity required for proficient interpreters at each level. Having a separate test for each certification level is essential to meet the demands of accurate, valid, and reliable assessment in a high stakes employment examination.

5.1 Written Test of English Proficiency

The written Test of English Proficiency (TEP) serves first as a screening tool to select candidates who have the greatest likelihood of success on the performance exams. Performance exams are more expensive to administer, so a screening tool makes the certification process more cost effective.

The TEP also functions as a measure of essential KSATs for qualified interpreters: the minimum level of English proficiency required for interpreters for the deaf. Moreover, it does so in a more effective manner than a performance test could, allowing for the assessment of a broader and more representative sample of a candidate's language proficiency in a shorter period of time. It thus provides better evidence of a candidate's proficiency in a format that is also fairer to candidates.

After its initial development, the TEP was piloted with 180 incumbent Texas interpreters who were representative of the larger population of interpreters in terms of ethnicity, gender, level of certification, and geographic spread. This type of representation was essential to provide strong evidentiary support for the validity of the pilot and its resulting analysis.

Following standard test development practice, the pilot consisted of considerably more test items than were intended for inclusion in the final test. The test items were divided among five distinct test sections: (1) Reading Comprehension, (2) Synonyms, (3) Usage and Grammar, (4) Sentence Completion, and (5) Antonyms. These sections are designed to assess candidates' proficiency at the lexical, syntactical, and discourse levels. Test items were written with strict attention to assessing the level of English proficiency identified in the job/linguistic analysis as the minimum level required for the work of interpreters for the deaf: 11th-12th grade level.

The pilot participants' responses were subjected to quantitative and qualitative analyses that scrutinized the test items, test sections, and the performance on the test as a whole. These analyses included item difficulty and discrimination measures, response coding, descriptive statistical analyses, intersectional correlation analyses, analyses of variance (ANOVA), and T-tests. On the basis of these analyses, the pilot test was reduced to include only the test items that best distinguished participants' levels of English proficiency. Similarly, the different sections were subjected to correlational analysis, which indicated that they assessed importantly different aspects of language proficiency. Finally, the reliability of the TEP was measured statistically using the Kuder-Richardson 20 (K-R 20) formula. The K-R 20 indicated that the TEP has an extremely high reliability (0.94 on a scale of 0-1, higher than 0.85 is considered exceptional)—which indicates that the TEP provides a highly consistent means of assessment of candidates' level of English proficiency.

Pilot participants were also extensively surveyed on their opinion of the TEP and its relation to the work they do as interpreters for the deaf. Overwhelmingly, they deemed the TEP to be representative of the level of proficiency required, as illustrated by some of their comments:

- *"The vocabulary sections were strong and diverse. This is such a vital area in day-to-day interpreting. I am glad to see it on the test. The grammar section [also assessed] another pitfall for interpreters that are vital in voicing."*
- *"I feel that the level of difficulty is sufficient for a basic English skills assessment and should definitely be accomplished successfully by anyone entertaining the interpreting field."*

- *"Overall, I would say the questions themselves are appropriate and sufficient enough to adequately grade/rate someone's level of proficiency in English."*
- *"Overall, I think that the test was very good and tested areas critical to the understanding of English that make interpreting into ASL possible."*
- *"I believe this test contained questions which were fair and represented areas of English proficiency which interpreters at all level should demonstrate."*

5.2 Interpreter Proficiency Examinations

Based on the findings of the job/linguistic analysis, the Expert Panel established three distinct levels of certification—Basic, Advanced, and Master—each with its own interpreter performance test, without which the validity and reliability of the certification process would be jeopardized. The first criterion for functional tests of proficiency is that the tasks included in the tests be based on a simulation of a real-life situation. The challenge in developing such tests lies in balancing authenticity with functionality, including constraints such as time and scorability. The empirical development procedure followed in creating these three exams ensures high validity in assessing the interpreting proficiency of candidates. Further, the careful training of test raters and the implementation of the UA's innovative objective scoring system provide strong evidence of the tests' reliability, and further enhance the evidentiary support for their validity.

Each test was designed to reflect the specific settings, modes of interpreting, and level of proficiency required for a proficient interpreter at that level, as determined empirically through the job/linguistic analysis. This design process conformed to the special requirements for language-intensive proficiency examinations (for example, Bachman & Palmer, 1996; Douglas, 2000). These requirements focus on the specifically linguistic, interactional features of the duties of interpreters, ensuring that the empirical basis for the tests is genuinely representative of the work performance by interpreters for the deaf. As such, all tests assess interpreting, transliterating, and sight translation. The complexity of each test is carefully controlled along three dimensions: **(1)** the speed of speaker/signer, **(2)** the complexity of topics/settings—as identified through the job/linguistic analysis, and **(3)** the complexity of the language used in the stimulus.

The scripting process for all three tests deserves special mention. Every section of every test is based on authentic discourse appropriate to the speaker, setting, and topic at hand. Each script underwent hundreds of hours of review and revision by the test developers, the Expert Panel, and other subject matter experts to ensure that each contained a representative sample of the relevant aspects of languages and language abilities required of interpreters, as identified through the job/linguistic analysis.

Each test was field tested with a stratified selection of three candidates. For each test, the three candidates were independently evaluated as "poor," "borderline," and "good" interpreters. These evaluations were then used as an independent criterion against which the candidates' performances on the exam were assessed. The candidates' performances on all three tests correlated very strongly with the independent evaluation of their proficiency, providing solid evidentiary support for the criterion validity of the tests.

The validity of the tests was further supported by the input of the candidates themselves. Specific to each level of the new exams, incumbents were asked to rank the following statements that directly relate to the exam's validity using a Likert scale from 1-5, with 5 being "agree strongly":

- The level of difficulty of the test is appropriate and reflects the interpreting proficiency required of a BEI-certified interpreter *of this level* in the course of his or her work.
- The complexity level of ***spoken*** *language* in the scenarios was appropriate and reflected what an interpreter of this level might actually encounter.
- The complexity level of ***signed*** *language* in the scenarios was appropriate and reflected what an interpreter of this level might actually encounter.
- The variety of sections is appropriate to assess the interpreting proficiency of candidates.
- The range of settings used in the scenarios is appropriate.

The candidates <u>unanimously</u> ranked <u>every</u> test as a 4 or 5 along <u>each</u> of the above dimensions. In other words, incumbent interpreters concluded that all three BEI interpreter performance tests authentically represented the nature and complexity of the work performed by proficient interpreters for the deaf. Below are a few representative comments:

- "The setting, subjects, and speakers were REAL."
- "All in all, the test was an excellent example of what … interpreters do on a regular basis!"
- "I truly felt challenged by the test and I feel that the data included represented what interpreters do."
- "Overall—Wonderful testing package."

<u>5.2.1 Objective Scoring System and Performance Test Reliability</u>

In addition to ensuring the validity of high stakes exams such as the BEI Interpreter Performance Exams, evidence for the reliability of the exams is essential. Adhering to detailed and professional test development standards has ensured that these performance examinations are highly reliable.

In addition, the implementation of the University of Arizona's objective scoring system greatly enhances the tests' reliability as well as their validity. In this system, a large number of specific, discrete language and interpreting features, identified by the job/linguistic analysis, are incorporated into the scripts. The Basic level exam includes 170 such features, the Advanced and Master exams each include 220.   Raters are trained to focus on these features, assessing candidates' performances on the basis of their rendering of only these features, according to specific and detailed guidelines. In this way, **(1)** candidates are assessed on the basis of a representative sample of the KSATs required of proficient interpreters, supporting the exams' validity, and **(2)** all raters evaluate candidates on exactly the same features, increasing the exams' reliability. While use of this scoring system requires considerable planning during test development, it results in a less demanding, more accurate and consistent assessment on the part of the raters.

The reliability of the tests was established during the rater training conference held following the tests' piloting. BEI raters were trained in the use of the scoring system and rated the pilot performance in teams of three. Whenever possible, each team included a deaf rater. There were approximately nine teams participating in the conference. Raters' scores were tracked and analyzed for both inter-rater and inter-team agreement, revealing extraordinarily high levels of agreement between the both the individual raters and the rating teams.

5.2.1.1 Criterion-Referenced Scoring

The BEI Interpreter Performance Exams, as well as the TEP, are criterion-referenced rather than norm-referenced exams, in terms of how their passing scores have been established. In norm-referenced exams, candidates' performances are judged relative to each other rather than against an external criterion that represents the minimum level of required proficiency, as is the case in criterion-referenced exams.

The passing score for the BEI exams was set empirically by the Expert Panel using the modified Angoff method, which is universally accepted as a valid means for setting passing scores for high stakes exams. This method balances the empirically established minimum level of required proficiency with fairness to candidates by taking into consideration criteria such as the following:

- Impact on the public health and welfare
- Consideration of test difficulty
- Standard Error of Measurement
- Characteristics of testing population
- Input of subject matter experts
- Comparability to other high-stakes performance exams
- Criterion validity from pilot performances

The result is a passing score that is fair, empirically based, and focused on ensuring that successful candidates possess the minimum level of proficiency required to responsibly serve the deaf community.

6. Concluding Remarks

The BEI certification process and tests have been meticulously crafted, with careful attention to ensuring their reliability, validity, and legal defensibility. Texas' implementation of this new system:

- Led to the overhaul of the certification system, reduced excessive variability, and ensured fair testing.
- Eliminated inappropriate competency standards and levels, thus improving delivery of service.
- Eliminated unreliable testing process and results.
- Set and improved educational standards for interpreters.

Most important, now that the BEI certification process has had 800 general administrations, it has proven itself efficient, cost-effective, and fair in certifying interpreters at three levels who are able to provide for the language service needs of the deaf community. In the end, this is also the strongest evidence of the validity of the BEI certification process.

7. About the University of Arizona National Center for Interpretation

The UA National Center is the major repository of the theoretical and practical aspects of specialized interpretation, its cognitive underpinnings, its ethical parameters, its best practice, its assessment, and the policy that guides its practice and assessment. The Center is committed to using this knowledge in support of its community—local, national and international—to ensure equal access to both civil rights and social services for all limited- and non-English members of the community, as it has for 30 years. Moreover, UA National Center is committed to doing so in a way which respects and supports the cultural and linguistic diversity that it sees as a profound asset within any community.

University of Arizona National Center for Interpretation Testing, Research and Policy
Roseann Dueñas González, Ph.D., Director
(520) 621-3615
ncitrp@u.arizona.edu

8. Works Cited

Bachman, L. and Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* Cambridge: Cambridge University Press.

González, R.D., Vásquez, V. & Mikkelson, H. (1991). *Fundamentals of Court Interpretation: Theory, Policy and Practice.* Durham, NC: Carolina Academic Press.

Joint Committee of the American Education Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Education Research Association.