Toward Extending the Educational Interpreter Performance Assessment to Cued Speech

Jean C. Krause
University of South Florida

Judy A. Kegl
University of Southern Maine

Brenda Schick
University of Colorado at Boulder

The Educational Interpreter Performance Assessment (EIPA) is as an important research tool for examining the quality of interpreters who use American Sign Language or a sign system in classroom settings, but it is not currently applicable to educational interpreters who use Cued Speech (CS). In order to determine the feasibility of extending the EIPA to include CS, a pilot EIPA test was developed and administered to 24 educational CS interpreters. Fifteen of the interpreters’ performances were evaluated two to three times in order to assess reliability. Results show that the instrument has good construct validity and test-retest reliability. Although more interrater reliability data are needed, intrarater reliability was quite high (0.9), suggesting that the pilot test can be rated as reliably as signing versions of the EIPA. Notably, only 48% of interpreters who formally participated in pilot testing performed at a level that could be considered minimally acceptable. In light of similar performance levels previously reported for interpreters who sign (e.g., Schick, Williams, & Kupermintz, 2006), these results suggest that interpreting services for deaf and hard-of-hearing students, regardless of the communication option used, are often inadequate and could seriously hinder access to the classroom environment.

Over the last few decades, educational services for children with hearing loss have changed substantially. Historically, most deaf and hard-of-hearing children in the United States were educated in residential programs, but since Public Law 94-142 was passed in 1975, these children have moved in large numbers to local public schools (Moore, 1992). As a result, it is estimated that more than 80% of K-12 students with hearing loss are now educated in the public school setting, and the majority of these students spend at least 40% of their day in regular classrooms alongside hearing students (U.S. Department of Education, 1999). This change has not only brought about an increased need for educational interpreters but also stimulated the development of tools to evaluate their performance. At the national level, the most notable development has been the Educational Interpreter Performance Assessment (EIPA; Schick & Williams, 1992).

Since its inception, the EIPA has served as an important research tool for examining the quality of educational interpreters. To date, studies using the EIPA have reported data for more than 2,000 educational interpreters nationwide. These studies (Schick, Williams, & Bolster, 1999; Schick et al., 2006) have been instrumental in identifying areas of need in the field of educational interpreting, revealing that the majority of interpreters who work in public school settings do not have sufficient skills to afford deaf and hard-of-hearing students adequate access to classroom communication. Without adequate access to this information, the value of inclusive education for many deaf and hard-of-hearing children is clearly diminished. As the Commission of Education of the Deaf (COED, 1988) pointed out in a 1988 report to...
Congress, federal law requires that “deaf students be integrated into regular classroom settings to the maximum extent possible, but if quality interpreting services are not provided, that goal becomes a mockery” (p. 103). Therefore, it is critical that any shortcomings in the skills of educational interpreters affecting access to classroom communication are identified reliably and addressed quickly. The EIPA, an established research tool with good validity and reliability (Schick et al., 2006), is well positioned to provide assistance in data collection efforts on both these fronts. In fact, analysis of data collected with the EIPA has already identified a number of skill areas that need improvement (e.g., Sign-to-Voice skills of interpreters working with younger children; Schick et al., 2006). As programs involving new models of interpreter training are introduced to address these skill areas (and others yet to be identified), the EIPA can also serve as an evidence-based mechanism for assessing the efficacy of these efforts, providing data that monitors changes in the quality of educational interpreters over time. Such data-driven information could help improve access to classroom communication for a great many deaf and hard-of-hearing children who use educational interpreters.

Cued Speech

CS, developed by Cornett (1967), is a system of manual signals (i.e., “cues”) designed to disambiguate phonemes confusable through speechreading alone. The cues are produced in synchrony with speech, or the visual mouth movements of speech, and consist of handshapes, representing groups of visually distinct consonants, combined with placements (located near the mouth), representing groups of visually distinct vowels. CS is a closed system: once the system has been mastered, it can be used to express any utterance that can be spoken. Consequently, CS interpreters can successfully convey unfamiliar vocabulary or foreign languages, often without prior knowledge of the topic or language. For these and other reasons, some parents of deaf children decide to use CS as their primary mode of communication.

The exact number of children using CS is unknown, but a national survey conducted annually by the American Annals of the Deaf (Program and Services Chart, 2003) showed that 14% of educational programs for deaf students offered a CS component in 2003. Moreover, the National Cued Speech Association reports that the vast majority of deaf children using CS are mainstreamed (Cornett & Daisey, 2001, p. 392), with a CS interpreter providing access to the auditory environment and voicing for the student as necessary. Because CS is not yet offered in interpreter training programs (ITPs), CS interpreters are often placed in classrooms after only minimal training and have little or no opportunity for evaluation and professional development. National certification opportunities are limited at best in most parts of the country, and no existing evaluation instrument for CS interpreters is geared exclusively to educational settings.

Although the number of educational interpreters using CS is relatively small, extending the applicability of the EIPA to include CS is appealing for several reasons. First, the EIPA is designed to be a flexible tool for evaluating what educational interpreters are expected to do in the classroom; it covers a wide range of communication options and is not limited to any one sign language or communication system. Broadening the applicability of the EIPA to include CS is thus a natural extension of its original design goals. Second, an extension to CS allows for additional research applications of the EIPA. Demographic data from the EIPA could provide information regarding the proportion of interpreters who use CS, and research aimed at analyzing the effects of language/communication system on interpreter skill would be more comprehensive because these studies could also include CS. In addition, data collected specifically on the skill levels of CS interpreters would be important for determining the typical quality of their services. This information is essential because virtually no data concerning the actual skills of these individuals is currently available. This type of data would also be helpful in assessing what special types of training are needed for this group or whether the current training methods are adequate. Finally, the No Child Left Behind
Behind Act (2001) mandates that only “highly qualified” staff can work with children in the public schools, and schools need some means of assessing the qualifications of interpreters using CS. Thus, tools are needed to evaluate all educational interpreters, not just those who use the most common communication options. Yet none of the evaluation tools currently available for interpreters who sign are directly applicable to those who use CS. As a result, states that need to evaluate interpreters who used the CS system must maintain a separate evaluation mechanism for them. In particular, states must look to one of two instruments that have been developed specifically for cued language transliterators by the Testing, Evaluation, and Certification Unit (TECUnit): (a) the Cued Language Transliterator National Certification Exam (CLTNCE), which has been administered nationally since 1988, and (b) the Cued Language Transliterator State Level Assessment, which has been available for purchase by state agencies from TECUnit since 1991. Available for nearly 20 years, these instruments are used by only about five states to establish minimum performance standards for educational interpreters who use CS (a handful of other states have introduced requirements calling for basic knowledge of the CS system, without any explicit evaluation of transliteration skills), and little change in the number of states with such laws has been apparent because the No Child Left Behind Act (NCLB) was passed in 2001. In contrast, the number of states requiring minimum performance standards for educational interpreters who sign has dramatically increased since NCLB was passed. Roughly 25 states now have some degree of performance standards in place for these educational interpreters, although some still rely on nationally recognized evaluation tools that were designed for adult community interpreting, such as certification from the Registry of Interpreters for the Deaf (RID) or the test designed by the National Association of the Deaf (NAD). With a growing awareness that specialized skills are required for interpreting classroom communication for children, however, states appear to be moving toward establishing performance standards based on the EIPA, a tool designed specifically for the educational setting. Of the 25 states with standards in place, 21 require a specified level of performance on the EIPA (many also set standards on content knowledge, continuing education requirements, and/or university degrees); in most of these states, a minimum score of 3.5 is required (Schick et al., 2006). At the national level, awareness of issues in educational interpreting is also growing, and the EIPA was recently adopted as a certification option by RID. As the EIPA is fast becoming the most widely adopted mechanism for evaluating educational interpreters, several states (e.g., Maine, Louisiana, New York) have expressed interest in a CS version of the EIPA.

Educational Interpreter Performance Assessment

It is generally accepted that interpreting for children in educational settings differs substantially from interpreting for adults, particularly when children are young and still acquiring language. Yet until the EIPA was established in 1992, there were virtually no instruments for evaluating interpreters that addressed issues pertaining to educational settings. Recognizing that many aspects of classroom communication are unique to K-12 education (see Schick, 2004), the EIPA is designed to evaluate an interpreter’s ability to convey pragmatic and prosodic information (why and how a message is delivered, respectively) as well as lexical information (what the message is).

When an EIPA is administered, two samples of an interpreter’s work are collected and submitted to the EIPA Diagnostic Center for evaluation: a Voice-to-Sign sample of the interpreter translating or transliterating spoken English in the classroom environment into sign communication and a Sign-to-Voice sample of the interpreter translating or transliterating what a deaf child signs into spoken English. Based on these samples, the EIPA evaluates 37 skill areas that fall into four broad evaluation domains: (a) Voice-to-Sign production (syntax, spatial grammar, and nonmanual aspects of prosody), (b) Sign-to-Voice production, (c) Vocabulary (range and depth of vocabulary, finger spelling, and numbers), and (d) Overall factors (aspects of interpreting that are discourse based, such as discourse mapping and cohesion). A three-member evaluation team, one of whom is deaf, rates the interpreter using a Likert scale of 0 (no observable skills)
to 5 (advanced skills) in each of these skill areas. Average scores are reported for each skill area and evaluation domain, and an overall average is computed. In addition, diagnostic feedback is provided detailing areas of strength and areas that should be targeted in professional development for each of the four evaluation domains.

As a measurement tool, the EIPA has been shown to have good reliability and validity (Schick et al., 2006). Interrater reliability is high, with correlations between teams ranging from 0.86 to 0.94 across the domains of evaluation. Coefficients of internal consistency within each domain are also high (ranging from 0.93 to 0.98), whereas interdomain correlations suggest that each domain taps a different aspect of an interpreter’s performance. As further evidence of validity, an examination of 42 interpreters with RID certification showed that individuals with that credential can be expected to score in the advanced range (4.0 or better) on the EIPA (Schick et al., 2006).

Initially designed to evaluate interpreters in real-life situations (interpreters filmed in their own classrooms), the EIPA has used videotaped stimulus materials since 2000 in order to standardize stimuli and facilitate meaningful comparisons of multiple EIPA scores (Schick & Williams, 2004). These videotaped materials are appropriate for evaluation of interpreters who use predominantly American Sign Language (ASL), typically viewed as the sign language of the adult Deaf community; predominantly Pidgin Sign English (PSE), the type of nativized English signing found among the adult Deaf community (see Davis, 2005 and Kuntze, 1990 for interesting discussions about nativized English signing); or Manually Coded English (MCE), the forms of English signing that were developed specifically to teach deaf students English in a more accessible form (e.g., Bornstein, 1990). For each of these communication options, two versions of the EIPA are available: one version is applicable to interpreters who work in elementary settings, and the other version is applicable to interpreters who work in secondary settings. Thus, the EIPA in its current form (see Schick & Williams, 2004 or http://www.classroominterpreting.org/ for a complete description of the EIPA tool and procedures) covers the majority of situations encountered by educational interpreters. If a CS version of the EIPA were also available at the EIPA Diagnostic Center, the applicability of the EIPA would be increased to an even greater number of educational interpreters. Such a development would be particularly useful for states seeking to evaluate CS transliterators in educational settings and would help ensure that an additional population of educational interpreters is held to appropriate competency standards.

In order to determine the feasibility of a CS version of the EIPA, a pilot CS test was developed and administered to 24 educational interpreters who work in K–12 settings, and the EIPA evaluation system was adapted to CS. Experts were consulted to establish appropriate evaluation rubrics, and two EIPA evaluation teams were established. This article examines the validity and reliability of the pilot test. In addition, it investigates the performance skills of educational interpreters who use CS, both in elementary and secondary settings.

**Methods**

**Participants**

The CS pilot test was administered to 24 CS interpreters from four states. Although interpreters were generally evaluated at only one grade level (elementary or secondary), one interpreter opted to take the test at both the elementary and secondary levels. As a result, 25 test performances were available for evaluation. Two states participated formally in pilot testing, and 21 of the 24 interpreters were from these two states. The remaining three interpreters were individuals who later trained to become evaluators. These individuals were never involved in evaluating themselves and had no exposure to test materials or evaluation procedures at the time of testing.

Demographic data for the 24 individuals who participated in pilot testing are summarized in Table 1. Most of the interpreters were Caucasian (92%), with one interpreter (4%) reporting membership in a minority ethnic group and one (4%) choosing to provide no information on race/ethnicity. A quarter of the individuals had either completed an ITP (12.5%) or earned a bachelor’s degree (12.5%). On average, interpreters reported 10 years of general interpreting
(SD = 6.9 years) experience and 8.4 years (SD = 6.0 years) of experience interpreting in an educational setting. Although a broad range of experience (minimum = 0.75 years, maximum = 24 years) was represented, median levels of experience were consistent with averages. These levels of experience are relatively high, representing about 2 years more experience on average than the educational interpreters who took the EIPA (ASL, PSE, or MCE) between 2002 and 2004 (Schick et al., 2006) and substantially more experience than the median experience (2–5 years) found in a survey of 222 sign language interpreters working in educational settings in the Midwest (Jones, Clark, & Soltz, 1997).

CS Pilot Test Development and Administration

In order to adapt the EIPA to include CS, it was first necessary to collect video materials appropriate for use in an EIPA evaluation. The EIPA scoring sheet was then modified to accommodate issues specific to CS interpreting, and the resulting Cued Speech EIPA pilot test (EIPA-CS) was administered to participants.

**Video materials.** For the EIPA, two types of video materials are required: (a) classroom materials designed to elicit the interpreter’s expressive product (usually Voice-to-Sign or, for CS interpreters, Voice-to-Cue) and (b) student materials designed to elicit the interpreter’s receptive product (usually Sign-to-Voice or, for CS interpreters, Cue-to-Voice). The classroom and student materials that were in use by the EIPA Diagnostic Testing Center at the beginning of the project period included two alternate test forms at both the elementary and secondary levels. Because the language or communication system used to transmit information to the deaf student is not likely to affect the types of classroom communications that must be conveyed by the interpreter (i.e., the oral language used by teachers and other students in the classroom), no modification to the EIPA classroom materials was required. The existing classroom videotapes (elementary classroom—Options A and B and secondary classroom—Options A and B) were used as the classroom materials for the CS pilot test.

The student materials, however, are specific to the language or communication system of the interpreter. Therefore, new student materials were required. To create the student materials, three deaf elementary students (0 girls, 3 boys; ages: 10 years) and three deaf secondary students (3 girls, 0 boys; ages: 13–16 years) whose primary mode of communication was CS were recruited from CS communities in the Midwest and New England. The children were interviewed individually for 60–90 min by the facilitator of the project, a proficient cuer. In order to ensure that the interview format was consistent with that used in creating the previous EIPA student videotapes for signing, Boys Town National Research Hospital (BTNRH) provided a videographer experienced in eliciting student materials for EIPA.

Because five of the six children were unfamiliar with the facilitator and none were familiar with the videographer, each interview began with roughly 15 min of activities designed to put the child at ease. Snacks were provided, and the child was given a tour of the recording area. The facilitator described her background in CS and her relationships with other cuers that were familiar to the child (e.g., a teacher at the child’s school). The purpose of the project was explained, and the child was given the opportunity to ask questions. Initial interview questions focused on
basic topics (e.g., family, favorite subject in school) until the child became comfortable with the camera and interview format. Questions then moved gradually from concrete to abstract topics, with the goal of collecting language samples that varied in length, contextualization, and complexity. The videographer assisted the facilitator in monitoring these aspects of the child’s responses; he suggested topics and follow-up questions as needed to elicit the necessary samples. He also recorded the child’s portion of the interview to digital videotape for subsequent processing.

Half of the interview was conducted with the child cueing and speaking simultaneously, and half of the interview was conducted with the child cueing without voice. That is, the facilitator and the child communicated silently by using cues synchronized with silent mouth movements. The reason for eliciting both of these modes is the varied nature of receptive cueing tasks facing educational CS interpreters and the changing expectations of the profession. Historically, most CS users have also had oral goals, and any receptive training that CS interpreters received was based on the expectation that they would have access to audiovisual (AV) information (i.e., that students would speak and cue simultaneously). However, there is now growing support for using CS to convey English in bilingual–bicultural programs (LaSasso & Metzger, 1998), where speech is not necessarily a goal. Consequently, an increasing number of settings will require educational CS interpreters who are capable of voicing CS presented in a visual-only (VO) modality, that is, students who cue with only silent mouth movements rather than with audible speech accompanying the cues. Despite the changing landscape, CS interpreters who do not possess this skill but have strong AV Cue-to-Voice skills are still adequate for many settings, and some settings (in which deaf students who use CS have oral goals and highly intelligible speech) may not require Cue-to-Voice skills at all. Therefore, it was determined that it would be most appropriate to include two receptive cueing tasks in the student video materials: AV Cue-to-Voice and VO Cue-to-Voice, at least as long as these separate skills remain required of some jobs but not others. By evaluating both AV and VO Cue-to-Voice performance, the EIPA-CS pilot test provides a mechanism for ensuring that a CS interpreter in a given job has the specific type of receptive skills necessary for that job.

After the student interviews were completed, the digital videotapes were reviewed by the authors to evaluate the quality and quantity of language samples elicited from each child as well as his/her presence on camera. Based on this information, it was determined that materials elicited from one secondary girl (age: 16 years) and one elementary boy (age: 10 years) were best suited for use on the student stimulus tapes. The materials from these children contained a sufficient number and variety (in length, content, contextualization, and complexity) of language samples so that it was possible to create student stimulus tapes that were comparable in difficulty to existing EIPA student tapes. The materials from the other children were not appropriate for a variety of reasons: one child was shy on camera, and his materials lacked language samples of adequate length and complexity; two children had very intelligible speech, such that no materials of adequate difficulty were available for the AV Cue-to-Voice portion of the test; and one child’s age (13 years) meant that her language samples were neither appropriate for the elementary nor secondary level. For each of the two children selected, student test materials were developed that contained both an AV and a VO testing segment. Each testing segment consisted of a complete set of EIPA test materials, comparable to that found on signing versions of the EIPA. As a result of containing two separate (AV and VO) sets of testing materials, the CS student stimulus tapes were somewhat longer in duration than their signing counterparts: the test tapes were approximately 40 min (signing versions: ~25 min) in length.

In order to create the testing segments for the two selected children, each child’s portion of the interview was first transcribed in its entirety. From the transcriptions, roughly 20 min of AV and 20 min of VO materials were identified for use on the tapes. In keeping with the procedures used to develop the student materials for signing versions of the EIPA, these materials were selected so that a fictional interview could be constructed that progressed from shorter and more concrete language samples to longer and more abstract samples. Thus, the chronological order of the original
interview was not necessarily maintained in the fictional interview. As such, the facilitator’s questions from the original interview were not always appropriate for the fictional interview. Therefore, fictional questions from an “interviewer” were scripted in order to facilitate transitions between selected language samples. In addition, minor modifications to the EIPA instructions (written and spoken) were required: references to “signing” were replaced with “cueing,” and instructions were added to facilitate the transition between AV and VO segments.

In order to ensure that the look and feel of the EIPA student tapes was maintained, BTNRH provided the EIPA backdrop for filming and recorded the necessary audio materials (i.e., the modified spoken instructions and the fictional interview questions) to digital audio files, using an actor whose voice had also been used on other EIPA materials. BTNRH then provided all audio files as well as the EIPA music and EIPA video graphics to the Media Innovation Team at the University of South Florida, who performed the final video editing. Two complete sets of tapes were produced: (a) elementary: warm-up and elementary: test and (b) secondary: warm-up and secondary: test. Each warm-up tape contained a short amount of material that was not included on the final test tapes. In an EIPA testing situation, such tapes are typically used by interpreters to select between the two student options available for a given grade level and to become familiar with the signing style of the student selected. Although only one set of student test materials was developed at each grade level for the CS pilot test, the warm-up tapes were still needed in order to allow CS interpreters the same opportunity to become familiar with the student’s cueing style and voice during the EIPA warm-up period (just prior to testing). The development of warm-up tapes for the pilot test also allows additional student tapes to be introduced seamlessly into the EIPA-CS test whenever resources become available to develop them. For consistency with test tapes, the CS warm-up tapes included an AV warm-up segment and a VO warm-up segment. As a result, the student warm-up tapes were again somewhat longer in duration than their signing counterparts (cueing versions: ~8 min; signing versions: ~5 min).

Scoring sheets. One purpose of the EIPA score sheet is to itemize all skills that are assessed (on a scale of 0 to 5) in an EIPA evaluation. For the signing version of the EIPA, 37 skills are listed on the score sheet, organized under four general domains: Voice-to-Sign (i.e., expressive product), Sign-to-Voice (i.e., receptive product), Vocabulary, and Overall factors. For the EIPA-CS pilot test, modifications to the existing EIPA score sheet were required because CS-based interpreting involves a somewhat different skill set than sign-based interpreting. Thus, any skills that were not applicable to CS interpreting (e.g., “location/relationship using ASL classifier system” in the Voice-to-Sign domain), including the entire domain of skills pertaining to Vocabulary, were eliminated. Other skills, however, were indeed applicable to CS interpreting and were not changed (particularly skills related to conveying prosody in the expressive and receptive domains, e.g., “stress/emphasis for important words or phrases” as well as some skills in the Overall factors domain, e.g., “demonstrates process decalage appropriately”). Finally, a number of CS-related skills (e.g., “appropriate use of alternate cueing hands”) were introduced to the score sheet, and the name for the third domain was changed to “Intelligibility.” In addition, Cue-to-Voice was subdivided into two domains so that both AV Cue-to-Voice and VO Cue-to-Voice performances could be evaluated. After the initial draft of the EIPA-CS score sheet was constructed in this manner, additional input was sought from three expert consultants (two transliterators, one consumer; all were certified Instructors of CS, and all were either pursuing or had completed Masters degrees) to ensure that the final skill list was comprehensive, addressing all basic CS-interpreting competencies necessary to convey classroom discourse (but not skills in specialty areas such as music, foreign languages, and regional dialects). A summary of the final version of the EIPA-CS score sheet is shown in Table 2.

The procedures used for tabulating scores from the EIPA-CS score sheet were analogous to those used for the signing versions of the EIPA: scores for a given domain were obtained by averaging the scores for skills within that domain, and overall performance scores were obtained by averaging the four domain
scores. One small difference, however, was in the calculation of the Cue-to-Voice domain score. Because the EIPA-CS Cue-to-Voice domain consists of two domain scores (AV and VO), a composite Cue-to-Voice score was obtained by calculating a weighted average of the scores for the two Cue-to-Voice domains: the larger of the two scores (AV or VO) was weighted by a factor of 0.75, and the smaller of the two scores was weighted by a factor of 0.25. Thus, an interpreter who received an AV score of 5 and a VO score of 1 would receive a Cue-to-Voice score of 4.0 (5 × 0.75 + 1 × 0.25). This composite Cue-to-Voice score was not displayed on the score sheet, which reported only the actual AV and VO domain scores. It was used simply to reduce slightly the effect of a poor score in one of the two Cue-to-Voice domains on the overall performance score. Such an adjustment is warranted because CS-interpreting jobs typically require either AV Cue-to-Voice skills or VO Cue-to-Voice skills but not both (more commonly, only AV skills would be required), and many interpreters can be placed successfully with minimal skills in one of the two areas (more commonly, with minimal skills in the VO Cue-to-Voice area).

**Pilot test administration.** After the video materials and score sheet were modified, the EIPA-CS pilot test was used to conduct 25 tests (14 elementary and 11 secondary). The testing procedure was largely the same as that used for the signing version of the EIPA. First, each interpreter provided some demographic information and indicated whether she/he planned to take the test at either the elementary level or the secondary level. The interpreter was then asked to select a classroom option after reviewing brief written descriptions of the two classroom options available at that level. Of the 14 CS interpreters tested at the elementary level, 7 selected Option A and 7 selected Option B; of the 11 interpreters tested at the secondary level, 7 selected Option A and 4 selected Option B. Unlike the signing version of the EIPA, the interpreter was not given a choice regarding student materials because only
one option was available at each level. Once the materials were selected, the interpreter was given a warm-up period in which to review (a) a detailed written description of the classroom option selected and (b) the Cue-to-Voice warm-up tape. The warm-up period was followed by two 40-min test periods, each separated by a short break. The interpreter determined which testing period would be used for assessing Cue-to-Voice performance and which would be used for assessing Voice-to-Cue performance.

Development of Evaluation Procedures

For the EIPA-CS pilot test, evaluation procedures that were specific to CS were developed with input from the same three consultants who provided feedback on the score sheet. It was necessary to develop evaluation procedures for each CS-based interpreting skill that was introduced to the EIPA score sheet. In order to facilitate the development of such procedures, the consultants and facilitator viewed the 16 test tapes from one of the participating states, looking for common features associated with interpreters at specific skill levels in each of the areas to be evaluated. The observations and discussions generated from these tapes led to precise written descriptions of interpreter behavior (i.e., rubrics, measurable and quantifiable whenever possible) that were associated with each of six scores available to evaluators (0, 1, 2, 3, 4, and 5) for each skill to be evaluated (Appendix A describes what each of these scores corresponds to in terms of overall skill level; for reference, standard EIPA skill descriptions are available at http://www.classroominterpreting.org or see Schick et al., 2006).

After specifying the details of the evaluation procedures in this manner, the consultants then served as the first evaluation team (Team 1) and met to begin rating tapes. The meeting followed procedures established for signing versions of the EIPA: team members first evaluated the interpreter individually and then discussed the interpreter’s performance to arrive at the final rating. During this period, Team 1 rated 12 tapes (6 elementary and 6 secondary). As they gained experience using the rubrics that had been developed, modifications were made to the evaluation procedure. When a modification was made to an evaluation procedure for a particular skill, all tapes that had been scored previously were rerated in that skill area. Given this iterative procedure, each tape took roughly 3 h to rate on average.

Rater Training and Reliability Assessment Procedures

In order to collect intrarater (i.e., intrateam) reliability data, Team 1 met again roughly 3 weeks later and rated 14 test tapes (7 elementary and 7 secondary). Six of the tapes had been rated previously (3 elementary and 3 secondary) by Team 1 in the earlier meeting. The remaining eight tapes were new. Thus, Team 1 graded 20 unique tapes (12 tapes in the first meeting and 8 new tapes in the second meeting). During the second meeting, the team was noticeably faster and more confident in their ratings. On average, each tape took roughly 2 h to rate. The team members themselves reported feeling more confident that their ratings were consistent with the evaluation rubrics due to increased familiarity with the evaluation procedures. They also commented that learning how best to apply the rubrics had continued throughout most of the initial session. Consequently, it was decided that additional intrarater reliability data were needed, in order to minimize the effects of learning. Therefore, Team 1 met a third time roughly 3 weeks later and rated seven test tapes (4 elementary and 3 secondary), all of which had been previously rated at the second meeting. During this final meeting, each tape took roughly 1½ h to rate. In all, Team 1 assigned 33 scores (12 in Session 1, 14 in Session 2, and 7 in Session 3). As shown in Table 3, the scores corresponded to 20 unique tapes and 13 repeated measures (six between Sessions 1 and 2 and seven between Sessions 2 and 3) that were available for reliability analysis.

Based on the experiences with Team 1, a workshop was designed to train a second team of evaluators (Team 2). The 1½-day evaluation training workshop consisted of a lecture intermixed with opportunities for the team members to practice evaluating sample materials. Members of Team 1 assisted the facilitator with the development of the lecture materials that were used in the workshop. In addition to a copy of the lecture materials, Team 2 also received a rater’s manual specifying the details of the evaluation procedures, which was developed to serve as a reference manual both during and after the workshop. Two members of Team 1 also attended the workshop and
were available to answer any questions that arose, as well as to provide personal insights on the evaluation process. The format of the workshop was designed so that it can be repeated in the future, in the event that the CS pilot test is adopted by the EIPA Diagnostic Center and additional teams of evaluators are needed.

Table 3  Number of tests rated at each team meeting

| Team/meeting     | Number of tests scored |          |
|------------------|------------------------|----------|
|                  | Total | New  | Rerated |
| Team 1/first     |       |      |         |
| meeting Elementary | 6    | 6    | —       |
| Secondary        | 6    | 6    | —       |
| Total            | 12   | 12   | —       |
| Team 1/second    |       |      |         |
| meeting Elementary | 7    | 4    | 3       |
| Secondary        | 7    | 4    | 3       |
| Total            | 14   | 8    | 6       |
| Team 1/third     |       |      |         |
| meeting Elementary | 4    | 0    | 4       |
| Secondary        | 3    | 0    | 3       |
| Total            | 7    | 0    | 7       |
| Team 2           |       |      |         |
| Elementary       | 7    | 4    | 3       |
| Secondary        | 5    | 1    | 4       |
| Total            | 12   | 5    | 7       |
| Total (all team  |       |      |         |
| meetings)        |       |      |         |
| Elementary       | 24   | 14   | 10      |
| Secondary        | 21   | 11   | 10      |
| Total            | 45   | 25   | 20      

In order to collect interrater (i.e., interteam) reliability data, Team 2 rated 12 test tapes (7 elementary and 5 secondary), seven of which had been rated previously (3 elementary and 4 secondary) by Team 1. On average, each tape took roughly 3 h to rate. Table 3 summarizes the number of tests that were rated by each of the teams and the number that were rerated in order to assess intrarater and interrater reliability (evaluators’ tests were never rerated). For the 25 tests, the two teams together assigned 45 scores (33 scores were assigned by Team 1 and 12 scores were assigned by Team 2).

Results

The EIPA-CS ratings for the 25 tests that were administered are shown in Figure 1. For tests that were scored more than once, error bars indicate the range of actual scores awarded (e.g., by Team 1 and Team 2 or by Team 1 at two different times). The 25 pilot test scores thus ranged from 2.5 to 4.5, although the 45 actual scores assigned by the two teams ranged from 2.1 to 4.6. Of all 45 scores that were assigned to tests, the mean score was 3.45 ($SD = 0.58$). This distribution of scores suggests that the EIPA-CS evaluation
was able to differentiate a wide range of interpreter skill levels.

The results of the initial evaluations of these 25 tests served a number of purposes. First, validity of the pilot test was assessed by examining the equivalency of test forms and the construct validity of the EIPA-CS instrument. Second, both intrarater and interrater reliability were measured. Third, skills with poorly specified evaluation procedures were identified by an item analysis that identified wide variations in evaluator scores. Finally, the performance skills of educational interpreters who use CS were analyzed, the first such data of its kind for this group of professionals.

Validity

Equivalency of test forms was evaluated by examining the breakdown of scores across grade level and classroom option. Table 4 shows close agreement in mean scores and in $SD$s both for the elementary and secondary levels and for options A and B at each level, suggesting that the test is equivalent in difficulty across level and classroom option for CS, just as it is for the sign language options.

Table 4  Average test scores across grade levels and classroom options

| Grade level | Total | Option A | Option B |
|-------------|-------|----------|----------|
| Elementary  |       |          |          |
| $N$         | 14    | 7        | 7        |
| $M$         | 3.33  | 3.44     | 3.23     |
| $SD$        | 0.46  | 0.54     | 0.36     |
| Secondary   |       |          |          |
| $N$         | 11    | 7        | 4        |
| $M$         | 3.58  | 3.47     | 3.72     |
| $SD$        | 0.68  | 0.71     | 0.65     |

Construct validity of the EIPA-CS test was evaluated by examining two subtypes of construct validity: discriminant and convergent validity. Both types of validity reflect the degree to which test items are organized appropriately into the five major content domains (i.e., the test “constructs”) listed on the EIPA-CS score sheet: Voice-to-Cue, Cue-to-Voice (AV), Cue-to-Voice (VO), Intelligibility, and Overall factors. Discriminant validity measures the degree to which each domain represents a different theoretical construct, whereas convergent validity measures the degree to which test items within a domain are related to the same theoretical construct. To assess discriminant validity, the domain scores for the 25 tests were used to examine correlations between domains. As shown in Table 5, the interdomain correlations ranged from no correlation to moderate correlations, indicating that the domains are evaluating largely different skill sets and that each domain contributes unique variance to the overall EIPA score.

To assess convergent validity, measures of internal consistency for each domain were calculated. Internal consistency is an estimate of the proportion of the variability in scores that is the result of differences in the skills under evaluation (and not a result of test items that do not pertain to the theoretical construct, or unclear evaluation procedures, etc.). A high coefficient of internal consistency indicates that the individual test items are homogeneous; that is, each item is related to the same theoretical construct and contributes in a consistent way to the overall score for the domain. Table 5 shows Cronbach alpha estimates of internal consistency for each domain. A Cronbach

Table 5  Interdomain correlations and internal consistency (Cronbach alpha) for each domain

|                  | Voice-to-Cue | Cue-to-Voice (AV) | Cue-to-Voice (VO) | Intelligibility | Overall factors |
|------------------|--------------|-------------------|-------------------|----------------|-----------------|
| Voice-to-Cue     | 1.00         | 0.11              | 0.34              | 0.56           | 0.37            |
| Cue-to-Voice (AV)| 1.00         | 0.58              | -0.01             | 0.56           |                 |
| Cue-to-Voice (VO)| 1.00         | 0.18              | 0.72              |                |                 |
| Intelligibility  |              | 1.00              |                  | 0.47           |                 |
| Overall factors  | 1.00         |                   |                  |                | 1.00            |
| Internal consistency | 0.78      | 0.92              | 0.93              | 0.81           | 0.74            |
alpha coefficient (essentially the average of all split-half correlations) above 0.70 is considered acceptable, and a value of 0.90 is considered to be very good (Schick et al., 2006). Acceptable levels of internal consistency were obtained for all domains, with very good levels obtained for Cue-to-Voice (AV) and Cue-to-Voice (VO). These results demonstrate that items within a domain are homogeneous and contribute to measuring the same skill set or construct. Taken together with the data on discriminant validity, these data lend support for the construct validity of the EIPA-CS instrument.

Reliability

To assess the reliability of overall performance scores, the pairs of scores obtained for tests that were scored more than once were analyzed. These scores, summarized in Table 6, allowed for assessment of interrater reliability (scoring differences between Team 1 and Team 2) as well as two types of intrarater reliability: “early” (scoring differences between the first and second meetings of Team 1) and “late” (scoring differences between the second and third meetings of Team 1). Test–retest reliability was assessed first by examining the hypothesis that the two sets of scores obtained for a given set of tests had equal means and variances (strictly parallel model). Chi-square tests for the goodness of fit found no statistically significant ($p < .05$) evidence to suggest that tests were scored differently by the two different teams ($p = .80$) or by Team 1 over time, either early ($p = .78$) or late ($p = .48$). Then, intrarater and interrater reliability were assessed by determining the interclass correlation coefficient (ICC), a measure of the agreement between pairs of scores. This measure showed high intrarater reliability, both early (ICC = 0.93, $p = .01$) and late (ICC = 0.86, $p = .02$). Interrater reliability was also fairly high (ICC = 0.76), with a $p$ value of .06 which suggests that the data show a strong trend toward this level of reliability, although statistical significance was not reached. Thus, more interrater reliability data would be required in order to verify whether 0.76 is an accurate reflection of the interrater reliability.

Reliability for each of the five domain scores (Voice-to-Cue, Cue-to-Voice [AV], Cue-to-Voice [VO], Intelligibility, and Overall factors), summarized in Table 7, was assessed in the same manner. With one exception, chi-square tests found no statistically significant ($p < .05$) evidence to suggest that tests were scored differently in any of these five domains, either by the two different evaluation teams or by Team 1 over time (early or late). The only exception was Cue-to-Voice (AV) scores, which did differ significantly between teams (but not within Team 1 over time). Regarding the degree of agreement in scoring, Intelligibility and Overall factors showed high interrater reliability (ICCs of 0.84 and 0.87, respectively) as well as high intrarater reliability, both early (ICCs of 0.86 and 0.94) and late (ICCs of 0.74 and 0.90), whereas Cue-to-Voice (AV) and Cue-to-Voice (VO) showed high intrarater reliability only. Given that high intrarater reliability was obtained for Cue-to-Voice (AV) and Cue-to-Voice (VO), it is expected that interrater reliability can be improved in these domains simply by providing Team 2 with further training on the evaluation procedures for these areas. To improve reliability for Voice-to-Cue domain scores, both teams may need additional

### Table 6: Pairs of overall performance scores (across teams) for each test scored more than once

| Test | Team 1 | Team 2 | Test | Team 1 (first) | Team 1 (second) | Test | Team 1 (second) | Team 1 (third) |
|------|--------|--------|------|---------------|----------------|------|----------------|----------------|
| INT1 | 3.7    | 3.3    | INT1 | 3.0           | 3.7            | INT3 | 4.1            | 3.3            |
| INT3 | 3.3    | 3.7    | INT9 | 2.8           | 3.7            | INT4 | 2.6            | 2.8            |
| INT11| 3.3    | 2.9    | INT11| 3.3           | 3.3            | INT13| 3.7            | 3.6            |
| INT18| 2.6    | 2.8    | INT18| 2.1           | 2.6            | INT14| 3.2            | 3.4            |
| INT20| 4.0    | 4.0    | INT22| 3.4           | 3.0            | INT17| 3.6            | 3.3            |
| INT22| 3.0    | 3.6    | INT24| 4.6           | 4.5            | INT19| 3.4            | 3.5            |
| INT23| 2.9    | 3.2    | INT25| 4.2           | 4.1            |

$M$ (SD) 3.25 (0.49) 3.36 (0.43) $M$ (SD) 3.19 (0.84) 3.29 (0.74) $M$ (SD) 3.54 (0.52) 3.43 (0.40)
training or it may be possible to refine evaluation procedures for certain items that were difficult to rate. In order to lend insight into whether individual items in Voice-to-Cue (or other domains) were difficult to rate, an item analysis was conducted.

Item Analysis

Individual items listed on the score sheet (see Table 2) were analyzed by comparing the pairs of item scores obtained for tests that were scored more than once. Items that were difficult to rate were identified by determining the maximum difference among 14 difference scores. The 14 difference scores were derived from 14 of the pairs of tests collected for the reliability analysis: specifically, the seven pairs of tests used to measure interrater reliability and the seven pairs of tests used to measure intrarater reliability. The six pairs of tests used to measure intrarater early reliability were excluded from this item analysis in order to minimize possible learning effects experienced by Team 1, who had reported an impression that learning occurred throughout the initial evaluation meeting and may have resulted in inconsistent ratings. Although learning effects were not apparent in the reliability of overall performance or domain scores, this impression could have stemmed from individual test items that they were still learning to rate. Therefore, these data were excluded. From the remaining 14 pairs of tests, a maximum difference score was determined for each item on the score sheet. The maximum difference score was derived by calculating the difference between each pair of scores and determining the maximum difference among the 14 difference scores. A difficult to rate item was defined as an item that had a maximum difference score of 2.5 or greater, which meant that in at least one instance, the two scores for that item (given either by Team 1 and Team 2 or by Team 1 at two different times) differed by 2.5 or half of the entire scale. Difficult to rate items that were identified and their corresponding difference scores are shown in Table 8.

Of the 44 items on the score sheet, 12 were identified as difficult to rate. In some of these cases, difficulty in rating items may simply indicate that teams needed more training on the evaluation procedures for these items in order to prevent such discrepancies. For example, five of the items identified as difficult to rate stemmed from interrater reliability differences only; these cases may reflect the need for additional training of Team 2 on these items. Team 2 had less experience overall with the evaluation procedures than Team 1, who not only had more experience evaluating tests over the course of three meetings but also helped design the score sheet and corresponding rubrics. Similarly, two maximum differences stemmed from intrarater reliability only, suggesting that Team 1 may need additional training on these items. In the remaining five cases (Cue-to-Voice [VO]: Items D, E, F, G, and H), however, both interrater and intrarater data maximum differences scores were 4.0 or higher, suggesting that these items were difficult to rate in at least one instance for each of the two teams. Notably, all five of these items were VO Cue-to-Voice skills. Additional training of both teams and modification of the evaluation procedures for these skills, to clarify any sources of confusion in the rubric, should also be considered in future versions of the test.

Performance Overview of CS Interpreters

With validity and reliability established, the performance skills of the CS interpreters who participated

Table 7  Summary of the reliability data collected for area scores

| Domain           | Interrater | Intrarater—early | Intrarater—late |
|------------------|------------|------------------|-----------------|
|                  | Fit        | ICC, $p$         | Fit             | ICC, $p$         | Fit             | ICC, $p$         |
| Voice-to-Cue     | 0.86       | $-0.19, p = .57$ | 0.43            | 0.52, $p = .21$  | 0.42            | $0.68, p = .10$  |
| Cue-to-Voice (AV)| 0.00*      | N/A              | 0.38            | 0.73, $p = .08$  | 0.46            | $0.96, p = .00$  |
| Cue-to-Voice (VO)| 0.74       | 0.17, $p = .41$  | 0.63            | $0.82, p = .04$  | 0.20            | $0.79, p = .01$  |
| Intelligibility  | 0.87       | 0.84*, $p = .03$ | 0.67            | $0.86, p = .03$  | 0.55            | $0.74, p = .05$  |
| Overall factors  | 0.73       | 0.87*, $p = .01$ | 0.94            | $0.94, p = .01$  | 0.13            | $0.90, p = .01$  |

* $p < .05$.

Note. N/A, not applicable. Bold font indicates ICCs that were statistically significant or approached significance ($p < .10$).
in pilot testing were examined. Overall results are shown in Figure 1; the dotted line represents an intermediate skill level of 3.5. This level is the minimal acceptable skill level established by many states for educational interpreters who use a sign language or system. Slightly more than half (56%) of the 25 interpreters tested had overall scores above this level. However, the data set includes four tests that were administered to individuals who were recruited as evaluators and opted to take the EIPA-CS test prior to participating in evaluator training. If these individuals (who were recruited as experts in the field and thus may not be representative of typical CS interpreters) are excluded, only 48% of the 21 interpreters that formally participated in pilot testing had overall scores above the minimal acceptable level.

In order to examine the skills of these 21 interpreters in more detail, Table 9 lists the means, medians, and SDs for this group in the five different domains of skill: Voice-to-Cue, AV Cue-to-Voice, VO Cue-to-Voice, Intelligibility, and Overall factors. As shown in Table 9, the interpreters often scored higher on intelligibility than on other factors that contribute to an accurately interpreted message. In addition, skills of the average interpreter did not meet minimum acceptable levels in either Cue-to-Voice domain, even in the AV Cue-to-Voice domain. In the VO Cue-to-Voice domain, the average interpreter’s skills were at a beginner/advanced beginner level, confirming that there has been little expectation historically for CS interpreters to develop skills in this area. Finally, SDs for the domains were large (ranging from 0.48 to 1.25), indicating that interpreters’ skills vary considerably within each domain.

**Discussion**

EIPA-CS: Status and Outlook

In this study, a pilot version of the EIPA for CS interpreters was developed in order to assess the feasibility of extending the EIPA to include CS. Results from this phase of initial development are promising. The construct validity of the instrument was quite high: interdomain correlations demonstrate that each domain measures a largely different skill set, and Cronbach alpha estimates suggest that items within
a domain exhibit acceptable levels of internal consistency. Test–retest reliability was also good: 15 tests were evaluated two to three times each, and no statistically significant difference between multiple evaluations was detected either between the two different teams or by Team 1 over time. Intrarater reliability was high both for overall performance scores (0.90, on average) and for four of five individual domain scores (0.85, on average), which suggests that the EIPA–CS test can be rated as reliably as signing versions of the EIPA (signing versions: 0.91 interrater reliability, on average; Schick et al., 2006). Interrater reliability was similarly high (0.86, on average) for two of five domains (Intelligibility and Overall factors). Although interrater reliability of overall performance scores was fairly high (0.76), it did not reach statistical significance.

One likely reason that intrarater reliability was higher than interrater reliability is that Team 2 had less experience overall with the evaluation procedures than Team 1, who not only met three times to evaluate tests but also helped to design the score sheet and corresponding rubrics. Because Team 1 had the opportunity to view many of the test tapes multiple times and to discuss the evaluation of many tapes more than once, it is not unexpected that reliability for this team would be high. Indeed, the reliability of Team 1 may represent an upper limit on the level of reliability that can be expected on the EIPA–CS, at least until reliability of difficult to rate items can be improved. Given that Team 2 was not involved in the design of the evaluation procedures, it seems reasonable that they would need somewhat more time to learn the evaluation procedures. Therefore, providing Team 2 with additional training before EIPA–CS testing resumes should improve interrater reliability. The format of the training workshop need not be changed, but the time frame should be extended to allow for more practice scoring of sample tapes. In addition, extended practice should be provided for those items that were identified as difficult to rate in the item analysis. Depending on the anticipated market for a CS version of the EIPA, additional evaluators could also be trained at such a meeting, which would provide a larger pool of evaluators.

If the CS version of the EIPA is officially adopted and offered through the EIPA Diagnostic center (negotiations with BTNRH are currently underway), work should continue in four areas. First, a second option for student tapes should be developed at each level, in order to make the testing process for CS interpreters fully parallel to the testing process for sign interpreters. Second, difficult to rate items should be further explored. Although these items do not appear to be affecting the overall reliability of the test, they do have a negative bearing on the internal consistency of the instrument. They could also be affecting the evaluation process, by causing teams to rate more slowly and/or experience frustration on these items. It is hoped that additional training and modification of rubrics for these items would improve scoring consistency. Third, because members of each EIPA–CS evaluation team are not typically located in one geographical area, it would be helpful to explore technological solutions that would allow team members to “meet” virtually (e.g., online, videoconferencing, etc.) to rate tests rather than traveling long distances to meet in the same physical location. Such technology would make evaluating tests more economically viable because evaluator travel expenses would not be incurred.

Lastly, and most importantly, additional validity and reliability data should be collected. In particular, more interrater data should be obtained in order to determine whether the estimate of 0.76 reflects the true interrater reliability of this instrument. Similarly, more intrarater data would be informative as well, because many of the confidence intervals around the ICCs were relatively large due to the small number of pairs of scores. These additional reliability data could be used not only to identify any remaining areas of need (i.e., individual items that remain difficult to rate) but also to establish criteria for acceptable variability of scoring between evaluators. Such criteria could be applied to future groups of potential evaluators, who could be asked to score sample tests and required to meet established reliability criteria before becoming eligible to operate in official capacity as an EIPA–CS evaluation team member. Potential evaluators who do not meet the reliability criteria could thus be easily identified and asked to undergo additional training (through peer, i.e., fellow evaluator, mentoring). Finally, to collect additional validity data,
CS interpreters could be evaluated with both the EIPA-CS and the TECUnit’s CLTNCE, in order to explore whether interpreters who qualify for national certification achieve higher scores on the EIPA-CS than interpreters who do not qualify.

Skill levels of CS Interpreters in Educational Settings

Although continued work (described above) could improve the EIPA-CS, the data from this study suggest that the pilot test is already in a position to provide reasonable information regarding skill levels of educational interpreters who use CS. The majority of CS interpreters who participated formally in pilot testing (52%) scored below an EIPA score of 3.5, the minimum proficiency level typically used for research applications of signing versions of the EIPA (e.g., Schick et al., 1999) and the level used by many states in establishing minimum performance standards for educational interpreters who sign. Although research is still needed to determine whether 3.5 is the actual minimum proficiency level needed to ensure access, it is virtually certain that a lower proficiency level would not be adequate (given the frequency and types of errors made by interpreters who receive a rating of less than 3.5). Thus, the picture that emerges for deaf and hard-of-hearing students who use CS is similar to that reported previously for their signing counterparts (e.g., Schick et al., 2006): many receive an interpretation that distorts or inadequately represents classroom communication.

Inspection of domain scores for pilot testing reveals that, in general, the intelligibility skills of CS interpreters were higher than other skills evaluated at a discourse level. Specifically, scores for the Intelligibility domain were substantially higher than scores for Voice-to-Cue product, a domain that contains numerous ratings for preserving prosodic and other types of information supportive to the message, and for Overall factors, a domain which contains ratings regarding the overall quality of the message. Although this finding could suggest that some CS interpreters may be focused in rapid discourse situations on “getting the cues out” at the expense of prosodic and other supporting information, another possibility is that some CS interpreters may not even be aware that prosodic and other supporting information should be preserved in the transliterated message. Given the paucity of training and professional development opportunities available to CS interpreters, the latter explanation is not unlikely. Regardless of the reason, the data suggest that a high degree of CS proficiency does not necessarily correspond to the ability to convey discourse effectively. This issue has profound implication for educational interpreters, given that child language theory has long speculated that the prosody and rhythm of a language provide a child with vital clues about the meaning of an utterance (Fernald, 1992; Jusczyk, 1997; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Wright Cassidy, 1989).

Further inspection of domain scores also reveals that the skills of the average CS interpreter who participated in pilot testing did not meet minimum acceptable levels in either of the two Cue-to-Voice domains. One contributing factor to poor scores in these domains may be that a number of interpreters opted to cue expressively while voicing for the deaf student. Although this practice is typically encouraged in the cueing community (nominally, to allow deaf consumers to monitor the accuracy of the interpreter's voicing) and is required on the national certification exam, it was not required on the EIPA-CS. Although interpreters were not explicitly downgraded for cueing expressively while voicing, their skills were judged on the quality of the spoken English output alone, reflecting our belief that the natural prosody of spoken English must be the priority for situations in which the primary consumers of the interpreter's output are hearing individuals (i.e., Cue-to-Voice situations). When expressive cueing accompanies speech, the speech frequently has a different rhythm and, in some cases, a different intonation, than typical spoken English; thus, the interpreter's Cue-to-Voice score may have been negatively affected to the extent that expressive cueing disrupted the natural prosody of spoken English. In addition, it is possible that the increased cognitive load of expressively cueing while performing a difficult task may reduce overall performance in the Cue-to-Voice domains. This possibility should be explored in future research, in order to assess the advantages and disadvantages of this practice. Finally, it should also be pointed out that the average interpreter's skills were at a beginner/advanced beginner
level in the VO Cue-to-Voice domain. Although VO Cue-to-Voice skills are not commonly expected of educational interpreters who use CS, some situations do require these skills. Formal training in this area is not widely available, and informal training through immersion in a cueing community is only possible in the rarest of circumstances, given the small number of people who use CS. If this type of skill is to be required in some educational settings, it is urgent that training in this area be offered to CS interpreters and that training materials be developed and made available for purchase.

Despite these shortcomings in specific skills, it is worth noting that the overall skills of CS interpreters in educational settings (based on this data set) do not appear to be substantially different from those of their signing counterparts. This fact is encouraging, given that the profession of cued language transliteration is only about 30 years old and that CS is not yet offered in ITPs. Another encouraging fact is that 14% of interpreters who formally participated in pilot testing scored 4.0 or higher. Clearly, there are excellent CS interpreters who work in K-12 settings. Finally, with the availability of the EIPA-CS instrument developed in this study, a research tool is now available for assessing CS interpreters nationwide. As information is gathered on the skills of these individuals, the extent and type of training needed for CS interpreters can be formally documented. Ultimately, it is hoped that such information can improve both the quality and the availability of training for these individuals. Even though the number of educational interpreters using CS may be relatively small, these research efforts are necessary in order to ensure that no child is left behind.

Funding

The U.S. Department of Education Office of Special Education Programs (Programs of National Significance grant H325 N010013 to B.S); the National Cued Speech Association (Transliterator Testing grant to J.K.).

Appendix A

Profile of Skills at Each Rating Level

**Level 1: Beginner.** Demonstrates very limited intelligibility with frequent errors in production. At times, cue production may be incomprehensible and lacks prosody and other supporting information. Individual is only able to interpret very simple Voice-to-Cue communication. Individual has difficulty conveying, comprehending, and interpreting cued messages; single words may be comprehended/interpreted, but effective communication is lost. An individual at this level is not appropriate for classroom interpreting.

**Level 2: Advanced beginner.** Demonstrates only basic intelligibility. Limitations in cueing speed and intelligibility interfere with successful communication. More fluent than a Beginner, but lack of fluency still greatly interferes with communication. In cue production, frequent errors and/or unclear cues and mouth movements are apparent. Some use of prosody and supporting information, but use is inconsistent and often inappropriate. Individual is able to read cues at the word level, but complete sentences often require repetitions and repairs. Both Voice-to-Cue and Cue-to-Voice interpretations demonstrate serious deficiencies in the message conveyed. Without considerable supervision and mentoring, an individual at this level is not recommended for classroom interpreting.

**Level 3: Intermediate.** Demonstrates moderate intelligibility, yet cueing speed and clarity would most likely be insufficient for complex interpreting situations. Some aspects of cue production may be incorrect even though it may not interfere with communication. Production of prosodic and other supporting information is emerging but may still be incorrect. Technical topics will most likely pose a great problem. May comprehend a cued message but may need repetition and assistance at times. Both Voice-to-Cue and Cue-to-Voice interpretations generally contain all the key points, but parts of the message may be missing. An individual at this level would be able to interpret basic classroom content but would demonstrate great difficulty conveying all information in the message and may have difficulty with interpreting rapid or technical information. An individual at this level needs supervision and additional training.

**Level 4: Advanced intermediate.** Demonstrates high intelligibility at most speaking rates with cue
production generally correct. Individual demonstrates paraphrasing strategies for conveying information when a speaker’s rate exceeds his/her maximum cueing rate. Cued messages are generally clear and consistent, but complex interpreting situations may still pose problems. Prosody is acceptable. Consistently includes other supporting information. Fluency may deteriorate when rate or complexity of input increases. Comprehension of most cued messages at a normal rate is good and Cue-to-Voice message conveys all keys points. An individual at this level would be able to interpret most classroom content but may still have difficulty clearly or accurately conveying information in some complex situations involving rapid speech, technical vocabulary, and multiple speakers.

Level 5: Advanced. Demonstrates high intelligibility at a wide variety of speaking rates, with paraphrasing strategies for communicating extremely high rates of speech. Prosody is skillfully conveyed. Individual correctly uses space and other techniques to incorporate fully all supporting information. Complex interpreting situations do not pose a problem. Comprehension of cued messages is very good. An individual at this level is capable of clearly and accurately conveying the vast majority of classroom interactions.

Notes

1. In general, the function of an interpreter is to translate between two languages (e.g., spoken English and ASL), whereas the function of a transliterator is to transfer information between two modes of the same language (e.g., spoken English or either signed English or cued English). Consequently, the term transliterator is often preferred in the cueing community (most typically, CS transliterator or cued language transliterator). However, the role of the CS/cued language transliterator goes beyond direct rendering of English grammatical structures and vocabulary as these individuals must also make decisions regarding how best to paraphrase fast speech, which information to convey when multiple talkers speak simultaneously, etc. Moreover, the term “interpreter” can be used in a generic sense to refer to an individual who works in the profession, as it is used in current EIPA terminology. For these reasons, we use the term “interpreter” throughout, except when referring to specific tests, individuals, or practices within the cueing community.

2. NAD and RID formed the National Council on Interpreting and recently developed the National Interpreter Certification test. With the availability of this new test, the NAD no longer administers their evaluation tool, and the RID will not administer their generalist test beyond December 31, 2008.

State requirements calling for the RID or NAD tests will ultimately need to be updated to reflect these changes.

3. The exclusion of the expressive cueing requirement in the Cue-to-Voice portion of this assessment was purposeful and is a major distinguishing factor between the EIPA-CS and other CS-interpreting assessments. The goal of interpretation is naturalness and message equivalence for the target language consumer. When an interpreter is voicing, the hearing consumer should, ideally, be able to close his/her eyes and receive the message without realizing it was mediated by a transliterator. Expressive cueing while producing the target spoken English interpretation increases the task demand to a point where the naturalness of the spoken English output can be compromised prosodically.

References

Bornstein, H. (Ed.). (1990). Manual communication: Implications for education. Washington, DC: Gallaudet University Press.
Commission of Education of the Deaf. (1988). Toward equality: Education of the deaf. Washington, DC: U.S. Government Printing Office.
Cornett, R. O. (1967). Cued Speech. American Annals of the Deaf, 11(2), 3–13.
Cornett, R. O., & Daisey, M. E. (2001). The Cued Speech resource book for parents of deaf children. Cleveland, OH: National Cued Speech.
Davis, J. (2005). Code choices and consequences: Implications for educational interpreting. In M. Marschark, R. Peterson, & E. Winston (Eds.), Sign language interpreting and interpreter education (pp. 112–141). New York: Oxford University Press.
Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), The adapted mind: Evolutionary psychology and the generation of culture (pp. 391–428). New York: Oxford University Press.
Jones, B. E., Clark, G. M., & Soltz, D. F. (1997). Characteristics and practices of sign language interpreters in inclusive education programs. Exceptional Children, 63(2), 257–268.
Jusczyk, P. W. (1997). The discovery of spoken language. Cambridge, MA: MIT Press.
Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczy, P. Q., & Wright Cassidy, K. (1989). How prosodic cues in motherese might assist language learning. Journal of Child Language, 16, 55–68.
Kantze, M. (1990). ASL: Unity and power. Communication Issues Among Deaf People: The Deaf American Monograph, 40, 75–77.
LaSasso, C. J., & Metzger, M. A. (1998). An alternate route for preparing deaf children for BiBi programs: The home language as L1 and cued speech for conveying traditionally signed languages. Journal of Deaf Studies and Deaf Education, 3, 265–289.
Moores, D. E. (1992). A historical perspective on school placement. In T. N. Kluswin, D. E. Moores, & M. G. Gaustad (Eds.), Toward effective public school programs for deaf
students: Context, process, and outcomes (pp. 7–29). New York: Teachers College, Columbia University.

No Child Left Behind Act. (2001). Pub. L. 107–110, Stat. 1425.
Program and Services Chart. (2003). American Annals of the Deaf, 148, 122–141.

Schick, B. (2004). How might learning through an educational interpreter influence cognitive development? In E. A. Winston (Ed.), Educational interpreting: How it can succeed (pp. 73–87). Washington, DC: Gallaudet University Press.

Schick, B., & Williams, K. (1992). The educational interpreter performance assessment: A tool to evaluate classroom performance. Paper presented at the issues in Language and Deafness: The use of sign language in educational settings: Current concepts and controversies, Omaha, NE.

Schick, B., & Williams, K. (2004). The educational interpreter performance assessment: Current structure and practices. In E. A. Winston (Ed.), Educational interpreting: How it can succeed (pp. 186–205). Washington, DC: Gallaudet University Press.

Schick, B., Williams, K., & Bolster, L. (1999). Skill level of educational interpreters working in public schools. Journal of Deaf Studies and Deaf Education, 4(2), 144–155.

Schick, B., Williams, K., & Kupermintz, H. (2006). Look who’s being left behind: Educational interpreters and access to education for deaf and hard-of-hearing students. Journal of Deaf Studies and Deaf Education, 11(1), 3–20.

U.S. Department of Education. (1999). To assure the free appropriate public education of all children with disabilities: Twenty-first annual report to Congress on the implementing of the Individuals with Disabilities Act. Washington, DC: U.S. Department of Education.

Received June 11, 2007; revisions received October 1, 2007; accepted October 15, 2007.