Advances in Social Sciences Research Journal – Vol. 11, No. 5
Publication Date: May 25, 2024
DOI:10.14738/assrj.115.17016.
Shah, A. P. (2024). Reliability, Sensitivity, and Specificity Measures of the Comprehensive Assessment of Accentedness and Intelligibility (CAAI) Test Battery. Advances in Social Sciences Research Journal, 11(5). 318-334.

Reliability, Sensitivity, and Specificity Measures of the Comprehensive Assessment of Accentedness and Intelligibility (CAAI) Test Battery

Amee P. Shah
School of Health Sciences, Stockton University, Galloway, New Jersey

ABSTRACT

Purpose: This paper aims to introduce and report on measures of validation of a diagnostic assessment, the Comprehensive Assessment of Accentedness and Intelligibility (CAAI) Test Battery that identifies communication issues related to a person's spoken accent. Due to scant availability of standardized assessments of accents, this test battery was developed to enable evidence-based diagnoses and determine severity levels of dialect- and accent-related communication difficulties (Shah, 2024). The present paper reports on psychometric testing conducted on the test battery with the resulting measures of reliability, sensitivity, and specificity of the CAAI. Methods: Native and nonnative speakers of English (N = 61) were administered the CAAI Test Battery, and their communication was evaluated and scored. Descriptive statistics were used to examine test sensitivity, specificity, interrater reliability, inter-item (internal consistency) reliability, and test-retest reliability. Results: The CAAI Test Battery meets high standards for *test sensitivity* in identifying accent-related communication concerns. *Test specificity* was high for not falsely identifying native speakers as having accent-related concerns. The instrument had strong i*nterrater reliability* across all 20 sections of the test (Pearson correlation coefficients ranged from 0.68 to 1.00, all highly significant at $p<0.01$). The test showed strong *inter-item reliability/internal consistency* with a large number of sections yielding moderate to high Cronbach's alpha coefficient range of 0.70 to 1.00. The test showed strong *test-retest reliability* with correlation coefficients ranging from 0.73 to 1.00, significant at $p<0.05$. Conclusions: With high sensitivity, specificity, and reliability, the CAAI Test Battery is found to be a stable and meaningful means to assess dialect- and accent-related communication concerns. The CAAI Test Battery helps fill a crucial gap in the area of accents/dialects that lacks other validated assessment measures. With a sensitive and reliable assessment, clinicians and teachers can identify an accurate baseline pattern of errors to address for accent management or pronunciation teaching and achieve effective outcomes in a short amount of time.

INTRODUCTION: STATE OF THE PRACTICES

Considering the growing diversity in the United States, the practices of accent management and teaching of pronunciation to English as Second Language Learners (ESL) have been increasing in prevalence and popularity over the last fifty years with a significant growth in the past two decades. Professional position papers in disciples of speech-language pathology and Teaching English as Foreign Language (TEFL) advocate for addressing accents and pronunciation as part of the professional curriculum. For example, the American Speech-Language-Hearing Association (ASHA) formally recognized the area of practice pertaining to dialect- and accent-

related services in two positions papers and the scope of practice (ASHA, 1983, 1985a, 2007). However, in contrast to the increase in clinical and teaching offerings to ESL clients, not much progress has occurred in providing evidence using standardized instruments and methods to move towards an evidence-based practice of accent management (See detailed report in Shah, 2024). A recent systematic study (Gu and Shah, 2019) showed that programs attempting to provide accent-related training do not use standardized assessments nor measure outcomes in a systematic manner. This trend has not changed in findings from national surveys from as long as two decades ago (e.g., Schmidt and Sullivan, 2003; Shah, 2005). As a result, this area of practice for speech-language pathologists as well as TESL/TEFL teachers remains subjective, potentially impacting ESL clients' progress, morale, as well as finances since they pay out-of-pocket for these services. As a crucial step towards toward regulating and standardizing this area of practice, a recent assessment framework was reported in Shah, 2024[1], namely, the Comprehensive Assessment of Accentedness and Intelligibility (CAAI) Assessment Framework. The CAAI Assessment Framework provided theoretically grounded and clinically-proven areas to assess and methods to assess using theory-driven, practical assessment material and examples. Going further, to continue to build an evidence base foundation, Shah (2004) also describes a testing instrument that builds upon this CAAI Assessment Framework, namely, the CAAI Assessment Battery. This Test Battery has been used with clients in a large number of clinics and organizations and has received peer-reviews and endorsement from ASHA, developing and offering it as a nationally-approved webinar training for SLP clinicians (Shah, 2010). Going further on the path of standardization, the present paper reports on attempts to validate the CAAI Assessment Battery. Specifically, owing to the lack of reliability or validity measures on existing tests and tools for accents, for example the *Phonological Assessment of Foreign Accent* (Compton, 2002) and the *Proficiency in oral English communication (POEC;* Sikorski, 2002), the present paper assesses for internal reliability, internal consistency, test-retest reliability, specificity, and sensitivity measures with the CAAI Assessment Battery. The following sections will first provide a short summary of the CAAI Assessment Battery and then report the details of the psychometric assessments and the corresponding results.

CAAI: Introduction and Orientation

The Comprehensive Asssessment of Accentendess and Intelligibility Assessment Battery (CAAI: Shah, 2007a), was developed to meet the needs and fill the gaps in the assessment of foreign-accented speech and dialectal variations. This test is designed to assess and diagnose dialect- and accent-related communication difficulties in an objective, quantitative manner. The objective of the test is to help make assessments and follow-up training/teaching of clients as evidence-based and data-driven as possible. The test is *comprehensive*, in that it targets a multitude of typical errors that range from syllable-level to discourse-level communication abilities. The areas tested are those that specifically pose communication difficulties for dialect and accented clients. With its comprehensive attempt to capture the errors, the test ensures that all communication issues are identified in the beginning, so subsequent teaching/training is effective.

The CAAI assessment battery includes a full set of stimuli material to conduct the assessment, as well as a test manual, an examiner manual, a scoring form, a response form, as well as a case history form. As mentioned earlier, the CAAI Assessment Battery uses a systematic framework

[1] The original version of the CAAI Assessment Framework first appeared in Shah, 2007b and Shah, 2009a)

URL: http://dx.doi.org/10.14738/assrj.115.17016

for targeting and assessing each area—the CAAI Assessment Framework. While the framework can be used with other examples outside of those used from the CAAI Assessment Battery, it is most efficacious and reliable if used together, as the present paper will demonstrate.

Table 1 (taken from the test manual of the CAAI, Shah, 2007a) shows a listing of the 22 sections of the CAAI Assessment Battery and the broad categories of assessment areas they represent.

Table 1: Broad categories of the areas of assessment, and specific section numbers underlying each broad category (taken from the CAAI Assessment Battery; Shah, 2007a).

Section No.	Section Title (I.E., The Measured Area)	Broad Categories	
1	Intelligibility relative to accentedness	**Baseline intelligibility and rate of speech**	
2	Intelligibility score & Rate of speech on narrative passage		
3	Sentence-level intonation	Suprasegmental aspects of varying length (from syllables to sentences)	**Speech production abilities**
4	Word-level intonation		
5	Lexical stress in single words of varying syllable-length		
6	Derivative stress in multisyllabic words		
7	Contrastive lexical stress		
8	Emphasis		
9	Sentence phrasing		
10	Phrasing contrast in sentence pairs		
11	Consonant word list (C)	Articulation of sounds (C, V, clusters)	
12	Consonant clusters word list		
13a	Vowels word list: General words (V)		
13b	Vowels word list: Specific words		
14	Phonological processes	Phonological patterns	
15a	Auditory Discrimination (paired contrasts)	**Speech perception**	
15b	Auditory Discrimination (labeling of single words)		
16	Prepositions	Prepositions	**Language areas**
17	Colloquial/idiomatic use of prepositions	Idioms	
18	Contrasting idiomatic phrases: expression & comprehension		
19	Comprehension of idiomatic expressions		
20	Advanced vocabulary	Vocabulary	
21	Conversational grammar & parts of speech (syntax, morphology, and semantics)	Grammatical and semantic skills	
22	Pragmatic problems	Social-pragmatic skills (including nonverbal communication)	

These 22 areas form the basis of the CAAI Assessment Framework (see Figure 1), described in detail in Shah (2024). The broad categories of the CAAI Assessment Battery as well as the underlying CAAI Assessment Framework include speech production, specifically, *articulatory* and *phonological* errors at the *segmental level* (consonant and vowel production) and at the *prosodic level* (e.g., stress, emphasis, and intonation). Speech perception is also tested: *Auditory and perceptual discrimination* abilities are tested, to make judgments about the underlying causes of accented speakers' production errors, and thereby helping to predict goals and target areas for intervention. In addition to the above speech production and perception issues, the test also assesses *language performance* in the grammatical, semantic, and pragmatic domains.

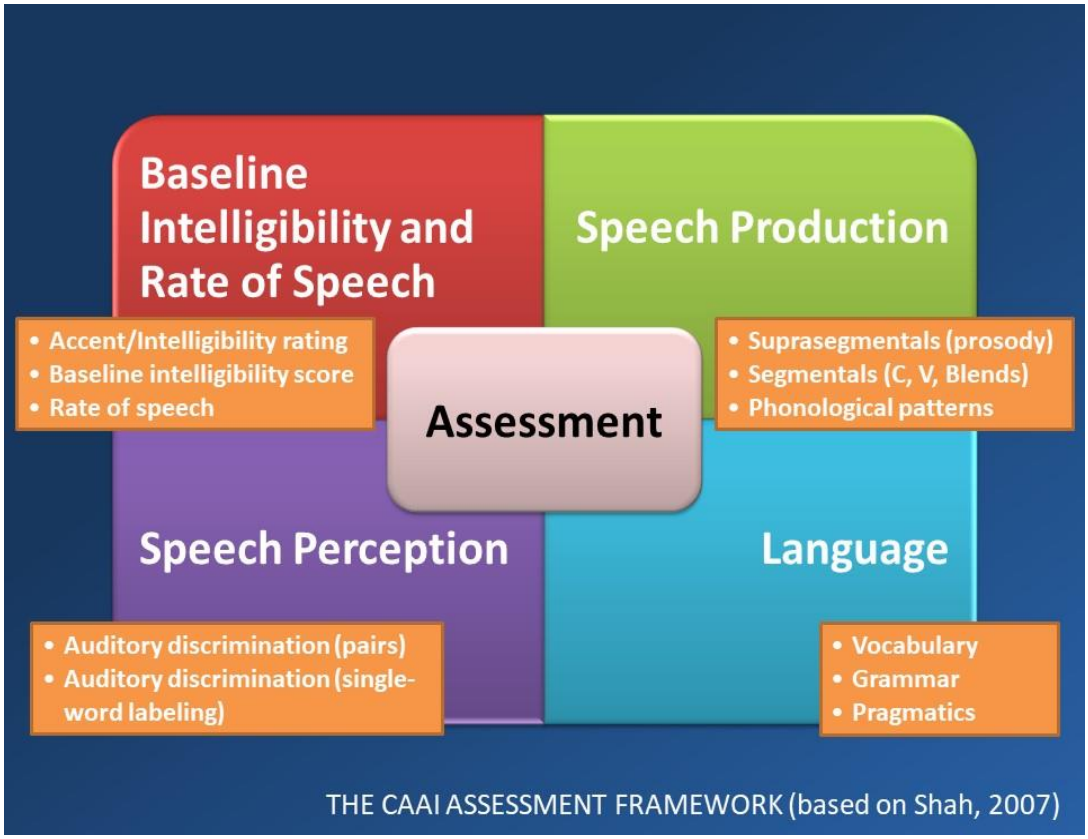

Figure 1: CAAI Assessment Framework (based on Shah, 2007a)

The larger objective of the CAAI Assessment Battery is to enable an evidence-based practice by developing a *systematic, scientific, quantitative approach* to the assessment of accents and dialects. To that end, the CAAI Test Battery provides diagnostic indicators and section scores in each area assessed and helps arrive at severity ratings for each area assessed. Specifically, numerical values for 20 sections (two qualitative sections, Section 21 on grammar, and Section 22 on pragmatics are excluded) can be compared to normative reference measures to see if each target is mildly, moderately, or severely affected in individual clients compared to native, English-speaking normed samples. These section-score numerical values can be compared before and after therapy to objectively evaluate progress. Additionally, prognostic indicators for therapy can be quantitatively estimated from the section scores. Refer to Shah (2024) for additional details of the theoretical premise of the CAAI Assessment Battery, its components, and detailed assessment procedures for using the CAAI Test Battery.

The purpose of the present paper is to describe the psychometric properties and analyses of the CAAI Assessment Battery. This paper describes the following statistical measures and their methods: test sensitivity, test specificity, and three areas of test reliability, including interrater, internal consistency/inter-item, and test-retest reliability.

METHODOLOGY FOR TEST VALIDATION

Early Pilot Procedures

The test stimuli were created and tested on 40 participants (10 native and 30 nonnative), over a two-year period. Based on the difficulties experienced during administration, issues faced during scoring, and other inconsistencies observed in participant performance, the test was revised to the present-day version (2007a). The revision involved the following changes:

1. The test was separated into sections instead of one continuous piece; the administration book, response sheet, and scoring book were all correspondingly divided into sections. The purpose was two-fold: ease of administration as well as independent sections that could be selectively omitted and/or used in any order, depending on the clinician's judgment of individual client needs. Separate section-by-section instructions were created, tested, and included in the later version of the CAAI (2007a) administration booklet as well as in the individual scoring form.
2. The entire test was re-formatted and additional seven items were included: a Spiral-bound *Test Manual, a Case History*, *Interview*, and *Language Background Questionnaire*; a Spiral-bound *Administration Stimuli* book; a Spiral-bound *Response sheets* book; a cut-out laminated *response-masker*; an *Individual Scoring Form*; and an *Individual Diagnostic Profile* form.
3. The response-masker was considered necessary for inclusion based on participants' performance during pilot testing. Participants' reading pace tended to be fast and various stimuli responses tended to run together, making it difficult to discern the word boundaries, clarity of segmental productions, or naturalness of prosodic patterns. The CAAI response-masker is now used by the examiner to selectively show participants the response items one at a time rather than the whole page at once, and it thus better manages the pace of responses.
4. Instructions were revised throughout the sections. Separate versions of instructions were provided for administration and scoring and included in the administration and scoring forms. Color-coding was used to differentiate various types of instructions for rapid administration and scoring.
5. Additional optional prosody contours were included based on the new ones produced by native speakers tested in the pilot study. Variable segmental and prosodic patterns produced by native speakers were included throughout the test as optional patterns to expect in native test subjects/clients and to be scored as correct based on this normative sample.
6. New items were included in the sections in which pilot testing did not capture a sufficient diversity of responses. Redundant and multiple items that appeared to not add new information and made the test longer were deleted.
7. Based on the response patterns seen in the nonnative sample, a separate section for vowels (Section 13b) was created. Similarly, the section on phonology (Section 14) was added to distinguish the phonological errors from production/articulation errors.

Normative Data Collection Details

Participants and Testers:

A total of 61 clients (native=28; nonnative=33) were each individually administered the post-pilot, updated CAAI Test Battery. Table 2 shows the summarized demographic details of the participants and the testers, which are described in detail as follows.

Native:

The native group consisted of participants aged 19-38 years (mean=25 yrs, S.D.= 4.76) with 11 males and 17 females who were monolingual English speakers residing in the Cleveland, Ohio area, and born and raised in a city of Ohio or its suburbs. Local regional dialect was permitted to capture the range of deviation from Standard American English to contrast with foreign-accent related deviations. These were typically undergraduate and graduate students (average= 14 years of education).

Nonnative:

The nonnative group consisted of participants aged 19-53 years (mean=32 yrs, S.D.= 9.83) with 18 males and 15 females who spoke 2-3 languages each including English, which was their second or third language. They were all residing in the Cleveland, Ohio area, and born and raised in a foreign country, including Korea, India, Saudi Arabia, Ukraine, Ireland, Argentina, China, Russia, Armenia, Japan, Peru, and Egypt.

These participants were selected from two participant pools: international graduate and undergraduate students on the Cleveland State University campus (the low proficiency group— with a distinct accent but sufficient fluency in English), and medical doctors and research scientists from the Cleveland Clinic (the high proficiency group— also with a distinct accent but advanced communication skills). The education level of this participant group ranged from undergraduate to doctorate (M.D. or Ph.D.). Their age of arrival (AOA) in the United States ranged from 6 to 41 years, and was on average 24 years, with a standard deviation of 9.9; this variable is used in the second-language literature to represent age of learning the second language (e.g., Flege, Munro, & MacKay, 1995). This wide range in the age of learning the language was expected to yield a range in the strength of the foreign accent as well as a range in the overall proficiency in English, thus helping represent a wide diversity in this normative pool. The participants' Length of Residence (LOR) in this country ranged from 1 to 24 years, and was on average 8.6 years. Once again, this variable was expected to capture a range of communicative proficiency in English as LOR has been shown to be predictive of difference in the degree of accents and overall L2 proficiency (e.g., Piske, MacKay, & Flege, 2001).

Testers:

Trained graduate students (n=4) from the speech-language pathology program at Cleveland State University administered the test, listened to the recordings, and conducted the scoring. These listeners were female, native, monolingual speakers of English residing presently, as well as born and raised in, the Cleveland, Ohio area. Their ages ranged from 22 to 24 years. All of the testers had undergone a basic training for using the CAAI Assessment Framework and following the CAAI test manual, test administrator's book, and the scoring form.

URL: http://dx.doi.org/10.14738/assrj.115.17016

Table 2: Demographic table of participants

Participants	N	Age	Gender	Education	Languages	AOA	LOR	Born & Raised
Native	28	Mean=25 years; S. D=4.76	M=11, F=17	Undergraduate or graduate students	Monolingual English	n/a	n/a	Ohio cities & suburbs
Nonnative	33	Mean=32 yrs; S. D=9.83	M=18, F=15	Ranged from undergrad to doctorate; most held M.D/Ph.D.	Various L1s; English was L2 or L3	Mean=24 yrs; S. D= 9.9	Mean= 8.6 years S. D=5.79	Korea, India, Saudi Arabia, Ukraine, Ireland, Argentina, China, Russia, Armenia, Japan, Peru, Egypt
Testers/ Scorers	4	Mean=23.2 yrs; S. D=.83	M=0; F=4	Graduate (SLP students)	Monolingual English	n/a	n/a	Cleveland, Ohio

Procedure

Prior to testing, each of the test administrators independently read the manual[2] for initial orientation and became familiar with the test and its administration. Each tester practiced administration on 1-2 people, which was audio and video recorded. Following this, the tester was provided a debriefing session by the Principal Investigator (Amee Shah) to provide feedback about the problems with administration, if any, and were given pointers to correct these. Once the method of administration was mastered, as determined by the P.I., the tester then collected data on all the native and nonnative participants, as described in the following paragraph.

The test administration with the participants began with a language background questionnaire (included in the CAAI battery), followed by the full administration of the CAAI battery. Each participant was tested with the entire CAAI battery (approximately lasting 1.5 hours, including the initial questionnaire/interview time). The assessment protocol involved elicitation of information through tasks such as answering simple questions, naming and/or pictures, and reading aloud words/sentences/passages. The testing took place in a sound-treated booth in the Speech Acoustics and Perception Laboratory at Cleveland State University. Audio and video recordings were obtained of the participants as they attempted the various areas of assessment. While participants' consent for the recordings was obtained prior to the testing, the recorders were kept out of sight of the participants so as not to let them become self-conscious and thereby less natural in their responses. The recordings were considered important for observing the details of articulatory and phonological differences as well as noting grammatical or pragmatic differences throughout the session, which may be difficult to capture in an on-line assessment session. The audio recordings were digitized, converted to .wav format, and opened later in a speech viewer program, Sound Forge (version 4.5, Sonic Foundry, 1991-1998), to

[2] In requiring the testers to read the manual on their own and orient themselves to the test with its help, it was expected that this exercise would serve as a test of the technical manual and revisions to the manual and/or other components of the test could be made accordingly.

enable seeing individual waveforms, and zooming in to see and hear details of smaller segments that the digital recorder would not allow. Individuals computing the scores used this speech viewer program instead of the playback option on the tape recorder to listen to the samples while scoring the CAAI Test Battery.

Scoring was conducted by two trained SLP student clinicians. Each judge/scorer independently assigned her scores. Scoring for each participant took on average two hours. The scorers then met and discussed discrepancies. Discrepancies in responses were resolved by a third listener, who was a trained clinician as well. Discrepancies helped determine reliability of test administration and stability of the instrument in its interpretation across two raters, as computed in the interrater reliability analysis below. Discrepancies were also used as a guide to identify weak items, to be later replaced, modified, or administered with better instructions, to control any ambiguity in their administration.

Sections 21 and 22 were not intended for a quantitative analysis; the scorers made qualitative notes of participants' grammar problems, e.g., "incorrect use of plural markers" in Section 21, and they noted the pragmatic issues seen throughout the session, e.g., "difficulty with topic elaboration and gives brief, non-detailed answers needing many prompts for details" in Section 22 results. For each of the remaining 20 sections, descriptive statistics were calculated from the responses, and a narrative summary was formulated to identify trends. Each of the 20 sections were scored across all the participants. A comparison of the foreign-accented with the unaccented speakers yielded baseline norms for accent-related speech deviations. Further statistical analyses helped arrive at various psychometric data, including test sensitivity and specificity, content and face validity, predictive validity, concurrent validity, inter- and intra-rater reliability, internal consistency, and test-retest reliability of each of the sub-tests. The following section reports on the results of the sensitivity, specificity, and reliability measures; all of the validity measures will be presented in a future publication (in the interim, Shah, 2009b provides preliminary validity findings).

RESULTS OF PSYCHOMETRIC ANALYSES

Test Sensitivity

Definition:

"A sensitive test is one that rarely fails to identify a disorder or disease…such a test is expected to cast a wide net, rarely missing people who have a disorder or disease" (Maxwell & Satake, 2006, p.119).

Method:

Sensitivity is computed as a proportion of people who test positive to those with a disease, disorder, or as in this case, accent-related communication difficulties.

Results:

For higher sensitivity standards, a test should be able to identify a large proportion of the nonnative speakers' accent-related concerns. Table 3 shows the mean scores of the native group versus nonnative on each of the 20 sections of the test.

URL: http://dx.doi.org/10.14738/assrj.115.17016

Table 3: Group Means and Standard Deviations (measures for test sensitivity and specificity).

Sec. No.	Section Name	Mean N	Std Dev N	Mean NN	Std Dev NN*
1	Accent rating (average)	1.03	0.11	3.20	0.99
	Accent rating (range on 5-pt. scale)	1-2		1-4	
2	Baseline Intelligibility Percent	100%	0.00	95.00	23.51
3	Sentence level intonation	97%	4.83	90%	14.14
4	Word Level Intonation	97.50%	7.91	90%	15.28
5	Lexical Stress in Single Words	97.33%	4.32	80.83%	10.74
6	Derivative Stress in Multisyllabic Words	95.42%	4.99	84%	11.93
7	Contrastive Lexical Stress	94.64%	7.94	72%	13.25
8	Emphasis	96.67%	4.30	84.44%	16.63
9	Sentence Phrasing	92.86%	10.10	88.57%	14.00
10	Contrasting Sentence Pairs	88%	12.29	86.42%	11.09
11	Consonants	99.38%	1.08	88.41%	9.72
12	Consonant Clusters	98.47%	2.32	91.19%	7.33
13a	Vowels	97.19%	3.96	80.37%	24.25
13b	Vowels	99.09%	2.87	89.09%	9.21
14	Phonological errors	97.04%	4.04	63.44	11.28
15a	Auditory Discrimination	98.9%	2.45	48.24	13.47
15b	Auditory Discrimination	98.4%	3.23	42.82	18.23
16	Prepositions	86.50%	8.83	82.00	10.77
17	Colloquial/Idiomatic Use of Prepositions	78.75%	10.29	38.33	24.78
18	Contrasting Idiomatic Phrases	82.96%	20.09	42.35	30.16
19	Comprehension of Idiomatic Expressions	95%	5.83	56.11	23.27
20	Advanced Vocabulary	71.50%	19.16	52.22	25.95

*N=Native; NN=Nonnative

Comparison of the mean scores shows two consistent findings. First, the native scores are significantly different from, and consistently higher than, the nonnative scores [t (38) = 4.52, p<.001]; native participants clearly performed better than the nonnative participants. Secondly, and more importantly, the nonnative participants received a mean accent rating of 3.20 (on a scale of 1-5, 5 being highly accented and least intelligible) with a low standard deviation (s.d.) of 0.99, which shows, on average, that the test was strongly and consistently able to identify accented individuals. Furthermore, a range of 38% to 95 % in the nonnative scores across the test sections shows that the test was able to identify the range of accent-related communication difficulties. With two ways of tracking accent issues, namely accent rating and the individual section test scores, the CAAI Test Battery rarely (if at all) misses individuals with accent-related difficulties, and is thus *highly sensitive*.

Test Specificity

Definition:

"A specific test is one that seldom erroneously identifies a person as having a disorder or disease when they do not" (Maxwell & Satake, 2006, p.119).

Method:
Specificity is computed as a proportion of people who test negative on the test without a disease, disorder, or as in this case, accent-related communication difficulties.

Results:
The ideal specificity standard for this test would be to ensure that no native speakers get falsely identified as nonnative speakers of English and/or as having speech or language issues related to their accent. The test should indeed help identify the presence of regional dialect-related concerns that need intervention. Table 3 shows that the native group consistently received accent ratings of 1-2, and on average, received an accent rating of 1 (on a scale of 1-5, where 1= Negligible difference from Standard American English speakers; and 2= Some trace of accent but completely intelligible). A low standard deviation of 0.11 indicates that this range represented all the native speakers in the group and that no one received a rating greater than 2. Thus, this test was consistently able to identify all the native speakers as native Standard American English speakers, with a mild regional accent, at most. Moreover, the baseline intelligibility percent was 100% for everyone in the native group (s.d.=0), indicating that native speakers were never falsely identified as having intelligibility concerns. Furthermore, individual section scores shown in Table 3 indicate that 13 of the 20 sections from sections 3-20 were scored as 95 % or greater in accuracy for the native group, suggesting absence of any communication difficulty, or at most, a mild regional variation of dialect. The remaining sections identified mild-moderate concerns related to language issues (sections 16, 17, 18, and 20), those related to prepositions, idiomatic expressions, and advanced vocabulary, respectively. Three sections related to grammar-related phrasing contrasts (Contrasting Lexical stress/section 7, Sentence phrasing/section 9, and Contrasting Sentence Pairs/section 10) also showed mild difficulty. These mild-moderate communication concerns in the language areas are consistent with normal variation in English proficiency expected among native speakers. The test was thus found to be *highly specific* as it does not misidentify native speakers as having communication concerns that need intervention. Mild variations in speech due to regional accents, where present, were correctly identified.

Interrater Reliability
Definition:
Interrater reliability refers to the extent to which two or more testers obtain the same result when using the same instrument to measure a concept. This measure helps assess the accuracy of judgment on the part of the tester. More importantly, this measure helps to verify whether a test can yield comparable results across (qualified) testers.

Method:
As described in the procedures section above, each participant was tested and rated by two separate listeners. These testers/raters were considered "qualified" as they had undergone training with the CAAI Assessment Framework (like the one described in Shah, 2024) and using the CAAI Test Battery manual, administrator book and scoring book. Without baseline training, it is possible to bias the testing with individual opinions, variable levels of experience, and individual demographic variables.

Each tester/rater conducted the scoring independently, and then discussed the findings with each other to arrive at discrepancies. Means were calculated for each participant's individual

URL: http://dx.doi.org/10.14738/assrj.115.17016

section scores. Pearson correlation coefficients were analyzed (for mean scores pooled over items within each section and the native and nonnative participants) to compare correlations between the two testers/raters. Each section was considered separate in its purpose, so correlation coefficients were obtained for each section separately.

Results:

Table 4 shows the three sets of correlation coefficients computed: Interrater reliability, Cronbach's alpha, and test-retest reliability correlation coefficients. Columns 1 and 2 show the section numbers and names, respectively. As seen in column 3, the interrater correlation coefficients ranged from 0.68 to 1.00, and were all highly significant ($p<.01$). These coefficients are statistically *meaningful,* as the r^2 ranges from 0.46 to 1.00, indicating that the variance shared between the two raters equals 46 to 100.

Table 4: Correlation Coefficients: Interrater reliability, Cronbach's alpha, and Test-retest reliability.

Sec. No.	Section Name	Interrater Correlation Coefficient*	Cronbach's alpha	No. of Items in each section	Test-retest Correlation Coefficient**
1	Accent rating (average)	0.95	1.00	12	0.98
	Accent rating (range)	0.95	n/a	n/a	n/a
2	Baseline Intelligibility Percent	1.00	n/a	1	1.00
3	Sentence level intonation	0.71	0.46	10	0.82
			(0.52 if item 5 removed)		
4	Word Level Intonation	0.77	-0.20	4	0.86
			(0.10 if item 3 removed)		
5	Lexical Stress in Single Words	0.94	0.51 12 **(0.55 if variable 12 removed)**		0.88
6	Derivative Stress in Multisyllabic Words	0.99	0.75	9	0.90
			(0.77 if variable 7 removed)		
7	Contrastive Lexical Stress	0.88	0.84	20	0.94
8	Emphasis	0.91	0.72	10	0.89
			(0.74 if variable 3 removed)		
9	Sentence Phrasing	0.68	0.45	7	0.92
			(0.58 if variable 2 removed)		
10	Contrasting Sentence Pairs	0.77	0.22	10	0.81
			(0.341 if variable 15 removed)		
11	Consonants	0.93	0.90	65	0.89
12	Consonant Clusters	0.88	0.86	58	0.93
13a	Vowels	0.87	0.91	17	0.79
13b	Vowels	0.85	0.95	22	0.80
14	Phonological errors	0.92	0.74	13	0.73
15a	Auditory Discrimination	0.95	0.62	91	0.78

15b	Auditory Discrimination	0.94	0.56	75	0.75
16	Prepositions	0.96	0.38	19	0.88
			(0.478 if variable 9 removed)		
17	Colloquial/Idiomatic Use of Prepositions	0.95	0.72	8	0.97
18	Contrasting Idiomatic Phrases	0.80	0.80	17	0.93
19	Comprehension of Idiomatic Expressions	0.88	0.78	12	0.97
20	Advanced Vocabulary	0.99	0.79	20	1.00
		* all sig at .01 Level ** all sig at less than .05 level			

Internal Consistency

Definition:

Internal consistency of a test is a measure of how well items that measure similar traits are correlated. Cronbach's alpha is a type of correlation used typically to determine internal consistency of a measure. While a good test is expected to have a generally high internal consistency, it is typically believed in the social sciences that a good test would likely show only moderate correlation (0.70 to 0.90) among items (e.g., Nunnally & Bernstein, 1994; Streiner & Norman, 2003), as higher correlations would suggest that some of the items are redundant whereas lower correlations may suggest that items may be measuring different traits and should not be included under the same category.

Method:

As mentioned in the previous section of Interrater reliability, the two testers/raters assigned scores independent of each other for each participant that they had tested. Each participant was thus scored by one pair of testers/raters, who then met and discussed the discrepancies in the scoring, and where possible, resolved the discrepancy with a mutually agreed upon score. For the scores that could not be resolved, a third rater (with the same training and experience as the other raters) was used to help break the tied scores. As a result, one consolidated score was available for each item per section for each of the native and nonnative participants. Cronbach's alpha correlations were tested for the spread of scores (n=61 native and nonnative participants) for each test item with the corresponding scores of the other items within each section. For example, section 1 consisted of 12 items, and Cronbach's alpha correlation was estimated over these 12 items. Table 4, column 4, shows the Cronbach's alpha correlation coefficients to show the internal consistency across the number of items included (column 5) in each section of the CAAI.

Results:

As shown in Table 4, column 4, Cronbach's alpha ranged from 0.22 to 1.00 (except for section 4, where a negative correlation was seen). As predicted, a large number of the sections (n=10 of the 20 total sections examined) yielded moderate correlations (0.70- 0.90), showing *good internal consistency* of the items within those sections. Three sections yielded very high correlations (0.91-1.00) and were thus considered to have redundancies with at least some of the included items. However, these items do not warrant elimination as they add additional

URL: http://dx.doi.org/10.14738/assrj.115.17016

clinical information, and since they comprise only two sections, it was determined that they will not take too much of the total test-administration time. A small number of sections (n=6 of the total 20 sections) yielded low correlations. With weaker items removed, a re-scaling suggested markedly increased Cronbach's alpha coefficients (an option called, "scale if item deleted" within the Reliability Analysis/Cronbach's alpha analysis, as provided by SPSS statistical package). These weak items and the improved coefficients for these low-correlation sections are indicated in bold font under Column 4 in Table 4. While at first glance it may appear that the items with the low correlations are testing a variable other than the overall section the items are part of, the inclusion of these seemingly weak items do not actually weaken the test reliability. In fact, it appears that it is these items, per se, that help identify accent-related speech and language difficulties that would otherwise be missed, as the scores on those items were sizably lower than on the other items. Indeed, these items may be essential to that section. The fewer the number of test items within a section, the greater the role each item plays in drawing out the communication difficulties. For example, section 4 shows a negative correlation, presumably due to the small size of that section (n= 4 items), which then makes each item essential and clinically significant. In other words, lower statistical correlations in this case are offset by the clinical significance of these items. The latter premise is substantiated by the validity data using the CAAI Test Battery (Shah, 2009b).

Overall, the test has a high *internal consistency* between the items within each section. There is sufficient diversity between the items, so that they are not redundant in measuring the same exact trait. Instead, the diversity across the items in a section helps identify the underlying problem in each test area through a set of distinct examples. The moderate-high correlation coefficients show that the items are indeed related within that category, and are not completely random.

Test-Retest Reliability

Definition:

The test-retest reliability measure helps estimate the consistency of the test over repeat administrations. Participants are administered the same scale on two different occasions to compare the consistency across testing and scoring methods.

Method:

An underlying assumption of this analysis is that the phenomenon being measured is stable and does not fluctuate over time, e.g., I.Q. However, the CAAI measures language skills that change rather quickly as part of the acculturation process of the nonnative speakers, exposure to English that the nonnative speakers gain daily, and their consequent ongoing learning and shaping in the second language. Studies have shown this change in second language abilities can occur over as short a period of exposure as four to eight hours (e.g., Rast & Dommergues, 2003); various studies have shown that a degree of length of residence in the nonnative country (LOR) is predictive of second language learning (Piske, MacKay, and Flege, 2001 for a review). Thus, even if a measure is reliable in the way that it lends itself to its administration over repeated occurrences, the measure would not show such stability if it were used to test a dynamic, changing population. Considering this concern, the test-retest reliability for the CAAI was measured in a modified procedure. Instead of re-testing the clients, video recordings of the client were made at the time of the first testing. These video recordings were then scored by one examiner immediately following the test administration (Time 1), and then re-scored,

blindly, by the same examiner a month later (Time 2). A month between testing was considered sufficiently long on the part of the examiner to override any practice effects of administering the test, scoring the test, and/or remembering any particular behaviors of the client. The rater did not receive any additional exposure to the test or to the participants, or any further training on the test administration during this period.

Results:

Correlation coefficients were obtained for the Time 1 and Time 2 scores for the entire corpus (n=61). As shown in Table 4, column 6, correlation coefficients for test-retest reliability ranged from 0.73 to 1.00, all of them being significant ($p<0.05$), indicating a *high test-retest reliability* of the CAAI.

DISCUSSION

While dialect- and accent-related communication concerns have become a routine area of practice in speech-language pathology and in English-as-second-language classrooms, there is a paucity of validated resources in this area of practice. As an important preliminary step, the Comprehensive Assessment of Accentedness and Intelligibility (CAAI) Test Battery was developed to enable evidence-based, data-driven, standardized assessment and therapy practices of dialect- and accent-related communication concerns. The availability of the assessment battery and clinical research developed because of the CAAI Test Battery will likely prepare clinicians and teachers to base their assessment (and teaching/managing) of foreign-accented speech in a manner that is better grounded in theory and supported by research data. The CAAI Test Battery helps clinicians to address questions related to a) what areas to test, b) how to test these areas, and c) how to quantitatively and objectively make diagnostic, prognostic, and therapeutic recommendations for foreign-accent therapy. A detailed Assessment Framework with areas to assess, methods to assess, and examples for assessment are described and published elsewhere (Shah, 2024)

As a crucial step towards standardization, validation efforts were made in an area of testing that lacks any reports of psychometric analyses. To collect normative data and test the psychometric properties of the CAAI Test Battery, two large studies were conducted to address reliability and validity measures, respectively. The present paper reports the methods and results of data collection and analysis to compute measures including test sensitivity, specificity, and three types of reliability, namely interrater, inter-item/internal consistency, and test-retest reliability. Results showed that the CAAI meets high standards of test sensitivity in identifying nonnative speakers of English with communication difficulties related to their foreign accent and intelligibility concerns. The test also helps identify which specific areas of communication the nonnative speakers may have difficulty with, and the extent of this difficulty. The test shows high specificity in that it did not misidentify native speakers of English as those with a foreign accent, and/or having accent-related communication difficulties when they did not have any regional dialectal variations from Standard American English speakers (SAE). For those native speakers with regional variations of dialect, the test was able to sensitively identify the type and amount of their dialect-related difficulties. The test showed high interrater reliability with correlations of 0.68 to 1.00, which were all highly significant ($p<.01$) and statistically *meaningful,* as the variance shared between the two raters equaled 46 to 100. The test showed high inter-item reliability/internal consistency as indicated by Cronbach's alpha range of 0.22 to 1.00. A large majority of the sections examined (n=14 of the

URL: http://dx.doi.org/10.14738/assrj.115.17016

20 total sections) yielded moderate to strong correlations (0.70- 1.00), representing *good internal consistency* of the items within those sections. Overall, there is minimal redundancy and sufficient diversity between the items, so that the different items help identify the underlying problem in each test area through a set of distinct examples, and the moderate-high correlation coefficients show that the items are related within that category, and not completely random. Test-retest reliability of the test was examined in a modified procedure which does not address test-retest administration, per se, but does help evaluate consistency across repeat scoring attempts. Thus, there is *high test-retest reliability*, or specifically, high consistency in repeat scoring attempts, with correlation coefficients ranging from 0.73 to 1.00, significant at $p<0.05$. Furthermore, with a month between testing intervals, the correlation coefficients do not simply reflect a practice effect on the part of the tester/examiner.

In conclusion, due to its high sensitivity, specificity, and reliability, the CAAI Test Battery is found to be a stable and meaningful means to assess dialect- and accent-related communication concerns. The CAAI Test Battery helps fill a crucial gap in the area of accents/dialects due to lack of any other validated assessment measure. With a sensitive and reliable assessment, clinicians and teachers can identify an accurate baseline pattern of errors to address for accent management or pronunciation teaching and achieve effective outcomes in a short amount of time. For example, Shah (2023a), and Shah (2023b) show two case study models that use the CAAI Assessment Framework and CAAI Test Battery to achieve effective outcomes in a low-proficiency of English speaker and a high-proficiency of English speaker in a short 8-12-week duration of sessions that were 60 minutes/week.

FUTURE DIRECTIONS

To supplement the reliability and sensitivity/specificity measures described in the present paper, a second upcoming paper will describe the methods and results of testing various validity measures of the CAAI Test Battery (preliminary results in Shah, 2009b). Among the questions for assessing validity asked were: 1) Is this test useful? 2) What is its face validity as evaluated by clinicians? 3) Does the test identify important communication difficulties in the participants? 4) How does it compare to an existing gold standard of intelligibility (concurrent validity)? 5) In the absence of a gold standard for accent-related communication issues, what is the construct validity of this test? and 6) Can this test predict severity of the communication difficulty and the amount of intervention required (predictive validity)? Finally, research is in progress to test various theoretical constructs that will likely serve well in making decisions about teaching and managing accented pronunciations. For example, the question of whether to prioritize prosodic deviations before segmental ones is a theoretical prediction that to date remains untested through clinical research. Continued research in this area will strengthen clinical practice around dialect- and accent-related communication concerns.

ACKNOWLEDGEMENTS

This research was supported in part by the Faculty Research Development Grant from Cleveland State University and Research and Professional Development Grant from Stockton University to Amee P. Shah. The author acknowledges the students and associates of the Speech Acoustics & Perception Laboratory at Cleveland State University and the Cross-Cultural Speech, Language, and Acoustics Lab at Stockton University for their assistance with data collection, interviews, and analyses. The author is grateful to the Office of Diversity at the Cleveland Clinic for assistance with recruiting participants for this project. Statistical consulting with Drs.

Michael Horvath & Steve Slane from the Department of Psychology at Cleveland State University and research affiliates at Stockon University is greatly appreciated. Portions of this peer-reviewed work were presented at the meetings of the American Speech, Language, and Hearing Association (Shah, 2005; Shah, 2007c), ASHA Connect, ASHA national webinars, Speechpathology.com, and Cross-Country Education Seminars.

References

American Speech-Language-Hearing Association. (1983, September). *Social dialects and their implications* [Position paper]. ASHA, 25, 23–27.

American Speech-Language-Hearing Association. (1985a, June). *Clinical management of communicatively handicapped minority language populations* [Position paper]. ASHA, 27, 29–32.

American Speech-Language-Hearing Association. (2007). *Scope of practice in speech-language pathology*. http://www.asha.org/policy/

Compton, A. (2002). *Phonological assessment of foreign accent*. San Francisco: Carousel House.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication*, *16* (1), 1-26.

Gu Y., & Shah A. (2019). A systematic review of interventions to address accent-related communication problems in healthcare. *Ochsner Journal*. 2019 Winter;19(4):378-396. doi: 10.31486/toj.19.0028

Maxwell, D. L., & Satake, E. (2006). *Research and statistical methods in communication sciences and disorders*. Clifton Park, NY: Thomson Delmar Learning.

Nunnally, J. & Bernstein, I. (1994). Psychometric theory. New York: McGraw-Hill.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*, 191-215.

Rast, R., & Dommergues, J. Y. (2003). Towards a characterization of saliency on first exposure to a second language. In *EUROSLA Yearbook* (Vol. 3, pp. 131-156). Amsterdam: John Benjamins Publishing.

Schmidt, A. M., & Sullivan, S. (2003). Clinical training in foreign accent modification: A national survey. *Contemporary Issues in Communication Science and Disorders, 30*, 127-135.

Shah, A. P. (2005). *Accent modification: Is it efficacy-based? Results from a nationwide survey*. Paper presented at the Annual conference of the American Speech-Language-Hearing Association (ASHA), San Diego, CA, abstract in *ASHA Leader, 10* (11), p. 86, 0513.

Shah, A. P. (2007a). *Comprehensive Assessment of Accentedness and Intelligibility (CAAI)*. EBAM Institute, L.L.C. https://caaiassessment.com/

Shah, A. P. (2007b). *A data-driven approach to comprehensive assessment of foreign-accented speech*. Paper presented at the Annual conference of the American Speech-Language-Hearing Association (ASHA), Boston, MA, abstract in *ASHA Leader,* 12 (11), pp. 105, 0683.

Shah, A. P. (2009a). Adopting evidence-based practices in accent modification: A data-based assessment model. *Speech Acoustics and Perception Laboratory Working Papers on Cross-Language Research, No. 5,* 1-41.

Shah, A. P. (2009b). Validation study of the Comprehensive Assessment of Accentedness & Intelligibility (CAAI). *Speech Acoustics and Perception Laboratory Working Papers on Cross-Language Research, No. 4*, 1-23.

Shah, A. P. (2010). Comprehensive Assessment of Foreign-Accented Speech. A peer-reviewed, national web seminar produced by the Professional Development office of the American Speech & Hearing Association, Rockville, MD. ASHA.org. Web seminar offered as a Continuing Education course from 2011-2017.

Shah, A. P. (2023a). Dr. JN: An Adult Nonnative Speaker of English: High Proficiency. *The Communication Disorders Casebook: Learning by Example*, 411-420. Plural Publishing, Inc. San Diego, CA.

Shah, A. P. (2023b). Ms. PW: An Adult Nonnative Speaker of English: Low

Proficiency. *The Communication Disorders Casebook: Learning by Example*, 421-443. Plural Publishing, Inc. San Diego, CA.

Shah, A. P. (2024). Assessing Accented Speech with a Data-Based Assessment Framework: The Key to Evidence-based Accent-Management. Advances in Social Sciences Research Journal, 11(5). 135-152.

Sikorski, L. (2002). *Proficiency in oral English communication (POEC)*. Santa Ana, CA: LDS & Associates.

Sound Forge (1991-1998). Version 4.5 [Computer software]. Madison, WI: Sonic Foundry. http://www.sonicfoundry.com/.

Streiner, D. L. & Norman, G. R. (2003). Health Measurement Scales: A Practical Guide to Their Development and Use. New York: Oxford University Press.