Accuplacer Companion in a foreign language context: An argument-based validation of both test score meaning and impact

Robert C. Johnson
Nursing Foundation Program, University of Calgary in Qatar
A. Mehdi Riazi
Department of Linguistics, Macquarie University, Sydney, Australia

Use of a single, standardised instrument to make high-stakes decisions about test-takers is pervasive in higher education around the world, including English as a foreign language (EFL) contexts. Contrary to longstanding best practices, however, few test users endeavour to meaningfully validate the instrument(s) they use for their specific context and purposes. This study reports efforts to validate a standardised placement test, used in a US-accredited, higher education institution in the Pacific, to exempt, exclude, or place students within its Developmental English Program. A hybrid of two validation structures – Kane's (1992, 1994) interpretive model and Bachman's (2005) and Bachman and Palmer's (2010) assessment use argument – and a broad range of types and sources of evidence were used to ensure a balanced focus on both test score interpretation and test utilisation. Outcomes establish serious doubt as to the validity of the instrument for the local context. Moreover, results provide valuable insights regarding the dangers of not evaluating the validity of an assessment for the local context, the relative strengths and weaknesses of standardised tests used for placement, and the value of argument-based validation.

Key words: assessment, testing, language testing, validity, validation

Robert C. Johnson, University of Calgary in Qatar, PO Box 23033, Doha, Qatar. E-mail: johnsor@ucalgary.ca.

Introduction

Best practices in educational testing clearly call for ongoing validity investigations for any assessment used to make important decisions about test-takers. This imperative comes not only from the recommendations of testing researchers (for example, Kane, 1992; Messick, 1989) and test publishers, but ethical and professional codes of conduct (AERA, APA, & NCME, 1985, 1999; Joint Committee on Testing Practices, 2015) and stated requirements of higher education accreditors (ACCJC & WASC, 2010). Despite this longstanding recommendation (and in the face of increasing use of tests to inform important decisions about individuals, programs, schools, and entire education systems), many in the literature lament a chronic lack of validation efforts. Of particular concern is the dearth of such efforts by test users, who should be investigating both the meaning of test outcomes and impact of test use within their particular context (Bachman, 2005; Kunnan, 2003; Xi, 2008).

Recently, it has been suggested this lack of in situ test validation may be a contributing factor in soberingly limited student success in basic/remedial/developmental English and mathematics programs, as well as ESL programs at junior and/or community colleges in the US. A number of recent reports, for example, point to disappointing program completion rates, student progress within specific programs, and demonstrated skills gains (Bailey, 2009; Bailey, Jeong, & Cho, 2010; Martorell & McFarlin, 2011; Offenstein & Shulock, 2011). Hughes and Scott-Clayton (2011) suggest a substantial part of the blame may lie with the dearth of investigations into the suitability of the placement instruments these institutions use, for their particular students, courses, programs, and educational objectives. Additionally, they propose that the use of a single, multiple-choice instrument, such as Accuplacer or Compass (Hughes & Scott-Clayton, 2011; Sullivan, 2008), to inform placement decisions at the majority of colleges, further contributes to the problem.

Until the current study, the host institution, like the vast majority of its US-accredited college brethren, used a standardised instrument, Accuplacer Companion (in addition to a locally designed and marked writing sample), to inform placement decisions about incoming students (Hughes & Scott-Clayton, 2011; Sullivan, 2008) but had never sought to validate either instrument for its context, learners, and purposes (Hughes & Scott-Clayton, 2011), as required by its accreditors, the Accrediting Commission for Community and Junior Colleges and the Western Association of Schools and Colleges (ACCJC & WASC, 2010). Also similar to many US-accredited colleges, student placement, retention, advancement, and achievement had been identified as ongoing problems requiring immediate action (Bailey, 2009; Bailey, Jeong, & Cho, 2010). In an

attempt to address these issues in the local context (described later), academic administrators, staff, and faculty members agreed to the establishment of an on-going validation effort for all instruments used to inform placement decisions at the institution. This paper reports the results of these efforts, as relates to Accuplacer Companion.

Accuplacer Companion

Accuplacer Companion (AC) is a multiple choice (4-option), standardised test, and is the paper-and-pencil version of Accuplacer OnLine, a widely used, web-based, adaptive placement test.

The paper-based version of AC was used by the institution due to the lack of electricity at some testing locations, and considerable variation in experience with computers amongst test-takers. The English subtests of AC used by the institution – Reading Comprehension and Sentence Skills – were designed for use with students for whom English is the language in which they are most proficient, or, as the test developers describe it, students 'for whom English is the best language' (College Board, 2003, p. A-11). These sections are intended to distinguish between such students who could directly enter into a credit level English course (in an English as a Second Language [ESL] context) and those who would best benefit from a semester of remedial English beforehand. As some readers may be aware, 'ESL' subtests, designed for use with English Language Learners, are also available for AC. The decision to use test sections intended for use with 'native' or 'near-native speakers' is addressed in further detail later, but appears to have been made by administrators at the school to ensure student eligibility for US educational grants (by using an instrument approved by the US Department of Education) without awareness of the availability of 'ESL' versions of the subtests, in paper-and-pencil format.

Each AC English subtest used consists of 35 questions. The Reading Comprehension test 'measures a student's ability to understand what he or she has read' (College Board, 2003, p. 17). The publishers identify five content areas addressed in this section of the test: Identifying Main Ideas, Direct Statements/Secondary Ideas, Inferences, Applications, and Sentence Relationships.

The Sentence Skills subtest was developed to assess candidates' comprehension of sentence structure: 'how sentences are put together and what makes a sentence complete and clear' (College Board, 2003, p. 19). The three content areas covered are: Recognizing Complete Sentences, Coordination/Subordination, and Clear Sentence Logic.

Local context

The host institution serves approximately 850 students, the vast majority of whom are Micronesian (98%), English Language Learners (98%), reliant on financial aid (99.5%), and academically underprepared (92% of accepted students are placed in Developmental Education courses). Approximately 50% of all learners are first-generation college students, 52% are female, 48% male, and 75% are between 18-24 years old.

While the institution exists in an EFL context, English is the official medium of the college, and the language of instruction and evaluation in nearly all courses offered. Given this scenario, and with over 90% of all incoming students being placed in the program, there is tremendous pressure on the Developmental English Program (DEP) to prepare students for English-medium credit courses. This also necessitates well-functioning assessments contributing to beneficial placement decisions, as large numbers of misplaced students, and mixed-ability classes, are likely to only add to the challenges faced by learners and instructors alike.

The DEP comprises three, semester-long levels described in course outlines as 'pre-intermediate', 'intermediate', and 'pre-college'. Each level contains two courses: Reading and Writing, and Listening and Speaking. Since 2007, the college has used results from both AC and a locally designed and marked writing sample, equally weighted, to categorise candidates into one of five groups:

i. not currently prepared for any level of DEP or credit English courses,
ii. DEP Level 1,
iii. DEP Level 2,
iv. DEP Level 3, and
v. exempt from DEP and placed directly into introductory credit English courses.

Since its adoption, various stakeholders, particularly DEP instructors, have questioned the suitability of the placement system, believing it resulted in numerous misplaced students and mixed-ability classes. AC, in particular, was often pointed out as being overly difficult for the institution's applicants, and of questionable relevance to DEP courses. However, no validation study had been conducted to provide evidence upon which viable decisions could be made regarding the placement tests or system. The current study thus aimed to fill this gap.

Validation framework

The validation framework developed for the present study – a hybrid of Kane's (1992, 1994; Kane, Crooks, & Cohen, 1999) interpretive model and Bachman's (2005) and Bachman and Palmer's (2010) assessment use argument (AUA) – was decided upon for two main reasons. First, Kane's model was chosen as the basis for the interpretive part of the structure (i.e., test score meaning) because it is probably the most widely known and commonly used framework in educational assessment, serving as the basis for many other influential frameworks, including Chapelle, Enright & Jamieson's (2004, 2008, 2010) investigations into the validity of the Test of English as a Foreign Language (TOEFL), and Bachman's (2005) and Bachman and Palmer's (2010) AUA as well. As Kane's model has been so widely used and discussed, particularly in assessment outside language education, it was thought to facilitate communication amongst stakeholders (such as faculty members and administrators) at the institution, many of whom had only recent and developing experience in the area of educational assessment. Evidence relating to evaluation, generalisability, and extrapolation inferences, for example, seemed readily parcelled and explainable as scoring, reliability (or consistency), and relevance to student learning in institutional courses. However, the AUA was felt to offer greater detail and structure in its consideration of test utilisation and impact, including issues such as sufficiency, equitability, values, and consequences.

The resulting hybrid framework, including all claims, warrants, and rebuttals considered for the current investigation, is presented in Figure 1. Claims 1, 2, and 3, and their associated warrants, relate to Evaluation, Generalisability, and Extrapolation inferences from Kane's interpretive model. Readers will note that the Explanation inference has been left out of the framework. This later inclusion in Kane's model addresses whether test tasks engage the abilities and processes intended by the test designer. The focus of this study, however, is not whether the test assesses the construct intended by the designers. It is far more to do with whether the tasks and skills associated with the test are relevant to the courses into which students are being placed. As such, Kane's Explanation inference was omitted from the current validation framework. Turning to test utilisation, Claims 4 and 5, Decisions and Consequences, and their associated warrants, are derived from Bachman's and Bachman and Palmer's AUA.

Efforts were made to ensure the framework was as comprehensive as possible, and so the model, and its constituent claims, warrants, and rebuttals, as well as the types and sources of evidence, were not purely the design of the researchers. They are the outcome of substantial input and negotiations with

several constituents at the college, including instructors, department chairs, and academic administrators.

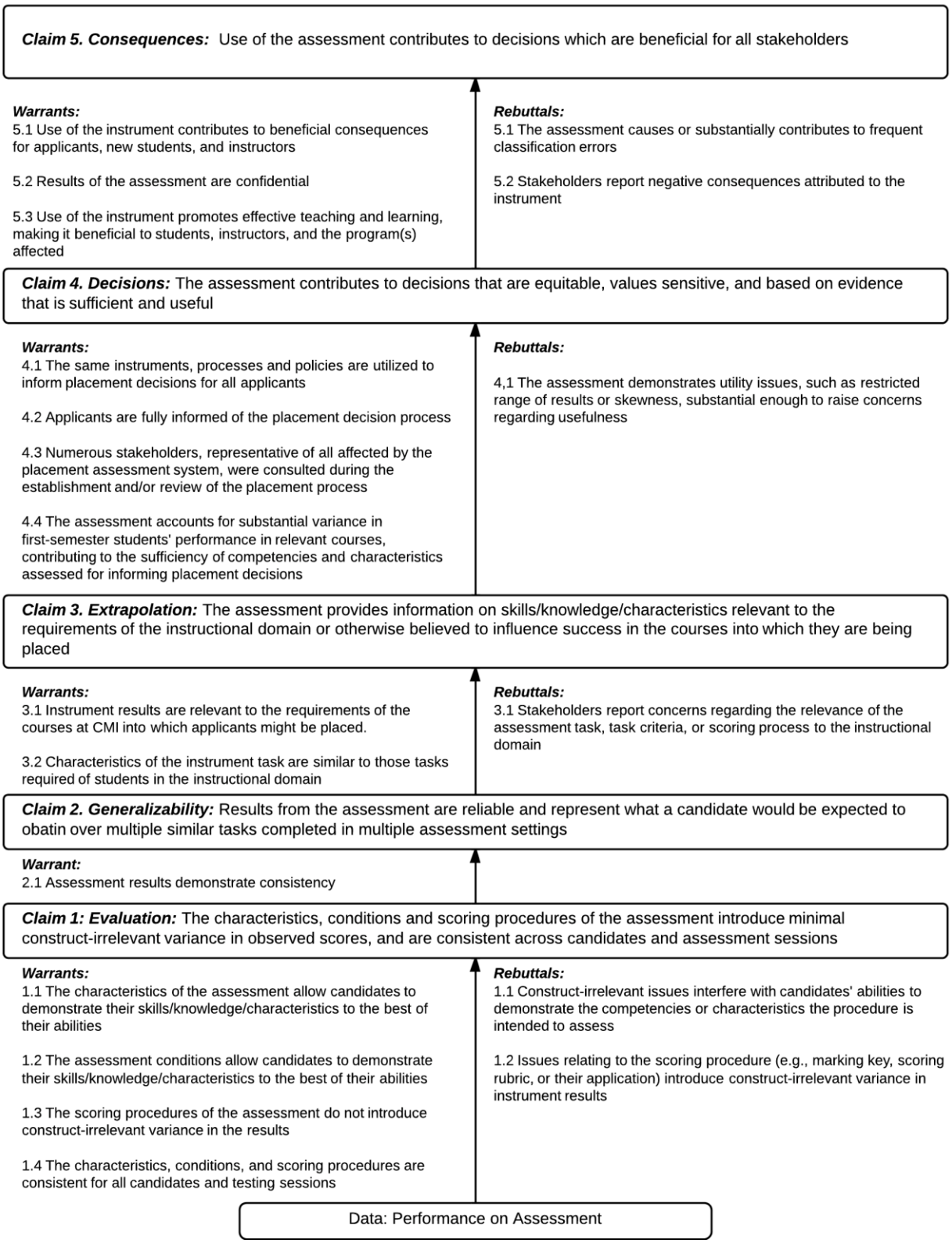

Figure 1. Validation framework.

Method

Participants

Evidence was gathered largely from three different participant groups. The first group comprised all applicants who completed the placement process over three consecutive academic years, regardless of whether they subsequently enrolled at the institution (n=2118). The second group consisted of students who were currently completing first-semester courses. Demographic data for this group was not available. However, the sample sizes (n=144 and n=160 for Reading & Writing and Listening & Speaking courses, respectively) relative to the student population (approximately 850), and the near-homogeneous nature of the student body in a number of important aspects (e.g., 98% Micronesian, English Language Learners), would seem to limit the risk of the sample not being representative of the population.

The third group comprised DEP instructors (n=17) who had taught at the institution for between 1 year to over 10 years. All but one instructor (93%) reported confidence in their familiarity with the learning outcomes of the program's English courses, and most had experience teaching more than one level and more than one course in the DEP. Instructor familiarity with the program, courses, and learning objectives was important to establish if we are to ascribe value to their insights regarding various aspects of the validity argument, such as relevance of test tasks to the target language use (TLU) domain, the courses themselves.

Finally, in order to gain information or clarity regarding institutional policies and practices relevant to the placement test and testing procedures, individuals at the institution involved in various aspects of the process were occasionally consulted, through personal communication.

Measures

Table 1 provides a brief summary of the various data sources used throughout the study.

Table 1. Data sources.

		Data Source (and Sample Size)	Description
Placement test results	1	Aggregate AC results for all applicants (*n*=2118)	Automatically scanned, scored, and compiled by computer, for all candidates
	2	AC results for all participating first-semester students (*n*=304)	
Course results	3	Final English course results (*n*=304)	For all participating first-year students, provided by instructors, as a score out of 100 (not a letter grade)
Questionnaires	4	Applicant questionnaire (*n*=175)	Conducted post-exam, regarding perceptions of placement tests, including AC
	5	First-semester student questionnaire (*n*=90)	Regarding the appropriacy and impact of their placement
	6	English instructor questionnaire regarding placement tests (*n*=14)	Regarding functioning and consequences of placement tests, including AC
	7	English instructor questionnaire regarding student placement (*n*=14)	Soliciting opinions regarding ideal placement for first-semester students in their courses
Interview	8	English instructor focus group interview (*n*=14)	Regarding functioning and consequences of placement tests, including AC
Documents	9	Guidelines from test publishers Policies and procedures of the institution	Current institutional practices regarding the placement assessment process and guidelines, and relevant test publisher documents

Accuplacer results

Accuplacer scores were collected for two groups. First, aggregate results for all candidates over three consecutive academic years (n=2118) were provided by the Registrar's Office. For all candidates (whether admitted or not), total scores for the English section, and both of its subtests, were provided. Second, across the course of three semesters, the placement test results of new students, currently in their first semester at the college, were gathered (n=304).

Course results

These first-semester students' course results (for both Listening and Speaking (n=160) and Reading and Writing (n=144) courses) were also used to investigate the predictive capacities of the placement instruments for student performance in English courses. In order to avoid problematic issues with restricted range of

course outcomes in such estimates (Armstrong, 2000; College Board, 2003), rather than using final letter grades, instructors provided students' course results as a final percentage (i.e., a score out of 100). Additionally, where the amount of information provided by instructors allowed, efforts were made to also produce final course scores without the influence of: points for attendance and/or participation; and points for attendance and/or participation, and any missed assignments. This was done in an attempt to establish results more reflective of student abilities, and less influenced by such issues as time management, motivation, and other factors that the placement instrument was not designed to address.

Questionnaires

Four questionnaires were used during the study. One was completed by applicants to the college (n=175) upon completion of the placement tests. The second was completed by first-semester students (n=90) in the final weeks of their English courses, in order to get their insights on the appropriateness and impact of their placement. Demographic data collected with both questionnaires suggest very close approximation to the student body at the institution, reported earlier, suggesting both participant samples are representative of the target population.

The two remaining questionnaires were completed by 14 of the 17 DEP instructors. The intention of the first questionnaire was to gather insights regarding the relevance of the items and tasks on AC to the skills and knowledge required of students in DEP courses. The second questionnaire, completed towards the end of each semester, asked instructors their opinion as to where each first-semester student in their course should ideally have been placed, based solely on relevant language skills.

Focus group interview

The faculty members who completed the instructor questionnaires also participated in an hour-long, semi-structured focus group interview. The interviews were audio recorded and transcribed. Copies of AC were provided to all participants. The interview sought to gather insights regarding various aspects of the placement instruments, including relevance to the TLU domain.

Documents

A variety of documents from the publishers of AC, College Board, were reviewed for relevant content. These included manuals for test users (College Board, 2003), as well as research on the predictive validity of the instrument

(Mattern & Packman, 2009). Additionally, a number of relevant policy and procedure statements at the institution were also reviewed.

Data analysis procedures

Descriptive statistics and score distributions were established as a means of informing the utility rebuttal of the decisions claim for AC. Kuder-Richardson 21 formula was used to estimate internal consistency. As only the subtest scores – Reading Comprehension and Sentence Skills – and total score for each candidate were available, variance for specific test items could not be determined. As such, typically preferred methods for estimating internal consistency, such as Cronbach's alpha, were not possible. Estimates of common variance (r^2, also sometimes referred to as coefficients of determination) were used to approximate the amount of overlapping variance between variables such as test scores and course results. These were calculated by squaring Pearson correlation (r) results. Finally, Chi-square was used to analyse differences in questionnaire responses.

Results

Evaluation claim

The evaluation claim, in Figure 1, asserts that the characteristics, conditions and scoring procedures introduce minimal construct-irrelevant variance (CIV), and are consistent for all individuals and assessment sessions. Three sources of evidence were considered in order to inform the evaluation warrants. These were relevant research published by the test developers, current institutional policies and procedures, and stakeholder insights regarding the instrument and its administration.

Warrant 1.1: Test characteristics

The publishers of Accuplacer assure users of both the adaptive OnLine version and its paper-based derivative, Companion, that all items included in the instruments have been rigorously investigated for differential performance between examinees both in terms of gender and ethnic background, including 'Asian-Pacific Islanders' (College Board, 2003), and that no items found to be problematic were included in the final versions of the instruments. As differential performance amongst groups may indicate disparities in familiarity of content or other issues not related to the target construct, such findings would normally provide backing for Warrant 1.1. However, these studies were conducted in an ESL context, with students more proficient in English than any other language, whereas the host institution is in an EFL context, serving

predominantly English Language Learners. It is entirely possible, therefore, there may be questions, answer options, or other texts in the instrument that contain language, cultural references, or other presumed background knowledge which may compromise test-takers' abilities to comprehend the question or task and, therefore, neither engage nor assess the competencies intended. This may, instead, support a rebuttal against the evaluation claim, if supported by evidence, such as examinee or instructor opinion.

During the focus group interview, instructors presented what appeared to be a uniform position that 'Accuplacer is too difficult for students' and that the 'level of language and vocabulary… are far too advanced to be accessible to the vast majority of applicants.' Furthermore, there was widespread agreement that 'students accurately placed in Level 1 would not understand very many of the questions of the English subtests.' Most instructors seemed to feel the majority of applicants were probably 'guessing for most of the questions.'

A majority of the instructors (86%) disagreed or strongly disagreed with the statement 'Most applicants to the college will be able to understand the texts (such as the instructions, prompts, questions, etc.) of the instrument.' Contingency table analysis found a Chi-square value, after Yates correction for violation of the assumption of a minimum of 5 participants in all cells (X^2_{Yates}), of 5.786, and effect size of 0.643 (df=1, p<.05). Chi-square (X^2) effect sizes (Φ) range from 0 to 1, with results approximating 0.3 considered moderate and those of 0.5 or above indicating a strong degree of association between the variables in question (Rea & Parker, 2011). As such, instructors can be said to strongly, and significantly, reject the idea that applicants are likely to comprehend the texts of the Accuplacer exam.

A number of instructors, in the open-ended 'comments' section of the questionnaire, identified content presenting potential cultural bias (for example, references to 'King Kong,' and 'the American dream'), and expressed concerns that the language of the test, in general, was far too difficult in for the majority of the institution's students and applicants.

Instructor opinion, then, would seem to rebut Warrant 1.1, suggesting comprehension of the instrument texts could be introducing construct-irrelevant variance.

Turning to test-takers themselves, of the 115 surveyed who expressed a non-neutral opinion (i.e., chose a response other than 'neither agree nor disagree'), the significant majority (65%) (X^2=10.652, Φ=0.304, df=1, p<.01) agreed with the statement 'I understood the AC English test instructions and questions.'

However, 35% of test-takers providing a non-neutral response (25% of all respondents total) stated they did find instructions and questions confusing. This represents a sizable portion of candidates whose scores may have been influenced by a factor not related to the intended construct. As such, the evidence does not support Warrant 1.1.

Warrant 1.2: Test conditions

Institutional policy states there is to be no time limit for the completion of placement assessments. As time constraints are likely to influence test-taker abilities to demonstrate relevant skill(s), this policy would seem to support the warrant of minimal CIV introduced by testing conditions. However, as faculty and other stakeholders identified timing as potentially problematic, an item was included in the test-takers' questionnaire about whether they had sufficient time to 'carefully read and answer all of the questions'. Here again, while the significant majority of the 133 test-takers providing a non-neutral response agreed with the statement (63%)(X^2=9.211, Φ=0.263, df=1, p<.01), a significant number (37%) reported they did not have enough time. This finding was corroborated by additional comments on the questionnaire, such as 'Give more time for students to take the test,' and 'add much more time.'

The evidence, then, would seem to suggest some form of time limit has in fact been imposed, at least from the perception of a substantial number of examinees, and thus does not back Warrant 1.2.

Warrant 1.3: Scoring Procedures

As the institution employs automated scanning, marking, data entry, and data processing (including computing placement recommendations), there would seem little opportunity for inconsistencies in scoring procedures to introduce CIV, barring perhaps, errors in the marking keys or some other aspect of the process.

Publications from the test developers (College Board, 2003) assure AC users that the items, answers, and answer options are carefully created and checked by experts in the field of entry-level credit and remedial college English. While reports of errors in the answer keys are not entirely unknown (CCCAA, 2007), they would appear to be quite rare. Further, no instructor reviewing the instrument as part of the focus group interview process reported finding problems with any item, such as more than one, or no, best possible answer, for example.

Presuming no errors in the scoring key provided by the publishers, the evidence supports the warrant that scoring procedures do not introduce CIV.

Warrant 1.4: Consistency across test-takers and sessions

The existence of an established institutional protocol that all placement test proctors are to follow was considered evidence in support of the consistency of test administration for all examinees and across testing sessions. Further, primary responsibilities for test proctoring had been held by the same two staff members since the adoption of the current placement assessment system and instruments. This, it could be argued, gives further likelihood of consistency than if these responsibilities for oversight of the testing sessions rotated amongst several different individuals. Additionally, during the faculty focus group interview, instructors who had served as supplemental proctors during testing sessions reported the perception these procedures are followed consistently across testing sessions and locations.

While conditions across testing sites is an issue that needs to be investigated in future, the evidence considered here, along with the automated scoring processes for AC, supports the warrant for consistency across test-takers and sessions.

Generalisability claim

As indices of reliability offer insight into the apparent consistency of scores across samples of observations, they provide evidence relevant to the generalisability claim (Kane et al., 1999). The estimate of internal consistency, via the KR-21 formula, was 0.76 for the total AC score. Given the known overly conservative nature of KR-21 (Brown, 2005), this result was deemed sufficiently close to the traditional criteria of 0.80, and thus held to support the generalisability claim.

Extrapolation claim

The extrapolation claim asserts that the instrument provides evidence regarding candidate competencies (and/or other characteristics) relevant to the tasks required of students in the instructional domain, or otherwise believed to influence student success in the courses into which they are being placed.

Warrant 3.1: Relevance to the instructional domain

Given that the English language courses are the Target Language Use (TLU) domain, substantial overlap between instrument results and the outcomes of these courses would be powerful evidence that the competencies assessed by the test are relevant to those required for student success. Table 2 reports common variance between AC results and DEP course outcomes.

Table 2. Coefficients of determination between AC scores and final course results

Course	Final Result		Level 1	Level 2	Level 3
Listening & Speaking	Unadjusted	r^2	0.194**	0.035	0.088
		N	93	41	23
	Adjusted 1	r^2	0.158**	0.037	0.07
		N	89	41	22
	Adjusted 2	r^2	0.246**	0.044	0.018
		N	84	41	22
Reading & Writing	Unadjusted	r^2	0.040*	0.181	0.386**
		N	100	20	24
	Adjusted 1	r^2	0.159*	--	0.329**
		N	29	0	24
	Adjusted 2	r^2	--	--	0.353**
		N	0	0	24

*significant at .05 level; **significant at .01 level
-- no results due to lack of participants for this cell
Adjusted 1 - final course results with any influence of scores for attendance or participation removed
Adjusted 2 - same as Adjusted 1, but with any influence of missed assessments also removed

As the instrument is designed to assess the reading and sentence-related skills of those for whom English is a 'best language', we might not be surprised to find AC scores most overlap final course results in the most advanced (Level 3) Reading and Writing (RW) course of the program. Coefficients of determination (r^2=0.33 to 0.39, p<.01) are higher than the 0.05 to 0.22 typically reported in other predictive validity studies at colleges in the US (Mattern & Packman, 2009). Perhaps unsurprisingly, given the instrument does not measure oral/aural skills, Level 3 Listening and Speaking (LS) course results were not predicted to any significant extent by AC scores.

Results from the instrument showed insignificant common variance with Level 2 LS or RW course results. Again, because the instrument is designed for use with native or near-native speakers, we might not expect it to predict final results for courses addressing 'intermediate' or 'pre-intermediate' English language learner skills. Perhaps oddly, then, AC results did show significant, and somewhat substantial, predictive capacity for not only Level 1 RW course outcomes (r^2=0.04 to 0.16, p<.05), but also LS course results (r^2=0.16 to 0.24, p<.05). It is unclear why such a pattern in common variance with course outcomes would occur.

Overall, however, the findings would seem to rebut the extrapolation, given the instrument demonstrated significant common variance for the outcomes of only three of the six DEP courses into which new students are placed.

Warrant 3.2: Instrument task(s) are similar to the instructional domain

During the focus group interview, instructors (all of whom had just reviewed AC, and had a copy available for reference) appeared unanimous in the opinion that the instrument tasks are generally dissimilar to the objectives and requirements of DEP courses. More specifically, instructors felt the instrument addressed 'parts of language, not whole language,' and required critical thinking and language skills often well beyond what is expected of students in DEP classes, and which the institution's 'credit level students would struggle with.'

On follow-up questionnaires, the majority of instructors (75%) disagreed with the statement that AC 'ask[s] students to do the same sorts of things they will be expected to do in their classes at [the college],' though this outcome was not found to be statistically significant (X^2_{Yates}=2.083, Φ=0.417, *df*=1, *p*=0.149).

In the 'comments' section of the questionnaire, one instructor repeated the concern raised in the focus group that the instrument addresses 'parts of language, not whole language.' Another felt 'most of the test is comprised of subtleties that we would expect to distinguish between native English speakers,' but which had little relevance to DEP or even entry-level credit English courses. None of the comments offered positive aspects of the instrument in relation to the extrapolation claim.

While the questionnaire item results were not statistically significant, the bulk of instructor responses to the item, comments offered on the questionnaire, and opinions expressed in the focus group interview expressed doubts as to the relevance of AC tasks to those required of students in English courses. As such, the evidence cannot be said to support the extrapolation claim.

Decisions claim

The decisions claim asserts that placement decisions are equitable, values sensitive, and based on evidence that is sufficient and useful. Four warrants and related rebuttals were addressed.

Warrant 4.1: Equitability

All 2,120 candidates for whom data was available would appear to have been excluded, placed, or exempted from the DEP based on placement instruments

outcomes and relevant policies of the college. Further, current institutional policy clearly states all applicants are to complete the same placement instruments, and be placed by the same placement assessment system decision-making procedures. As such, the evidence supports the equitability warrant.

Warrant 4.2: Full disclosure

While it might seem obvious that applicants to the college would know the purpose of the placement test(s) they are required to sit, there would appear to be no standing policy at the institution regarding informing examinees of how their results are used, how placement decisions are made, and what the potential outcomes might be. According to instructors who participated in the study, and to Student Services staff members consulted informally, test-takers are not made aware of this information at the testing sessions or at other times or through other means. Nor are examinees aware of the relative weighting of AC and the writing sample which also informs placement decisions, use of cut scores, or other aspects of the placement decision process.

To the understanding of both faculty and staff, the only information most test-takers receive is a final placement decision and a date to come to the school to register. As such, the warrant of full disclosure is not supported by the evidence examined.

Warrant 4.3: Stakeholder input

Numerous stakeholders, representative of all affected by the placement system, are to be consulted during the establishment and/or review of the placement process, including the selection of its constituent assessments (Bachman & Palmer, 2010). According to instructors in the focus group interview, and other stakeholders present at the time AC was adopted (such as Student Services staff, the former head of Institutional Research, academic administrators, and the former chair of the DEP, all consulted informally), the decision was made largely by executive administrators, and the establishment of cut scores and other implementational procedures were carried out primarily by the Institutional Research (IR) department. According to institutional documents, and members of IR consulted, these decisions were made largely with issues of comparability of results with other US-accredited institutions, and assurance of student eligibility for US educational grants, in mind. Little to no consultation with other stakeholders, such as academic administrators, faculty members, or students, would seem to have occurred.

Warrant 4.4: Sufficiency

As we are currently interested in AC alone (and not the functioning of its results in combination with writing sample outcomes), the question here becomes one of whether the instrument makes a substantial contribution to the sufficiency of the abilities assessed to inform beneficial placement decisions. As reported earlier, when considering Warrant 3.1 (relevance to the TLU domain), instrument scores were found to demonstrate significant common variance with the outcomes of only three of the six DEP courses into which new students are placed. As such, it cannot be said to contribute to the sufficiency of skills, knowledge, and other characteristics considered in the placement assessment system necessary to result in beneficial decisions.

Rebuttal 4.1: Utility issues

This rebuttal addresses the possibility that a placement instrument demonstrates utility issues that could raise concerns regarding its usefulness. Two sources of information were analyzed: score frequency distributions, and a review of the cut scores used to differentiate students into various ability categories.

Figure 2 shows the distribution of results for the 2,117 candidates who completed AC over the course of the study. Mean (20) and median (19) results are very low for a test with 70 total items. The estimate of skewness (1.281), relative to the standard error of skewness (0.053), indicates the distribution is significantly, positively, skewed. Skewness alone, however, does not establish whether or not utility is necessarily threatened. Further insight was sought from the cut scores, presented in Table 3, established by the institution in order to separate candidates into placement categories.

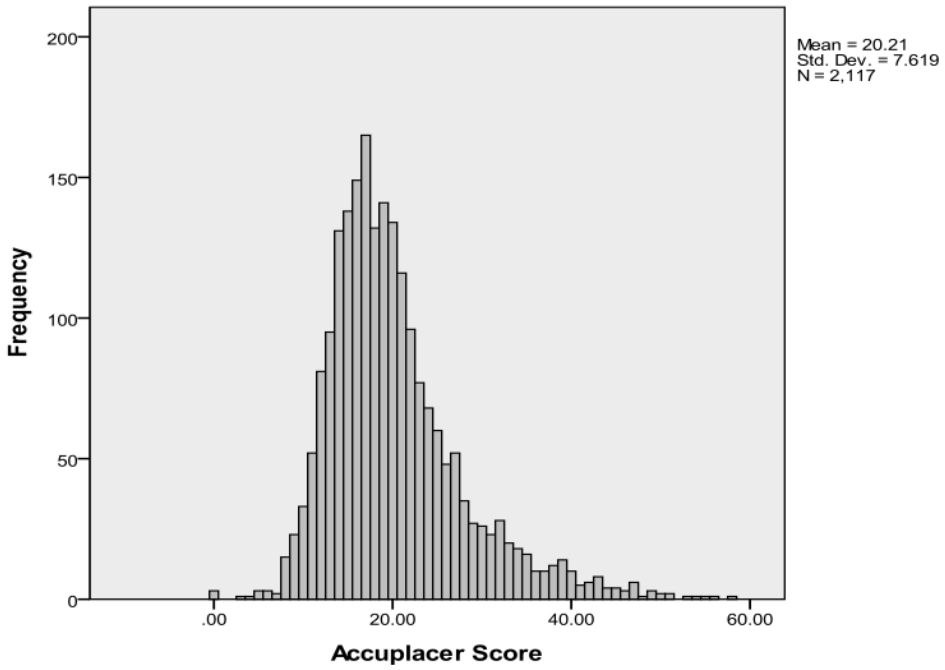

Figure 2. Frequency distribution of AC scores from Fall 2008 to Fall 2011 semesters

While the cut-scores reported are those used by the college since the adoption of AC, they are not the original cut scores developed and intended for use. A report from the first semester of the instrument's implementation details an immediate change in cut-scores, from those initially established by Institutional Research, to the current values. While the report does not indicate the original cut-scores, it does state the need for lowering them, as very few applicants qualified for enrollment in any English courses if the original ranges were to be used. As such, the new cut-scores were established in order to both admit sufficient numbers of the applicants to the school for that semester, to avoid a substantial drop in enrolment numbers, and to ensure at least some new students were placed in Levels 2 and 3 of the DEP.

Table 3. AC cut scores.

Cut Score Range	Placement Recommendation
43-70	Credit English
37-42	DEP Level 3
30-37	DEP Level 2
15-29	DEP Level 1
0-14	Not currently prepared for any English course at the institution

From Table 3, we also see that some placement categories are associated with very small score arrays. Ranges for Levels 2 and 3 are only seven and five points wide, respectively. Given that the standard error measurement for the

instrument is 3.80, these are probably dangerously restricted ranges upon which to base high-stakes decisions about candidates, and ones which may result in a substantial number of placement errors. This evidence would seem to support Rebuttal 4.1.

Consequences claim

Perhaps the most important claim of all is that the placement system, its constituent instruments, and the decisions informed by them, result in beneficial consequences for all affected.

Warrant 5.1: Beneficial consequences for individual stakeholders

Consequences for applicants

Table 4 summarises instructor and applicant (test-taker) responses to questionnaire items intended to gather opinion regarding AC's impact on examinees. Results indicate that instructors are unanimous or nearly unanimous in their opinion that AC is likely to negatively impact test-takers' perceptions of their English language abilities and their desire to pursue a tertiary education. Looking to the responses of the examinees themselves, however, the majority does not report experiencing negative effects with regard to perceptions of their language abilities, likelihood of being successful at the college, or desire to pursue a higher education. As the warrant pertains to examinee experience, firsthand feedback was felt to take precedence over instructor perceptions, and thus the evidence would seem to support Warrant 5.1.

Table 4. *Stakeholder opinions regarding AC's impact on examinees*

	Questionnaire Item	Group	*n*	Prop.	X^2	Φ	*df*	*p*
Instructors	The AC English subtests will have a positive impact on students' perceptions of themselves and their English language skills.	Agree[a]	0	0	--	--	--	--
		Disagree[b]	10	1.00				
		Total	10	1.00				
	The AC English subtests will have a negative impact on students' desire to pursue a postsecondary education at CMI or another institution.	Agree[a]	10	.91	5.82[c]	.727	1	.016
		Disagree[b]	1	.09				
		Total	11	1.00				
Test-takers	Taking the AC English test made me think I can be a successful student at CMI.	Agree[a]	94	.77	35.70	.541	1	.000
		Disagree[b]	28	.23				
		Total	122	1.00				
	Taking the AC English test made me feel good about my English abilities.	Agree[a]	93	.76	32.27	.512	1	.000
		Disagree[b]	30	.24				
		Total	123	1.00				
	Taking the AC English test made me want to study at CMI.	Agree[a]	121	.86	72.35	.716	1	.000
		Disagree[b]	20	.14				
		Total	141	1.00				

a Combined 'Agree' and 'Strongly Agree' responses
b Combined 'Disagree' and 'Strongly Disagree' responses
c X^2 with Yates correction as one cell violates assumption of minimum 5 participants
-- no result due to lack of participants in one cell

Consequences for new students

Evidence considered for this aspect of the warrant included student performance in the courses into which they were placed, and first-semester student opinion (solicited via questionnaire) as to the accuracy and impact of the placement decision. Much of the evidence was troubling. For example, 37% of new students did not pass the English courses into which they were placed, making it the most common outcome, and 15% of first-semester students reported being placed in a level too difficult for them.

However, as final placement decisions are the result of AC results, writing sample results, respective cut scores for the two instruments, and decision-making policies of the placement system, we must remember the evidence considered here does not reflect the functioning of AC alone.

Consequences for instructors

Two sources of evidence were considered for evaluating this aspect of the framework: instructors' responses to the questionnaire soliciting their opinion

as to where, ideally, each of their first-semester students should have been placed; and opinions offered in the focus group interview.

According to questionnaire results, instructors identified 43% of new students in Listening & Speaking (LS) courses, and 32% in Reading & Writing (RW) as being in the wrong level for their abilities. Further, they felt 15% of students in LS and 14% in RW were in courses they did not have the language abilities to pass, regardless of time and effort dedicated to the course. This closely matched the 15% of new students who self-identified as being in a level that was too difficult.

During the focus group with faculty, many expressed the opinion that frequent student misplacement is at least partly to blame for the low success rates of many students, as many are over- or underwhelmed, and that mixed-ability classes resulting from the placement errors made teaching and learning more difficult in their courses. Some described the beginning of each semester a 'scramble' to try to identify and re-place students in the wrong classes for their ability levels while course changes could still be made at the college.

As with the opinions and performance of first-semester students, though, these outcomes reflect the functioning of the overall placement system, and not AC alone. However, as we see in Table 5, when asked specifically about AC, faculty members expressed the widely held opinion that the test is not useful for informing placement decisions at the college. Further, it was quite clear during the focus group interview that much of the frustration instructors felt was focused towards Accuplacer, with most holding the writing sample as the likely source of any useful placement information.

Table 5. Instructor opinion regarding the usefulness of AC.

Questionnaire Item	Group	*n*	Observed prop.	X^2	Φ	*df*	*p*
The AC English subtests are useful for choosing which applicants are able to enroll in English courses at CMI.	Agree[a]	1	.09	5.82❖	.727	1	.016
	Disagree[b]	10	.91				
	Total	11	1.00				
The AC English subtests are useful for placing incoming students in the Developmental or Credit level English classes best suited for their current language abilities.	Agree[a]	2	.18	3.27❖	.545	1	.070
	Disagree[b]	9	.82				
	Total	11	1.00				

a Combined 'Agree' and 'Strongly Agree' responses

b Combined 'Disagree' and 'Strongly Disagree' responses

❖ X^2 with Yates correction as one cell violates assumption of minimum 5 participants

Overall, particularly when viewed in combination with the previously reported views of faculty that AC tasks are not relevant to the TLU domain, results indicate that instructors view AC as negatively affecting themselves by misplacing several new students each semester.

Warrant 5.2: Confidentiality of results

Institutional policy establishes that placement instrument results are confidential and available only to the examinee. The lone exception to this rule occurs if instructors have a new student in their course that they believe has been misplaced. With the permission of the DEP chair, they may be allowed to review the placement assessments of the student. If the instructor feels a change is in the best interests of the student, the student must be consulted and agree to the move. If this happens, then placement materials are reviewed by the department chair and the instructor into whose class the student would transfer, as their consent is also required. Given this policy, the confidentiality warrant would seem to be supported.

Warrant 5.3: Promotion of effective teaching and learning

Instructors, during the focus group interview, complained of mixed-ability classes, 'scrambles' at the beginning of every semester to identify and move misplaced students, and a number of first-semester students being either under-challenged or, worse, having little chance of success. Questionnaire responses and focus group comments clearly indicate that faculty members perceive Accuplacer as the main problem.

In summary, there is both qualitative and quantitative evidence to suggest the use of AC is not positively impacting on teaching and learning in the courses into which students are being placed.

Discussion

Figure 3 provides a summary of the validation framework, restated in light of the evidence and the main findings informing the warrants and rebuttals relating to each claim.

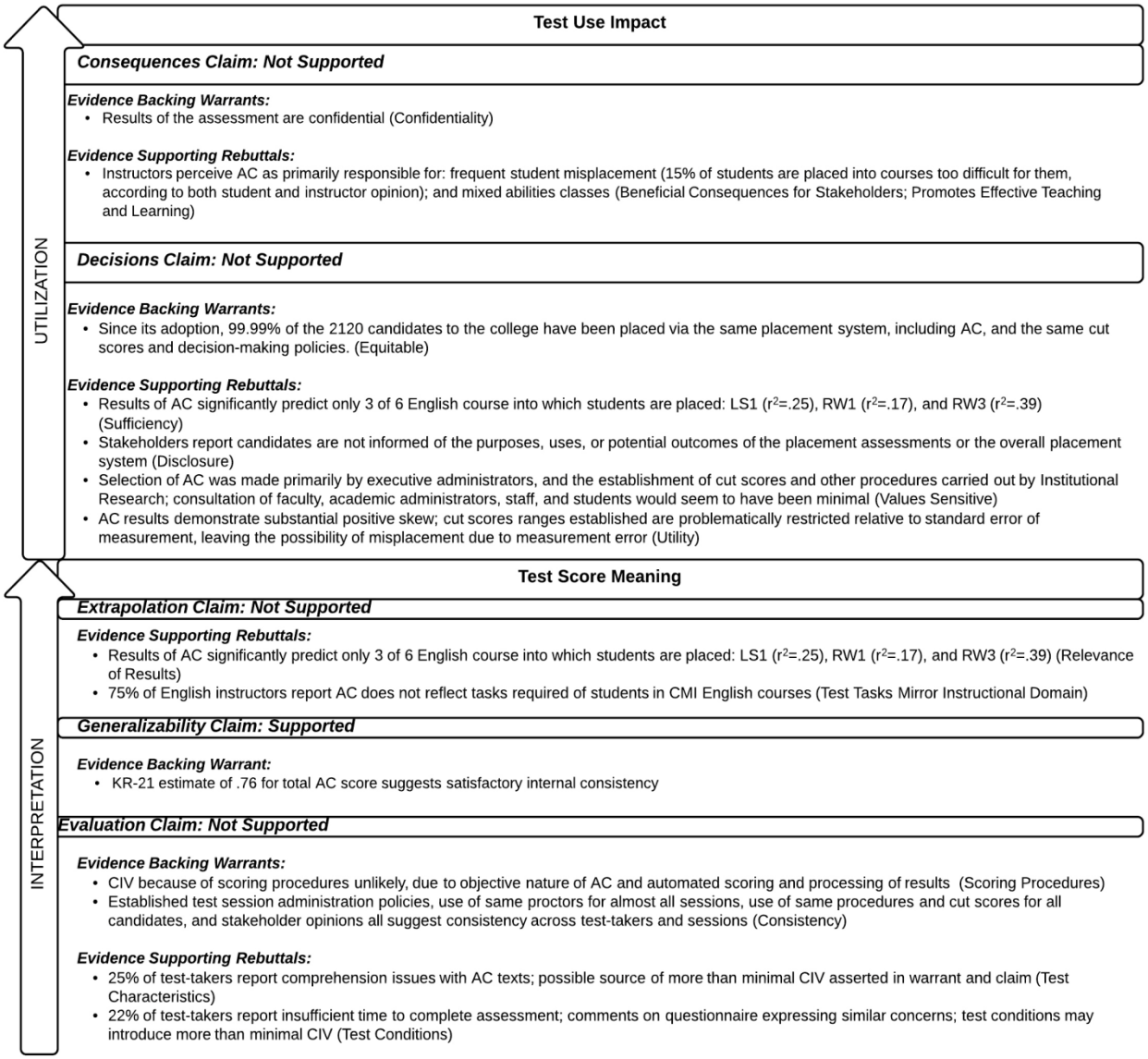

Figure 3. Validation framework restated in light of evidence

With only one claim – generalisability – supported by the evidence, results do not bode well for the validity of AC, as employed at the host institution. Despite the reliability of the instrument and the consistency in its administration and scoring, the number of test-takers reporting difficulties understanding its texts (25%), and expressing concerns they did not have enough time to do their best (22%), suggests a level of CIV in observed scores too considerable to support the evaluation claim.

With regard to the extrapolation claim, the instrument did not demonstrate significant overlap of variance with student performance in half of the English courses into which it is used to place students. This would appear to confirm

the widely held perception reported by instructors, that the tasks of the instrument are not relevant to the TLU domain (English courses).

Overall, then, concerns regarding substantial possible CIV and lack of relevance of the abilities assessed by the instrument to those required by DEP courses, suggest the meaning of the test scores produced by AC are not well-suited to the context and needs of the institution.

Turning to test utilisation, other than equitability – all applicants must take AC and be placed by the same cut scores and other placement-decision mechanisms – none of the evidence considered suggested AC contributes to decisions that are values sensitive and based on evidence that is sufficient and useful. Evidence relating to the decisions claim, such as descriptive statistics, estimates of skewness, and a review of the cut scores for the instrument, suggest the test is too difficult for the target test-takers. In addition, cut score ranges (as small as 5 points wide, for example) are far too close to standard error estimates (3.80) for placement errors to be unlikely.

For the consequences claim, the requirement of confidentiality would appear to have been met. However, poor success rates for first-semester students (37% do not pass the English course into which they are placed), and questionnaire responses from both instructors and first-semester students alike suggesting 15% of students are placed into courses too difficult for their abilities, suggest significant problems. While these results do not reflect the functioning and impact of AC alone (writing sample results and cut-scores also influence placement decisions), instructor opinion clearly indicates they perceive AC as the primary problem.

Conclusion

The validation of AC for the local context resulted in a number of valuable outcomes for the host institution. After years of disagreement, the evidence collected led to an immediate resolution amongst faculty, staff, and administration, that an alternative instrument better suited to the institution's applicants, students, and learning outcomes was needed.

The framework used for the study proved of further value in appraising potential replacements, and helped find an instrument with stronger initial performance in key areas such as common variance with DEP courses, relevance to the TLU domain (according to instructors and students alike), and a range of scores that, unlike AC, did not implicate utility issues.

Additionally, the investigation helped to identify placement testing policy- and procedure-related strengths and weaknesses. For example, consistency in testing sessions and test administrators was felt to be a strength that should be continued, as was the policy of not having a time limit for the test. However, a substantial number of test-takers reporting not having enough time to do their best on the test was troubling, and solutions such as making sure the absence of a time limit was clearly explained in the examinees' first language were suggested.

Importantly, because of these benefits, the institution and its constituents agreed to continue validation efforts for all placement assessments, as well as other high-stakes instruments used at the college. Most important of all, it is hoped that these changes, and the ongoing evaluation efforts, will contribute to improved beneficial outcomes at the institution, especially those that impact teaching and learning, and student success.

At a broader level, this study may offer important insights to the language and educational testing community. It has offered a viable argument-based framework from which others may work, and hopefully contributes to diminishing the longstanding void of validation efforts, especially amongst test users. Results of this investigation would also seem to support the position of Hughes and Scott-Clayton (2011), presented earlier, that use of a standardised instrument not validated for the unique students, needs, and learning outcomes of the programs and institutions using them, are probably contributing to the ongoing retention and learning problems reported at these organisations.

Finally, with increasing demands for accountability, and more prevalent use of tests to inform decisions about individuals, programs, schools, and entire education systems, the need for addressing questions of how, why, and when assessments are used, the consequences of their use, and the ethical obligations of test developers and users, has never been greater. This study demonstrates the critical role validation, particularly by local test users, utilising argument-based frameworks, has to play in addressing these questions, and assuring ethical, meaningful, and beneficial assessment in education.

References

ACCJC & WASC. (2010). *Guide to evaluating institutions*. Accrediting Commission for Community and Junior Colleges & Western Association of Schools and Colleges. Retrieved from http://www.accjc.org/wp-content/uploads/2010/09/Addendum-to-Std-IIID_Guide-to-Evaluating-Institutions-w-Cover.pdf.

AERA, APA & NCME. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

AERA, APA & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Armstrong, W. B. (2000). The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community College Journal of Research and Practice*, *24*(8), 681-695.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford University Press, USA.

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community college. *New Directions for Community Colleges*, *145*, 11-30.

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*(2), 255-270.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

CCCAA. (2007). *California Community Colleges Assessment Association Test-development feasibility project.* Sacramento, CA: California Community Colleges Assessment Association.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2004). *Issues in developing a TOEFL validity argument*. Paper presented at the 26th Annual Language Testing Research Colloquium, Temecula, CA.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

College Board. (2003). *Accuplacer OnLine technical manual.* College Board. Retrieved from http://professionals.collegeboard.com/profdownload/accuplacer-program-manual.pdf.

Hughes, K. L., & Scott-Clayton, J. (2011). *Assessing Developmental Assessment in Community Colleges: A Review of the Literature* (No. 19). Community College Research Center: Teachers College, Columbia University.

Joint Committee on Testing Practices. (2015). *Code of fair testing practices in education.* Retreived from http://www.apa.org/science/programs/testing/fair-code.aspx

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425–461.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Kunnan, A. J. (2003). Test fairness. In M. Milanovic & C. Weir, (Eds.), *Select Papers from the European Year of Languages Conference, Barcelona* [pp. 27-48]. Cambridge: CUP.

Mattern, K. D., & Packman, S. (2009). *Predictive validity of ACCUPLACER scores for course placement: A meta-analysis* (College Board Research Report No. 2009-2). Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2009-2-predictive-validity-accuplacer-scores-course-placement.pdf.

Messick, S. J. (1989). Validity. In Linn, R. L. (Ed.), *Educational measurement* (3rd ed.). NY: Macmillan.

Martorell, P., & McFarlin, I. (2011). Help or hindrance? The effects of college remediation on academic and labor market outcomes. *Review of Economics and Statistics*, *93*(2), 436-454.

Offenstein, J., & Shulock, N. (2011). Political and policy barriers to Basic Skills Education in the California Community Colleges. *American Behavioral Scientist*, *55*(2), 160 -172.

Rea, L. M., Parker, R. A. (2013). *Designing and conducting survey research: A comprehensive guide.* NY: Wiley & Sons.

Sullivan, P. (2008). An analysis of the National TYCA Research Initiative Survey, Section II: Assessment practices in two-year college English programs. *Teaching English in the Two-Year College*, *36*(1), 7-26.

Xi, X. (2008). Methods of test validation. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (pp. 2316-2335). Boston, MA: Springer.