Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context

By

Robert Charles Johnson

B.A. (hons.), Simon Fraser University, Canada, 1995

M.A., University of Birmingham, UK, 2001

Research carried out at the College of the Marshall Islands.

Submitted December 20, 2011 to the Department of Linguistics, Division of Linguistics and Psychology, Macquarie University.

This thesis is presented for the degree of Doctor of Philosophy in Applied Linguistics.

To my sister, Kathy. We love and miss you.

Table of Contents

APPENDICES

Abstract

Use of a single, standardised instrument to make high-stakes decisions about test-takers is pervasive in higher education. Contrary to longstanding best practices encouraged by researchers, professional organizations, test publishers, and many accrediting bodies, however, few, if any such institutions have endeavoured to meaningfully validate the instrument(s) they use for their specific context and purposes. The current study attempted to address this void by developing and applying an argument-based validation framework for two widely adopted placement assessment methods – a standardised placement test (Accuplacer), and a locally developed and marked writing sample –utilised by a 2-year higher education institution in the Pacific.

A hybrid of two validation structures – Kane's interpretive model and Bachman's assessment use argument – was implemented in order to assure a balanced focus on both test score interpretation and test utilization. Various types and sources of evidence informed the study, including instrument outcomes, student course results, institutional practices and policies, test publisher data, and the opinions of stakeholders gathered via focus group interview and questionnaires.

Results are argued to provide insights regarding a number of current issues in the literature, including: i) debates regarding the relative strengths and weaknesses of standardised tests and locally developed and marked writing samples for informing placement decisions; ii) the value of locally conducted validation efforts to evaluate the performance and impact of an institution's chosen assessment instruments, and identifying opportunities for improvement; and iii) the need for further argument-based validation studies, particularly those which attend to both test score interpretation and the long-neglected area of test utilization, to be carried out wherever assessments are used to make decisions which impact stakeholders.

STATEMENT OF CANDIDATE

I certify that the work in this thesis entitled "Assessing the Assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context" has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university of institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

The research presented in this thesis was approved by Macquarie University Ethics Review Committee, reference number: HE31JUL2009-D00048 on September 11, 2009.

Robert Charles Johnson (40783995)

December 15, 2011

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of a number of people. My wife, Susan, has made many sacrifices, including having to share her husband with this dissertation. I cannot thank her enough for her support, understanding, and patience.

For the rest of my family, your encouragement was a tremendous help, as was your understanding when I needed to be unavailable.

The guidance, support, and encouragement of my dissertation supervisor, Mehdi Riazi, was essential. I am grateful for his knowledge, cheer, and knack for knowing when I needed reassurance, and when I needed a gentle push back to reality.

Finally, I must acknowledge the help and inspiration I received from my friends, colleagues, and students in the many places I have lived and taught throughout the time I worked on this dissertation, and the coursework which came before it. Shukron. Kommol tata. Xie xie. Terima kasi.

CHAPTER ONE

Introduction to the Study

1.0 Introduction

This chapter will familiarize readers with the fundamental elements of the study, beginning with the central problem the investigation seeks to address, introduction of the context (the College of the Marshall Islands, henceforth CMI) in which the study was conducted, and the tests used to inform placement (and, effectively, admissions) decisions for incoming students to CMI. The importance of the investigation, for local stakeholders, test users elsewhere, and potential insights it may offer to the field of testing and validity studies, will then be presented. Having established the purpose and context of the study, we will then move on to explicitly stating the specific research questions. The chapter concludes with an account of the known limitations of the investigation.

1.1 Problems addressed by the study

The local impetus for the study came from longstanding concerns, expressed at various times by faculty, staff and academic administrators, regarding the suitability and performance of the assessments used to place, exclude, or exempt incoming students in the Developmental English Program at CMI.

Since 2007, the college has used results from a combination of a standardised, multiple-choice instrument — Accuplacer Companion — and a locally designed and marked writing sample (both of which are addressed further later in this chapter) to categorize candidates into one of five groups:

i. not currently prepared for any level of CMI English courses (and, therefore, not currently eligible to enroll at the college);
ii. Developmental English Level 1;
iii. Developmental English Level 2;
iv. Developmental English Level 3; or

v. exempt from Developmental English and placed directly into credit English courses

It is important to note that while one of the categories renders the applicant ineligible for registering at the college, the official position of the institution is that the placement system does not result in admissions decisions. While some stakeholders have argued a placement decision that renders the applicant ineligible for enrolment is effectively the same thing, the counter-argument is that the decision is not permanent, and any applicant unhappy with their placement outcome may re-take the tests at any future testing session, and can enroll at any time once they have demonstrated the English ability considered necessary to be successful in one of the educational programs currently available at CMI.

Since its inception, the current placement system has been a source of concern amongst various stakeholders. The use of Accuplacer, in particular, has been a contentious issue, with many expressing worries regarding its suitability for CMI applicants and relevance to the institution's courses and programs.

In self-study reports to its accreditors, the College of the Marshall Islands (2008, 2009) identified low pass rates and low retention of new students as major issues the institution was working hard to address. In the Fall 2009 semester, for example, final grade distributions in Developmental Reading and Writing courses and credit level Composition courses indicated only 46% of new students achieved a grade of C+ or higher, and some 37% did not pass, making a failing grade the most common result for new students in these courses. These results did not include students who ‘dropped' the course or were withdrawn due to lack of attendance, which are also identified as problems at the college. Recently, it was suggested these low indicators of student success might be at least partly due to issues with the current placement assessment system (PAS), as many felt it was misplacing a large number of new students, creating classes with mixed abilities that made instruction and learning more difficult, and perhaps resulted in large numbers of students being in courses in which they were either academically overwhelmed or in which they were underwhelmed and perhaps bored or frustrated. Until the current study, however, little empirical evidence had been gathered to corroborate or refute suggestions that the placement instruments were problematic.

Additionally, the study was intended to address another issue: assuring that the college was meeting its ethical and professional obligations to its constituents and its accreditors. Best practices in educational testing recommended by researchers (for example, Messick, 1989; Kane, 1992), test publishers (College Board, 2003), required by ethical and professional codes of conduct in educational assessment (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985, 1999; Joint Committee on Testing Practices, 2005) and expected in the standards of CMI's US accreditors (Accrediting Commission for Community and Junior Colleges & Western Association of Schools and Colleges, 2010) all clearly call for the ongoing investigation of any assessment used to make important decisions about students and/or applicants to the college. Like many of its peers, however, until this study CMI had never undertaken such an investigation, let alone had an ongoing program of “assessing the assessments” (Hughes & Scott-Clayton, 2011). This study, then, also represents an initial attempt at creating such a program, establishing an initial framework and focus for the ongoing validation of the instruments currently comprising CMI's placement assessment system.

1.2 Importance of the study

This study was felt to be of importance not only for the stakeholders of CMI, but test users elsewhere, particularly at institutions using standardised tests and/or local performance assessments to inform placement decisions regarding incoming students.

Additionally, it is suggested the study may offer insights of significance to the ongoing development of validity studies and the practice of argument-based test validation.

1.2.1 Importance of the study for CMI

The investigation was considered extremely important for stakeholders at CMI for a number of reasons. First, results from Accuplacer Companion (AC) and the writing sample (WS) are, other than the prerequisite possession of a secondary school or General Educational Development (GED) diploma), the only sources of information used to exempt, exclude, or place students within the Developmental English program at the college. Placement exams are already considered high-stakes because they impact applicants' initial paths towards a tertiary education, and often their referral to extracurricular sources of support (such as extra tuition, mentors, academic skills programs, and so on) (Hughes & Scott-Clayton, 2011). Adding the potential outcome of ineligibility for enrolment, even if only temporarily, only raises the stakes, and makes the ongoing evaluation of the suitability and performance of the instruments informing this decision that much more important.

Placement decisions also affect teaching and learning at the college, and therefore impact all parties — students, instructors, programs, and the institution itself — and their abilities to meet their educational goals. If students are frequently misplaced, for example, the result can be classes of mixed abilities, making it more difficult for instructors to tailor and execute effective lesson plans addressing the same or similar needs for an entire class of students. It can also mean many students in courses for which they are under- or over-prepared, and possibly result in high levels of frustration, disinterest, attendance and retention issues, and disappointing academic results (Bailey, 2009). This, in turn, can impact students in terms of valuable resources, such as time and finances. Students placed in courses that overwhelm them, and who struggle academically as a result, for example, can, possibly even after their first semester at CMI, be placed on academic probation or even suspension. Such circumstances come with the risk of not only delayed studies for a semester or more, but may also impact eligibility for financial aid.

It was hoped, then, that the study would result in greater understanding of the suitability and performance of the current placement instruments. Through insights into the relative strengths and weaknesses of the composite placement assessment system at the college, it was felt that avenues for improving the decisions and outcomes of the placement process could be identified and pursued. Additionally, it was hoped that the current study would be an initial template for ongoing, future evaluations of the placement system and its component instruments, hopefully leading to a program of perpetual investigation, analysis and improvement for all constituents.

1.2.2 Importance of the study for test users elsewhere

CMI is certainly not the only 2-year college, or educational institution in general, which struggles with issues relating to student placement, or retention and achievement. Recent studies indicate soberingly limited student success — whether defined as progress within a specific program, program completion rates, or demonstrated skill improvement — in basic skills, remedial/developmental English (and mathematics), and English as a Second Language (ESL) courses at junior and community colleges throughout the US (see, Bailey, 2009; Bailey, Jeong & Cho, 2010a, 2010b; California Community Colleges Chancellor's Office, 2009; Martorell & McFarlin, 2011; Offenstein & Shulock, 2011). While this has led to many questioning the pedagogical effectiveness of these programs, Hughes and Scott-Clayton (2011) suggest the widespread lack of local investigations into the suitability and performance of the placement instruments for these institutions' particular students, courses, programs, and educational objectives, may be at least partly to blame.

Hughes and Scott-Clayton also suggest the use of a single, multiple-choice instrument — typically Accuplacer or Compass (Brown & Niemi, 2007; Hughes & Scott-Clayton, 2011; Sullivan, 2008) — for informing placement decisions at the majority of US colleges may also play a significant part in the ongoing struggles of their programs and students. The relative contributions of standardised tests and locally designed and marked performance assessments, particularly writing samples, is a long-standing and highly polarized debate in education and assessment circles (Hillocks Jr, 2002; McLeod, Horn, & Haswell, 2005; Sullivan, 2008; White, 1990). While several researchers and educators suggest standardised instruments offer insufficient insights into the myriad abilities and characteristics likely to influence student success in college programs (Armstrong, 2000; Hillocks Jr, 2002; McLeod, Horn, & Haswell, 2005; Murphy & Yancey, 2008; White, 1990), others are of the opinion that writing samples or other subjective assessments are unnecessary or even counterproductive to the placement decision process (Saunders, 2000; Sullivan, 2008).

The current study may offer a rare opportunity to compare the relative advantages and disadvantages of widely used representatives of standardised tests and local performance assessments — Accuplacer and a writing sample – used in the same context, for the same purpose, and with the same participant pool. Further, we may be afforded the opportunity to evaluate specific qualities of these instruments which have been particular points of contention in the literature, such as reliability, relevance to the instructional domain, and their sufficiency for the placement decisions being made.

The current study at CMI, then, may be able to offer important insights for stakeholders at colleges, or educational institutions in general, which currently struggle with issues of placement test suitability and performance, which, according to a recent survey of 2-year colleges, is the vast majority (Sullivan, 2008). Should it prove fruitful, and provide enlightening information leading to the improvement of the placement system and resulting decisions and consequences at CMI, perhaps it could also serve as impetus for similar studies being developed at other institutions.

1.2.3 Insights for validity and validation studies

It is also suggested that the current study is important because of the ongoing need in linguistic and educational assessment for more applied, argument-based validation studies. Over the past three decades, there have been considerable shifts and developments in the investigation of validity and validation, from the largely conceptual work of establishing the nature of validity (Cronbach, 1988, 1989; Kane, 1992; Messick, 1989; Shepard, 1993; Spolsky, 1981), to the creation of validation frameworks for guiding investigative work (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2004; Kane, 1992, 2004, 2006; Kane, Crooks, & Cohen, 1999; Mislevy, Steinberg, & Almond, 2002, 2003) culminating in the opportunity and great need for widespread application of those frameworks for assessment design and evaluation (Bachman, 2005; Bachman & Palmer, 2010; Chapelle et al., 2008).

With increasing demands for accountability coming from education officials, government agencies, accreditors, policies such as the US's “No Child Left Behind”, and more widespread use of tests to inform (or entirely base) decisions about individuals, programs, schools, and entire education systems, the need for addressing questions of how, why and when assessments are used, the consequences their use creates, and the ethical obligations of test developers and users has never been greater or more widespread. “Turning these challenges into accomplishments will depend upon the willingness and capability of language testers to apply the knowledge and skills acquired over the past half century to the urgent practical assessment needs of our education systems and societies” (Bachman, 2007, p. 1). It is hoped, then, that this study may offer some momentum for this push to turning these challenges into accomplishments, as Bachman puts it, through contributing to the application and development of argument-based validation studies, particularly of high-stakes, widely used assessment tools such as Accuplacer, and locally marked writing samples. This study may well represent the first such investigation, addressing both interpretation and utilization, of Accuplacer instruments — the most widely used placement tests amongst US colleges.

Finally, as Xi reminds us, echoing the sentiments of Cumming (2004), there is a need for the expansion of the knowledge base of language (and, one could argue, particularly higher education) assessment, “to include research on contexts and learner populations other than academically-bound young adults at universities in English-speaking countries. These unique contexts and populations may present us with new challenges in developing validation research paradigms and methods” (Xi, 2008, p. 139). It is suggested, therefore, that this study is also of importance because it involves a context — the Marshall Islands, Micronesia, and the Pacific Islands in general — and learner population not typically addressed in the assessment, testing, or education research literature (or, if they are, they are generally lost amongst the rather problematic classification of “Asian & Pacific Islander”, in which Pacific Islanders often make up a small minority, and Micronesians and Marshallese even less so, and their unique needs and issues are often washed out and not identified or addressed).

1.3 Context of the study

This section will provide information regarding the context in which the current study took place. We will start with the broad context of the college, its mission, the background in which it exists and operates, and a description of its student population. A brief review of the Developmental English program, its course structure and course objectives will then be presented, followed by an account of the current placement assessment system and the constituent instruments employed to exempt, exclude or place students within the program.

1.3.1 The College of the Marshall Islands

The College of the Marshall Islands is a small, 2-year college, serving approximately 800 students in the tiny, Pacific atoll nation of the Republic of the Marshall

Islands (RMI). The college exists amidst a backdrop of poverty and economic stagnancy, with unemployment rates of 30-50% (Bright & Chutaro, 2007) and rising, and the vast majority of income for the country coming from foreign aid. While the population of the nation is only some 60,000, booming birth rates and ever increasing numbers moving to the urban centres of Majuro and Ebeye (already amongst the highest population concentrations in the world with 38,000 and 82,000 people per square mile, respectively) (Bright & Chutaro, 2007; United Nations, 2004) mean already strapped health, sanitation, and education systems are only becoming more strained. In addition, the Compact of Free Association with the US government will soon expire, and the millions in guaranteed financial aid from the nation's biggest benefactor are soon to disappear. As the official higher education institution of the nation, there is a dire and immediate need for the college to do everything it can to help improve the education and employment opportunities of the Marshallese people.

The Official Philosophy of CMI would seem to recognize this responsibility, and the need for education in the fields it offers, such as Nursing and Elementary Education, stating: “The regents, administration, faculty, and staff of CMI believe that quality education is essential to the well-being… of the Marshallese people as a whole, now and in the future. We are therefore committed to… educational content address[ing] the general and specific needs of the students, the local community and the nation” (CMI, 2008, p. 3). The college's Official Mission Statement further “recognize[s] the need to raise the standards of higher education in this nation to internationally required levels” (*ibid.*).

Whether or not the college has lived up to these lofty goals, however, is a common issue of debate amongst many, including members of the parliament, Ministry of Education, and the Marshallese population at large. While officially recognized by the Accrediting Commission for Community and Junior Colleges (ACCJC, a division of the Western Association of Schools and Colleges, or WASC), the school was placed on probationary status in 2003 (ACCJC, 2003), and soon thereafter on “show cause”, the last step before losing accreditation altogether. As, with very few exceptions, almost every CMI student is dependent on Pell grants – a US federal grant offered to eligible low-income students, loss of accreditation, and therefore ineligibility of students for US financial aid, could have potentially shut down the campus.

However, the hiring of a new president and initiation of extensive reforms, particularly the accelerated pursuit of internal analysis and review (i.e., assessment, in the program review or quality assurance sense of the term) would appear to have produced a positive turnaround for the college. In 2006, following an official inspection, the ACCJC upgraded the college's status to ‘probationary', and in 2009, the college once again was granted full accreditation (CMI, 2008, 2009).

College constituents and the local community celebrated the improved academic health of the college, and many attributed it to the considerable efforts of constituents on the complex, often painful, assessment endeavours. In July of 2008, based on the results of the institutions' recent Community College Survey of Student Engagement (CCSSE) CMI was recognized for ‘outstanding performance' in promoting high levels of student learning and retention (CMI, 2008). Perhaps most encouraging of all, though, was that both student enrolment and year-to-year retention rates reached record highs in the Fall 2008 semester. It was in this context of appreciation for internal analysis and improvement, that the current study was proposed, and widely supported, as a means of fulfilling accreditation requirements, ethical and professional obligations to campus constituents, and addressing lingering concerns regarding the suitability, performance, and impact of the current placement system and its component instruments.

1.3.2 The student population

According to the most recent self-study submitted to accreditors, the vast majority of CMI (2009) students are: Marshallese (96%), English Language Learners (98%), reliant on financial aid in the form of US Pell Grants (99.5%), and academically underprepared (35% of all applicants do not meet minimum English language skill estimates to be eligible for any CMI programs and 92% of accepted applicants are placed into Developmental English courses). Additionally, approximately 50% of all students are first-generation college students.

For many of the students, particularly those from the ‘outer islands', the move to live in an urban setting, and to pursue a tertiary education at CMI, represent significant challenges. Many will have limited, perhaps even not any, interactions with non-Marshallese before, have only used English in whatever English classes they may have had in primary or secondary school, and might have had little to no experience with computers or perhaps not even had access to electricity, let alone modern technology. Additionally, for many students, literacy in both English and Marshallese is a problem. Marshallese is traditionally an oral language, and while there have been efforts to develop, and teach in the schools, a common written form, none have enjoyed much success. The result is a majority of incoming students to CMI who must not only struggle to cope with the acquisition of the target language, but also do so without the benefit of similar literacy skills or knowledge in their first language to draw upon.

1.3.3 The Developmental English program

English is the official course of instruction at CMI, and all courses, save those dealing with Marshallese culture, history and language, or Japanese or Chinese language, are taught and assessed entirely in English. As such, all students need to be able to thrive in an English medium classroom in which they will be required to study, discuss, and write about abstract concepts and academic subjects. With the vast majority of incoming students being underprepared in terms of English language abilities, academic abilities in general, and often limited L1 literacy skills to draw upon, the Developmental Education Department, and particularly the Developmental English Program, plays an extremely important, and certainly challenging, role at the college.

They must also meet this challenge in an extremely limited timeframe, as, due to Pell Grant eligibility restrictions, students must complete their developmental preparation within approximately three semesters. For students placed in Level 1 of the program, which is the majority of incoming pupils, this leaves little room for any academic missteps, or they can find themselves out of Pell grant money and with developmental courses still required before they can transfer to a credit program.

It is largely around this financial aid time constraint that the Developmental English (and larger Developmental Education) program has been designed. Consisting of three, semester-long sessions, the program is comprised of three levels described in official course outlines as ‘pre-intermediate', ‘intermediate' and ‘pre-college' (CMI, 2010).

Each level is comprised of two courses, one Reading and Writing (RW) and one Listening and Speaking (LS). The General Outcomes and Student Learning Outcomes (SLOs) for all courses of the Developmental English program, as listed in Appendix A, demonstrate the scaffolding of the overall program and the interconnectedness of the outcomes and competencies addressed from one course level to the next. Likewise, general and specific outcomes for the first-semester credit English courses — Composition, and Speech — have been referenced in the design of developmental SLOs and curricula, and subcommittees of the Curriculum and Assessment Committee, called ‘Bridge Committees', occasionally meet to ensure curricular and assessment alignment from Developmental English through to credit English. It should be noted, however, that these outcomes are not considered finished. Their analysis, discussion, and revision is ongoing as part of the quality assurance program at CMI, and there are certainly areas that might be less than clearly established as yet, such as what is intended by terms like ‘pre-college' or determining what exactly is meant by ‘pre-intermediate' listening skills as opposed to ‘intermediate' or ‘pre-college', for example.

To date, no investigation as to the particular relevance of the placement instruments to the curricula or learning outcomes of Developmental English courses has been conducted. This is an area of need that will hopefully be expanded upon in future evaluations of the placement assessment system and its instruments.

1.3.4 The current placement assessment system

This section will briefly describe the placement assessment system at CMI, starting with an outline of the testing sessions in which candidates complete the placement instruments. A detailed description of both procedures, Accuplacer Companion and the writing sample, will then be presented, followed by a description of the decision-making process employed to determine where applicants start their studies with CMI.

1.3.4.1 Placement assessment sessions

At the time of the current study, CMI used two instruments to inform its placement decisions: Accuplacer Companion and a locally designed and marked writing sample. While applicants complete both sections of Accuplacer Companion, Mathematics and English, only the English subtest results impact the decision of whether they will be eligible for registration at the college. Applicants must be considered ready for at least Level 1 of the Developmental English program to gain entry to the college.

Candidates complete the placement instruments during scheduled testing sessions held at least twice per year at each of the numerous locations in both the urban centers and ‘outer islands’. Testing sessions usually take place in high schools or other public buildings with sufficient space and suitable facilities — desks, lights, etc. — for the procedure. All instruments chosen must be paper-and-pencil, to avoid problems with irregular or lack of access to electricity, and to avoid issue with the vast variance in applicant familiarity with computers or technology in general impacting student performance on the instruments.

There is no fee for candidates to take the tests, and while recent or imminent secondary school graduates are the focus of the sessions, any member of the public who has completed high school or obtained a GED certificate is welcome to take the tests. There is no limit to how many times a candidate can take the placement tests.

Recent and imminent high school graduates are the primary focus of these testing sessions, because the Ministry of Education of the RMI requires CMI administer their placement tests to all graduating secondary school students in the country. It is presumed this is done as a means for the ministry to gather performance/achievement measures for their schools and students; however the purposes for their requirements have never been clearly shared with the college.

Exam sessions range from 20 to 200 examinees, depending upon the location and time of year, and the college typically examines 450-600 potential applicants for fall student intake and another 150-200 for spring, though these numbers are rising (CMI, 2009). It should be noted that only a small portion of those tested ever apply to or attend CMI. All candidates are provided copies of Accuplacer Companion first, and then instructions for completing the test (use of provided pencils only, how to complete the answer key, no talking, etc.) are conveyed in written English and spoken Marshallese. There is no time limit on the completion of the instrument, and examinees are instructed to submit their exam and answer sheets when they feel they have done their best. After completing the Accuplacer test, students may then start on the writing sample. Again, there is no time limit imposed on this instrument. With the English and Mathematics sections to complete for Accuplacer Companion, the writing sample, and the completion of various other information sources for the Ministry of Education and CMI, testing sessions can last in excess of four hours.

All testing sessions are proctored by a minimum of two CMI representatives (more for larger testing sessions), typically one or more member of the registration and enrolment staff and a member of Developmental English faculty as well. At least one proctor must be fluent in Marshallese in order to provide overall instructions and guidelines. They are also free to answer questions pertaining to the instructions and regulations of the exam, but not to discuss anything that might unfairly help one examinee over another, such as the meaning of specific words or sentences in test items, readings, or answer options.

All copies of the instrument, answer sheets and any other materials required are brought in and taken out by the test administrators. All candidates' identification is checked before or during the exam process. More details about each of the placement tests are provided below.

1.3.4.2 Accuplacer Companion

Accuplacer Companion is a multiple choice (4-option), standardised test published by the College Board. The test has two sections, English and Mathematics, both of which applicants must complete, with results from each section respectively informing placement decisions into the Developmental English program and the Developmental Mathematics Program. However, only English placement decisions include the category of not currently being eligible for registration at CMI. The English test is designed to help higher education institutions determine whether new students would be better served starting in a developmental English course or directly entering credit level English. It is intended for use with students for whom English is a “best language” (College Board, 2003, p. 12) and is the paper-and-pencil version of Accuplacer OnLine, a web-based, adaptive placement instrument.

The English section of AC consists of two subtests, each with 35 questions: Reading Comprehension, and Sentence Skills. The Reading Comprehension test, not surprisingly, “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, all of which comprise between 12-29% of the items on this section of the instrument: “(a)

Identifying Main Ideas, (b) Direct Statements/Secondary Ideas, (c) Inferences, (d) Applications, and (e) Sentence Relationships" (*ibid.*). There are two types of questions in this subtest. In the first type, examinees are presented a reading text, typically narratives ranging from a few sentences to longer passages, and then a series of test items requiring the test taker to identify or demonstrate understanding of the main idea, secondary ideas, or make inferences based on the text. The other type of questions involve candidates reading two sentences and then answering a question about the relationship between them, such as whether the information presented in one supports, refutes or repeats the information or ideas presented in the other, for example.

According to the publisher, the Sentence Skills subtest addresses 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, in approximately equal frequency in terms of the percentage of test items addressing each, are: (a) Recognizing Complete Sentences, (b) Coordination/Subordination, and (c) Clear Sentence Logic. Candidates must answer two types of questions in this section. The first presents them an incorrect sentence and requires them to choose the best of four attempts at correcting it. The second type, "construction shift" questions (*ibid.*), present students a sentence and then ask them to choose the best option presented which rewrites the sentence but does not change the original meaning. Some items are intended to address examinee comprehension of the logic of the sentence and others the relationship between coordination and subordination.

For both sections of the instrument, publishers assure the test users that readings and items have been carefully selected, and tested with examinees from a wide variety of backgrounds, in order to eliminate or at least limit the potential for construct-irrelevant variance due to factors such as cultural bias, familiarity with the topic, offensive content, and so on, as well as for any typographical errors, logical problems, lack of a ‘best answer' or similar problematic issues.

1.3.4.3 The writing sample

The writing sample is an untimed, short essay in response to a brief prompt. The prompts are designed locally, by a small group of English course instructors, with periodic revisions and advice also sought from an outside consultant. In order to make prompts as readily understandable to all applicants as possible, they are kept relatively short, simple, and generic in topic, requesting such things as descriptions about future career plans, talking about someone they admire, or explaining why the applicant wants to study at CMI, for example.

Approximately 10-15 developmental and credit English course instructors participate in rating the writing samples each semester. All raters have experience teaching more than one level of English at CMI and are familiar with the student learning outcomes of the various courses. While rater training and norming sessions were conducted upon the adoption of the current rubric, no further rater training has occurred since. Certainly no rater training or norming occurred from the time the author joined the institution, at the start of the Fall 2008 semester, through to the end of the Fall 2011. Due to high turnover of faculty at the college, this means that the majority of the current instructors/raters employed by the college have never completed a training or norming session, and none would have participated in one recently. Raters are volunteers — participating is encouraged but not required — and are paid approximately $1 US per exam read. The majority of raters are native speakers of American English, though some raters at the time of the study were also native speakers from Canada, the UK, New Zealand, Fiji, and elsewhere. A smaller number of English instructors, and raters, could be described as ‘near-native speakers’ of English, as all have completed graduate degrees in English-medium educational institutions and all have prior experience living and working in English-speaking environments.

Each composition is rated by at least 2 English instructors. Rating is done according to a locally developed, analytic scoring rubric (see Appendix B). The criteria included are: Support, Organization, Sentence Variability, Diction, and Errors in Grammar. No further information or documentation explaining the criteria could be found. For each criterion, there are a number of descriptors intended to assist the rater in judging the writing sample, and the rater must check the descriptor they feel best describes the candidate's work. For example, the descriptors for Diction are:

A. Word choice inaccurate in much or all of the response
B. Word choice often inaccurate
C. Some inaccurate word choices
D. Word choice is mostly accurate
E. Words occasionally used inaccurately
F. Consistently precise in word choice

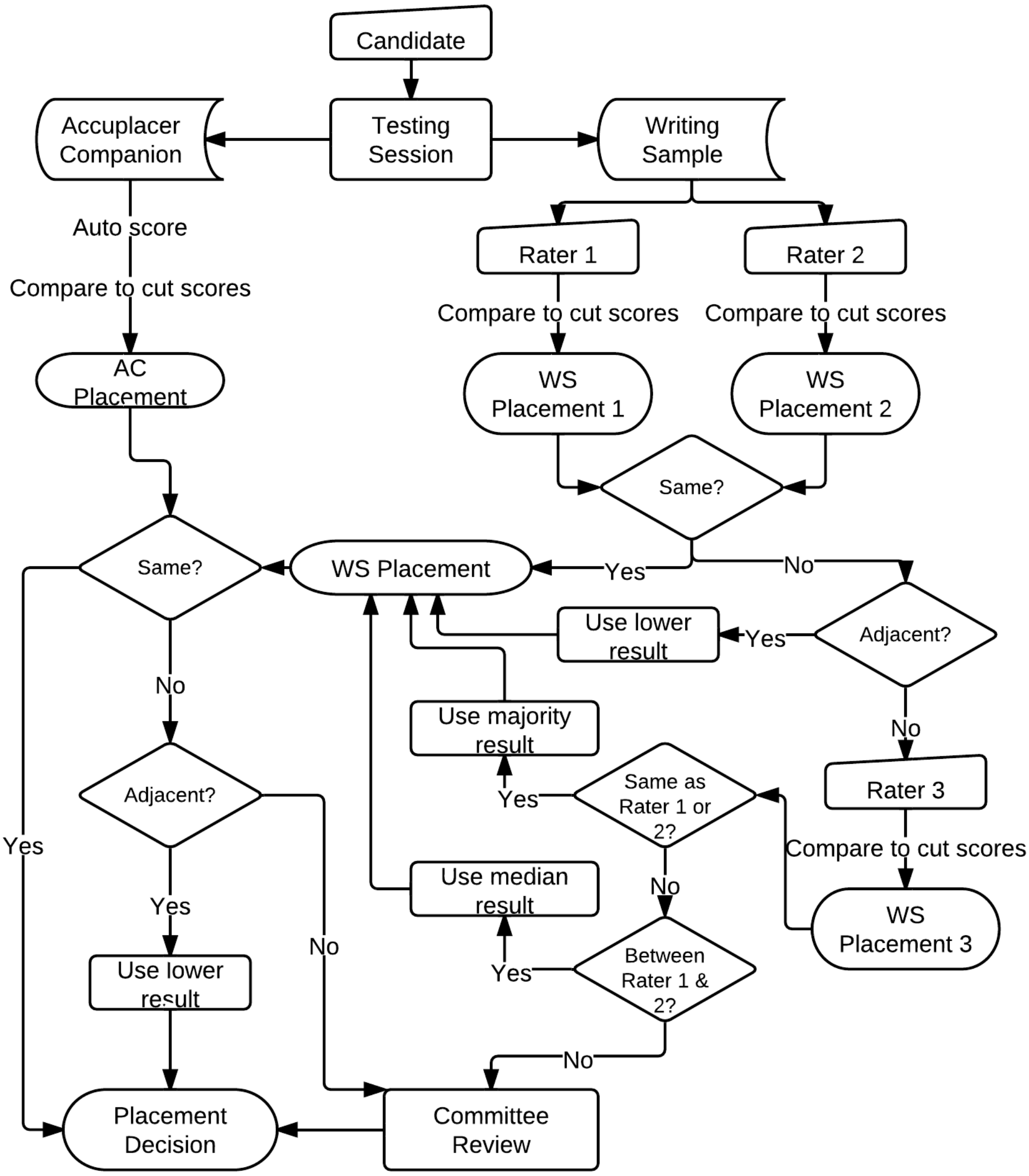
Each criterion has six descriptors to choose from, each corresponding to a score from zero to six. Results for all criteria are averaged for a final score out of 6 and then entered into a spreadsheet application.

1.3.4.4 The placement decision process

Figure 1.1 shows the series of steps, some automated and some requiring the judgment of raters or committee members, involved in reaching a final placement decision for a candidate.

Having completed both placement instruments, candidates' performances must be scored. The answer sheet for Accuplacer Companion is automatically scanned, scored, and entered into a spreadsheet application. The placement essay is conveyed to two raters, both of whom are English instructors at the college. All raters use the common, analytic rubric described in section 1.3.4.3 and results from both raters are added to the spreadsheet.

Figure 1.1: Placement Assessment System Flowchart



AC = Accuplacer Companion
WS = Writing Sample

Once data for AC and both WS raters have been entered into the spreadsheet application results for each are automatically referenced with the relevant, locally established, cut scores in order to determine a recommendation for student placement. For Accuplacer Companion, the cut scores at the time of the study were:

43-70	Credit English
37-42	Developmental English Level 3
30-37	Developmental English Level 2
15-29	Developmental English Level 1
0-14	Not ready for any CMI English course

The placement recommendation determined from the Accuplacer score is automatically generated by the spreadsheet via comparison of the result with the established cut scores. The process for producing the placement recommendation for the writing sample, however, can be much more complex.

For the writing samples, each rater's total score (the average of the scores for each criterion in the rubric) are compared with the following cut scores:

10.0-12.0	— Credit English
8.0-9.9	— Developmental English Level 3
6.0-7.9	— Developmental English Level 2
4.0-5.9	— Developmental English Level 1
0.0-3.9	— Not ready for any CMI English course

However, while the maximum score possible for WS results is six, the cut scores range up to twelve. No one remaining at the college was directly involved in the establishment of the cut scores, and none could offer an explanation as to why the rating scales would differ. Perhaps the original intention was to add the two raters’ results together and then compare the combined score with the cut scores.

As employed since Fall 2008, however, the WS result from each rater, upon entry into the spreadsheet, is automatically doubled, compared to the cut scores above, and then a placement recommendation for each rater’s results is generated. The two recommendations are then compared. If they are the same, that is the final WS placement recommendation. If the two judges' ratings result in different levels being recommended, but they are adjacent (e.g., Level 1 and Level 2), the lower of the two becomes the WS placement recommendation. If they are not adjacent, then a third reader is required to rate the essay. This rating is compared with the cut scores, and a third recommendation is generated. If two of the three recommendations agree, that majority recommendation is used. If the three are sequential (e.g., Level 1, Level 2 and Level 3), then the median recommendation (Level 2) is used. If none of these parameters fit, (e.g., Level 1, Level 3 and credit English were the three recommendations) then the writing sample, raters' scores (and rubrics, possibly with notes) and Accuplacer results are forwarded on to a placement committee to review and make the final decision as to where the student will be placed.

Once placement recommendations from both AC and the WS have been established, the two are compared. Similar rules apply here as for the comparison of the recommendations of the different writing sample raters. If the levels are the same, that is the final placement decision. If they differ by one level, then the lower of the two is used. If the two recommendations differ by two or more levels, all materials and scores are referred to a placement committee which will review the evidence and make a final placement decision.

Placement committees consist of a minimum of two English instructors at the college, and typically include the Chair of the Developmental Education department, though this is not mandated. These members review all available evidence — the writing sample, the raters' assessments, and Accuplacer Companion test and subtest results, and determine a final placement decision, which must be unanimous.

Having established the impetus and significance of the current study, and described the context in which it was conducted, the purpose and research questions for the investigation will now be presented.

1.4 Purpose and research questions

The primary purpose of this study was to construct and investigate an initial validity argument for the current instruments and overall placement assessment system employed by CMI to exempt, exclude or place students within their Developmental English program.

It was hoped that this investigation would shed light on a number of areas, most of which are believed to be of importance to CMI, as well as other institutions which use standardised instruments and/or writing samples to inform similar decisions, and some of which are thought to be relevant to the study of validity and practice of test validation. To address the purpose of the study, three research questions were formulated.

Research Question (RQ) 1: With regard to test score interpretation, what do the results of the study indicate relating to: a) the relative advantages of AC and the WS in the CMI context; and b) current debates in the literature regarding standardised instruments and WS's for placement purposes?

RQ 2: With regard to issues of assessment utilization, what do the results of the study indicate relating to: a) the relative advantages of AC and WS in the CMI context; and b) current debates in the literature regarding standardised instruments and WS's for placement purposes?

RQ 3: Did the articulation and investigation of the validity argument produce insights identifying opportunities for improving the placement assessment system for the benefit of stakeholders at CMI?

These research questions will be restated after the literature review, in order to link the issues illuminated in the literature and the foci of the current study. Additionally, each question, and the relevant evidence addressed in the study, will be discussed in Chapter 5, along with an examination of implications for future research revealed by the investigation.

1.5 Limitations of the Study

Before moving on the literature review, it is important to recognize the limitations of the current study. For issues which are more extensively dealt with in later sections, such as data and/or methodology issues which are attended to in Chapter Four, the materials and methods chapter, only brief summaries will be provided here.

One significant limitation for the current study was access to data. This restriction came about in two ways. First, while the institution and stakeholders were almost universally supportive of the investigation, being not only a researcher, but also an instructor at the institution placed limits on the sorts, and level of detail, of data that could be made available. While aggregate data for AC and WS results for all candidates could be provided, for example, individual responses for each AC test item was considered too much detail to ensure candidate confidentiality. This limited the options available for a number of data analyses, including internal consistency estimates, the breadth of the sample size for predictive validity estimates, and so on.

The second way in which access to data occurred was the lack of the data itself. Due to inconsistent policies, a longstanding lack of focus on record-keeping and/or internal self-evaluation until recently, and lost data due to hardware failure from salt air, power outages, power surges and other environmental factors, some data which would have helped inform many aspects of the current study were not available. As one example, investigating the influence of the different prompts used on the WS outcomes was not possible as the institution would not seem to have kept any record of which prompt was answered by the candidate.

Another limitation, not only of this study, but of nearly all studies involving predictive validity, is that factors known to influence course outcomes, such as motivation, study skills, punctuation, time management, and so on, are not accessible via one or more language assessment.

Finally, while the inclusion of multiple constituents provides advantages, such as maximizing perspectives considered and insights available, hopefully improving stakeholder buy-in and likelihood of continuation of validation investigations in the future, and resulting “ecological validity” of the instruments produced for data collection, it also has its drawbacks. The questionnaires used in the study, for example, cannot be said to focus on a specific construct, or reflect any particular theory or model. Rather, they are the product of negotiations between the researcher and other stakeholders, such as faculty and administrators, intended to gather data each represented group was interested in, while also keeping the instruments concise.

1.6 Overview of the dissertation

In the following chapters, a comprehensive review of changes and advances in validity and validation studies in recent decades will be presented, leading up to current theories and frameworks in the field. Subsequently, recent empirical studies utilizing these frameworks for validity investigations will be discussed. Upon completion of the literature review, a detailed description of the methods and materials used for the current study will be offered. The final two chapters of the dissertation will present the results of the various lines of investigation involved in this study and, subsequently, a discussion of their implications for the local context, CMI, for placement testing and validity investigations in a broader context, and possible directions for future research.

CHAPTER TWO

Literature Review: Recent Advances in Test Validity and Validation

2.0 Introduction

In discussions of testing and assessment, be it in the realms of psychology, education, or languages, the term *validity* has consistently been used to refer to the "quality or acceptability of a test" (Chapelle, 1999, p. 254). However, "[b]eneath the apparent stability and clarity of the term… its meaning and scope have shifted over the past years" (*ibid*). As the definition and assumptions we hold regarding validity shape how we approach *validation*, the process of gathering and evaluating evidence in order to establish the 'quality and acceptability' of an assessment, it, too, has undergone significant change in recent history.

The intent of this chapter, then, is to review developments in the conceptualization of validity and consequent practice of validation from the 1970s until today, in order to establish the chronological and theoretical background for the current study and the two modern validity frameworks it employs: Kane's (1992, 2004) interpretive argument and Bachman's (2005) Assessment Use Argument. To be certain, significant contributions and developments occurred in the study and application of validity prior to the 1970s. However, a more comprehensive investigation of the history of the field would be beyond the scope and purpose of this thesis. Further, in the interest of brevity, the works considered here are limited to those felt most influential to the study of validity and practice of validation, and most pertinent to the eventual establishment of Kane's and Bachman's respective models.

2.1 Validity and validation over recent decades

As summarized in Table 2.1, advances in the study of validity have come rapidly in recent history, resulting in a substantial rethinking of best practices in validation. From the widely held perspective of the 1970s, of validity as a collection of disparate factors – content, face, criterion (including concurrent and predictive), and construct validity –

validity is now widely perceived as a unified concept, with construct at its core and the careful construction and evaluation of comprehensive arguments the preferred means of its investigation.

Table 2.1: Conceptualization of Validity and Validation over Recent Decades

	Decade			
	1970s	1980s	1990s	2000+
View of Validity	Disparate factors: e.g., criterion, content, face, construct	Discrete factors, but emerging importance of construct	Unitary concept, with construct subsuming all other aspects	Unitary concept, but with increasing work on practical applications
View of Validation	Largely the establishing of correlational evidence with other criteria, particularly other instruments	Growing dissatisfaction with correlational approach	Hypothesis testing: i.e., the collection of data to approve or refute the validity hypothesis	Creation and investigation of comprehensive arguments (validation frameworks)

The traditional view of validity as disparate factors prevalent in the 1970s had held dominance in testing since before World War II (Angoff, 1988; Cronbach & Meehl, 1955). Validity was largely considered a constant property of the instrument itself, relatively unaffected by factors such as the context within which and the purposes for which the assessment was used. As a result, test validation focused principally on estimates of reliability, analyses of content, and how well results correlated with other tests believed to address the same or similar skills, knowledge or characteristics (Chapelle, 1999; Cronbach, 1971; Kane, 2001), and few, if any, discussions of the consequences of test use appeared in the literature (Shohamy, 1993).

The 1980s saw growing dissatisfaction with this paradigm (Cronbach, 1980, 1988) and increasing momentum behind a fundamental shift in the conceptualization of validity and practice of validation. Forays into the design of, and support for, tests of integrated skills (Klein-Braley, 1985; Klein-Braley & Raatz, 1984) and performance-based language assessment (Xi, 2008), particularly communicative competence (Bachman & Palmer, 1982), and the realization they required different validation procedures (relative to indirect, discrete-point, measures, such as multiple-choice exams, for example), led to increased attention to the still relatively undeveloped concept of construct validity (Clark, 1975; Palmer, Groot, & Trosper, 1981; Xi, 2004). The conventional, disparate view of validity was becoming increasingly seen as fragmented and incomplete, particularly since it failed to account for; i) what test scores, particularly those of traditional multiple-choice measures, actually mean; and ii) the consequences of test use and test score use for the various stakeholders affected.

Calls for alternative means of validation through something more akin to hypothesis-testing (Bachman & Palmer, 1982; Chapelle, 1999; Cronbach, 1980) began to appear, as did appeals for the inclusion of test use and test use outcomes as part of the purview of validity and validation discussions (Spolsky, 1981). By the end of the decade, concepts which would become central to discussions of validity, such as the idea that particular tests may be valid for one purpose but not another (Chapelle, 1999; Henning, 1987), the idea of test washback (Hughes, 1989), and the ethical implications of assessment (AERA, APA, & NCME, 1985; Canale, 1987) began to demand more attention in validity and assessment discussions.

This changing zeitgeist and foundational work set the stage for Messick’s (1989) galvanizing publication, “Validity”.

2.2 Messick's unitary validity

Messick (1989, p. 13) defined validity as “an integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". This conceptualization, as presented in Table 2.2, highlights three fundamental aspects of validity still widely held today.

Table 2.2 Messick's Facets of Validity (adapted from Messick, 1989, p. 20)

	Test Interpretation	Test Use
Evidential Basis	Construct validity	Construct validity + Relevance/utility
Consequential Basis	Value implications	Social consequences

First, it is unitary, with construct at its core and all other forms of validity, such as content and criterion, as well as reliability, as pieces of evidence supporting or refuting construct validity. Second, the legitimacy of inferences based on test scores is entirely dependent upon the veracity of the theories and assumptions behind the design and implementation of the test. Resultantly, validation must be the process of theory-testing and involve the on-going collection and analysis of various sources and types of evidence. Third, Messick's view of validity included test use, and the outcomes of test use, on individuals, groups, and society, solidifying these issues as central to discussions and investigations of validity.

Messick's unitary validity was extremely influential in educational and psychological assessment; so much so that Bachman (2005, p. 4) divides validity studies into two eras, the "pre-Messick" and the "Messickian". However, Messick's theoretical model (Table 2.2) was abstract in nature and not readily amenable for validation studies in practice. Considerable work was required to adapt the framework into something that could usefully guide validation studies for test researchers and test users (Shepard, 1993; Xi, 2008). This need resulted in subsequent development of validation frameworks.

2.3 Early validation frameworks: Focus on score-based interpretation

“Validation frameworks specify the process used to prioritize, integrate, and evaluate evidence collected using various methods” (Xi, 2008, p177). Such structures that are simultaneously theoretically sound and practical are valuable, and rare, commodities for guiding both the design and evaluation of assessment instruments (Chapelle, Enright, & Jamieson, 2008; Moss, 2007; Stoynoff, 2009).

Amongst considerable work put forward in this search for a functional bridge between theoretical conceptions of validity and applied validation studies, almost certainly the most influential in educational assessment has been Kane’s (1992) interpretive argument.

2.3.1 Kane's interpretive model

Kane (1992) proposed the idea of an interpretive argument as a framework for articulating a case for an instrument’s validity by establishing the inferences and assumptions associated with the score interpretation, and then gathering, evaluating, and presenting the relevant evidence. Kane, Cooks and Cohen (1999), as shown in Figure 2.1, presented the interpretive argument as a model with four components — observation, observed score, universe score and target score — and three inferences bridging one component to the next — scoring, generalization, and extrapolation.

Figure 2.1 Inferences and components of an interpretive validity argument (adapted from Kane et al., 1999, p.9)

Inferences		*Scoring* ⇨		*Generalization* ⇨		*Extrapolation* ⇨	
Components	Observation		Observed Score		Universe Score		Target Score

The first inferential bridge, scoring (later referred to as evaluation), involves transforming an examinee’s performance on an assessment (the observation) into an observed score. Assumptions here requiring supporting evidence include appropriateness and consistency in the scoring procedures and the conditions under which the performance was obtained. We must be assured at this stage that results represent the influence of construct-irrelevant factors as little as possible. Construct-irrelevant factors may include scoring procedure errors, confusion caused by poorly written instructions, references to information unfamiliar to the examinee, noisy assessment conditions, or anything else which may reduce the score's representativeness of the candidate's competency we are interested in, or their ability to demonstrate this competency to the best of their abilities (Kane et al., 1999).

Next, this observed score is generalized to the universe (sometimes called true) score. Based on measurement/generalizability theory, generalization relies on the assumption that the observed score is representative of what examinees would have obtained on multiple tasks similar to those of the assessment, gathered in a multitude of test settings.

Finally, extrapolation refers to the inference of what the examinee can do outside of the realm of the test itself, what Kane and his colleagues (1999) call the "target score", such as what the test-taker might be expected to be capable of doing in a language-related activity in real situations. Assumptions here requiring supporting evidence would include that the test tasks engage the abilities and processes intended by the test designer (later referred to as the explanation inference), and are relevant to the target domain (extrapolation) (Kane, 2004; Xi, 2008).

Kane's approach, then, views validity as the extent to which the inferences made on the basis of the assessment outcomes are appropriate. Validity is evaluated based upon the collective strength and coherence of an argument comprised of several rational conclusions, all supported by relevant theory and empirical evidence (Haertel, 1999; Marion & Pellegrino, 2009). This offers us at least four major advantages. First, compared with an approach which relies on individual pieces of evidence, as with the disparate factors approach of the 1970s and prior, or focuses solely on construct, as the *Standards* largely advocated in the 1990s (APA, AERA & NCME, 1999), the inferential argument guides researchers to consider and evaluate evidence relating to a comprehensive array of factors influencing instrument validity, beyond solely construct or criterion related issues; areas which may otherwise have been left unconsidered (Kane, 1992; Bachman, 2005). Second, in so guiding the consideration of validity investigations, it helps to limit the possibility of, consciously or not, seeking and/or utilizing evidence which fits with our pre-existing biases towards the instrument(s) in question (Marion & Pellegrino, 2009; Xi, 2008). As Cronbach put it, “falsification, obviously, is something we prefer to do unto the constructions of others” (1989, p. 153).

The third advantage of the inferential model, as Chapelle et al. (2010) point out, is that, compared with the construct-focused method advocated in the *Standards* (AERA, APA & NCME, 1999), the inferential argument offers the substantial advantage of a means of approaching validity investigations with instruments for which the intended construct is too complex to readily allow for a concise definition or to be the sole basis upon which to establish a claim for validity.

Finally, all of these advantages result in arguably the most important benefit of all: by providing greater guidance in the creation and evaluation of a comprehensive validity argument, and a means of addressing the validity of instruments with difficult to define constructs, Kane’s approach helped make validation an endeavour more feasible for a much broader population of educators and other test users (Marion & Pellegrino, 2009).

Within the realm of language testing, Kane’s model inspired influential structures for validation studies, including Chapelle, Enright, and Jamieson’s (2004, 2008, 2010) work with the TOEFL, and the Assessment Use Argument of Bachman (2005) and, later, Bachman and Palmer (2010), all of which are presented later in this chapter.

However, Kane's model, particularly in its early iterations, was limited in that it did not attend to issues of utilization, including impact on test takers and other stakeholders, which Messick's conceptualization clearly held as within the realm of validity. This was a void, which, in the realm of language testing, Bachman and Palmer (1996) attempted to address.

2.3.2 Bachman and Palmer's test usefulness framework

Bachman and Palmer (1996) presented a framework for test usefulness intended to make Messick's work more accessible to test developers and users. The usefulness framework subsists of six qualities, four of which – construct validity, reliability, authenticity and interactiveness – address test score interpretation and roughly overlapped with Kane's explanation, generalizability, extrapolation and evaluation inferences, respectively. The remaining two qualities, impact and practicality, attend to issues not yet addressed in Kane's framework: test utilization.

Impact refers to the consequences of test use, including effects on instruction (washback), on individual test-takers, instructors, courses and programs, and on the institution and perhaps society at large (Bachman & Palmer, 1996). Practicality, meanwhile, relates to the resources required by an instrument during whatever stages are relevant to the context, such as development, trialing, operationalization, administration, marking, and so on (*ibid*).

While not without its own limitations (discussed in the following section), the usefulness model addressed a considerable need and, in combination with its flexibility for the myriad contexts, purposes and types of evidence associated with language test use, quickly came to dominate empirical validation research in language testing (Xi, 2008). It has, for example, guided investigations into criterion-based measures and placement instruments (Bachman, 1996), performance-based assessments, such as writing samples

(Nakamura, 2004; Weigle, 2002) and both general ability and English for Specific Purpose instruments (Cellan, 2007).

2.3.3 Limitations of early validation frameworks

While the frameworks proposed by Kane (1992; Kane et al., 1999), Bachman & Palmer (1996), and others, were influential and important developments for validation studies, they were still fundamentally limited.

Xi (2008) points out that, while Bachman and Palmer's (1996) test usefulness structure assisted in establishing an evidence-based argument for the validity of a specific instrument, it did not provide "logical mechanisms to prioritize the six qualities and to evaluate overall test usefulness" and, subsequently, evaluations are "conveniently at the discretion of test developers and validation researchers" (p.179). Similarly, Kane's model also lacked such inherent logical structures and was open to, whether intentional or otherwise, selection of confirmatory evidence.

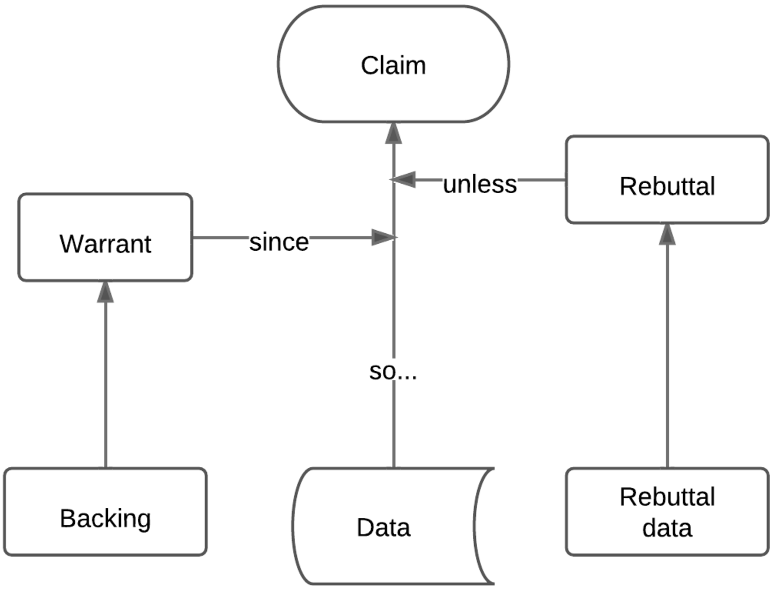
Further, while Bachman and Palmer's model included aspects of test use and consequences, Kane's (as yet) did not, and neither adequately addressed issues of test use (Bachman, 2005; Kunnan, 2003) which were increasingly gaining attention in the literature, such as ethical considerations (Lynch, 2001; Shohamy, 1997), fairness (Kunnan, 1998, 2000; Shohamy, 1998, 2001) and washback (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Shohamy, Donitsa-Schmidt, & Ferman, 1996).

Bachman (2005) suggests three major problems with any validation framework that does not address test use. First, even valid score-based interpretations provide no guarantee of the relevance, usefulness and sufficiency of the instrument for the projected use or decisions based upon its results. It is entirely possible for an assessment to be a valid indicator of the intended ability, yet used for a purpose it is not intended. Or, even if used for an intended purpose, to be an insufficient source of information for the "richness and breadth of abilities that are typical of content and performance standards, as well as representing the variety of ways in which teachers may interpret these standards" (Bachman, 2005, p. 14). Second, such frameworks provide no defense against the subversion of test-score interpretations for uses they were not intended. Third, a validity argument on its own offers no means for predicting or investigating the consequences of test use.

2.4 Toulmin's argument model

With regard to the limitation of validity arguments not utilizing a logical framework, a solution would seem to have been found with the adoption of Toulmin's (2003) structure for practical arguments. While certainly not new — Toulmin's model (Figure 2.2) had been in existence since at least the publication of *The Uses of Argument* some 45 years earlier (Toulmin, 1958) — the structure was first implemented in the realm of validity studies by Mislevy, Steinberg and Almond (2003) for their "evidence-centered design" (Mislevy et al., 2002, p. 3) framework addressing validity in the development of task-based language assessments.

Figure 2.2: Toulmin's Argument Model (adapted from Bachman, 2005, p.9)



Toulmin's model consists of 6 elements. Claims are the interpretations we intend to make from the observed data. The data is the information, such as a test score or the interpretation of test score. Warrants are the qualities of the data or methods used to establish the data which support the claim, and backing is the evidence, such as theory or empirical evidence justifying the warrants. Rebuttals, on the other hand, are alternative rationale which might refute our warrant(s) or claim, and rebuttal data is the evidence which supports this counterargument to our claim.

The use of Toulmin's structure assists in articulating an argument consisting of a chain of logical inferences — from test performance to interpretation and/or interpretation to decision, for example — explicitly laying out the assumptions upon which the inferences are based, and outlining both supporting and potentially refuting theoretical and empirical evidence regarding these assumptions. As such, it is extremely useful for articulating validity arguments, including associated justifications for validity claims and supporting evidence, while simultaneously outlining and addressing possible refuting rationales and evidence.

Following the lead of Mislevy et al., other influential researchers, such as Kane (2004), Bachman (2005), and Chapelle et al. (2004, 2008), similarly incorporated Toulmin's model into their own approaches. It would appear, then, that it has become the standard logical framework for validity investigations in educational and language assessment. Because it is the current standard in such studies, and due to the important benefits it adds to a validation framework, Toulmin's core concepts were implemented for the current investigation as well (see Chapter Four for further details on the theoretical and conceptual/validation framework of the study).

2.5 Kane's interpretive model addresses utilization

In later iterations of his approach, Kane (2001, 2002, 2004) included a fourth interpretive bridge, test use, and fifth component, decision, extending his model beyond the consideration of test score interpretation alone, to also include utilization, and thus offering a basis for developing arguments representing a more complete conceptualization of validity.

Figure 2.3. Kane’s Bridge Analogy Extended to Include Test Use

Inferences		*Scoring* ⇨		*Generalization* ⇨		*Extrapolation* ⇨		*Test Use* ⇨	
Components	Observation		Observed Score		Universe Score		Target Score		Decision

As noted when presenting earlier versions of Kane’s approach, the interpretive model has been a highly influential and widely applied framework for guiding the design and evaluation of assessments. Here we will discuss two recent examples of attempts to translate the interpretive model into language testing validation frameworks: Chapelle and her colleagues’ (2004, 2008, 2010) substantial work involving the Test of English as a Foreign Language (TOEFL), and Bachman’s (2005; Bachman & Palmer, 2010) Assessment Use Argument.

2.6 Translating Kane’s approach to language testing validation frameworks

While Kane’s interpretive argument framework certainly helped to ease the burden of validity investigations, and broadened the scope of potential researchers who could feasibly engage in such studies, constructing and evaluating an interpretive argument is still a complex and time-consuming task. As a result, there are few existing examples of comprehensive attempts to do so. One such example is the work of Chapelle, Enright, and Jamieson (2004, 2008, 2010), who, from 2000 to 2007, drew upon Kane’s model as they undertook the task of articulating and evaluating a validity argument for the Test of English as a Foreign Language (TOEFL).

2.6.1 Chapelle, Enright and Jamieson’s validation framework for the TOEFL

Chapelle, Enright and Jamieson (2004, 2008, 2010) established six inferences, with corresponding warrants and assumptions requiring evidence to support them: i) domain description; ii) evaluation; iii) generalization; iv) explanation; v) extrapolation, and vi) utilization. In addition to Kane’s five inferences, then, an additional element, domain description, has been added. However, as, Chapelle and her collleagues point out, Kane (2004, p. 141) does state that “if the test is intended to be interpreted as a measure of competence in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain (in terms of content, cognitive level, and freedom from potential sources of systematic errors) tend to support the intended interpretation". The inclusion of the inference in the argument, then, if adequately supported, offers justification for the position that observations of performance on the TOEFL provide evidence of relevant knowledge, skills and abilities in situations that are representative of target language use in the academic institutions using the instrument, and in which the examinees will be required to function in the target language (Chapelle et al., 2010).

Through seeking out supporting evidence for each of the assumptions relating to their warrants, Chapelle and her colleagues were able to develop a comprehensive (though perpetually evolving) argument for the validity of the TOEFL instrument, with incremental inferences and accompanying warrants and evidence, leading from the target language use (TLU) domain all the way through test score interpretation and continuing on to utilization and consequences of test use for various stakeholders.

While comprehensive, however, the validity argument produced by the researchers only incorporates confirmatory evidence, and does not include any mention of even potential rebuttals or refuting theory or evidence. Chapelle et al. (2010, p. 11) acknowledge this, stating "[d]evelopment of the TOEFL validity argument thus far has not yet reached the stage at which we can evaluate its utility for discussion of… rebuttal[s]." While some might suggest there is no reason potential rebuttals and refuting evidence must wait until after warrants and confirmatory evidence have been addressed, the researchers have done a great service to the myriad stakeholders affected by use of the TOEFL instrument. By creating and sharing this validity argument, they have established a common framework for the investigation and discussion of the various underpinning bases, and resultant overarching argument for, the validity — both in terms of test score interpretation and test utilization and impact — of one of the most widely used high-stakes assessment tools in the world.

Further, Chapelle et al.'s work, and that of others, into the ongoing process of articulating Kane's model, represents important advancements in the continuing efforts to transform Messick's concept of unitary validity into a practical, argument-based approach to validation (Mislevy et al., 2002, 2003; McNamara, 2006; Xi, 2008). Such efforts demonstrate that the interpretive argument approach is extremely useful for:

i) approaching the validation of instruments with extremely complex constructs, such as "academic English language proficiency", and still being able to articulate clear and compelling arguments about the relative strengths and weaknesses of various aspects contributing to the overall validity of the instrument and its use;

ii) identifying assumptions in need of support or further investigation in the argument for the validity of the instrument; and

iii) providing guidance which can help increase the likelihood of articulating a comprehensive validity argument, attending to all established inferences connecting observed performance through to score interpretation and continuing on through test utilization and consequences and, in doing so, reducing the likelihood of the considering only confirmatory or pre-existing evidence alone.

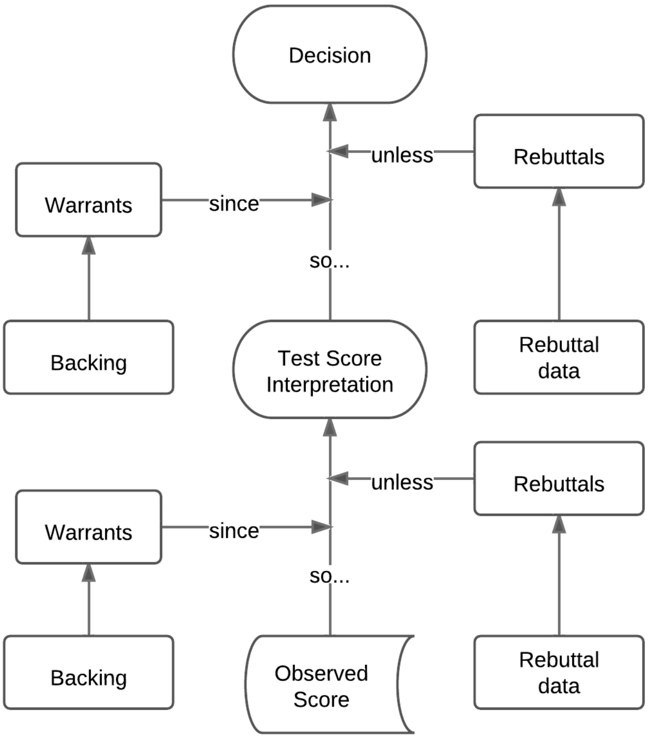
However, Messick's work established the centrality of *both* test score interpretation *and* test use to considerations of validity. Even in their modern iterations, frameworks like Kane's still focus primarily, or at least could be argued to provide far more structured guidance when it comes to, score interpretation rather than test utilization. As Bachman (2005, p. 4) puts it, "Although the fields of language testing and educational and psychological measurement have been discussing test use and the consequences of test use for well over a quarter of a century, neither field has, as yet, developed a comprehensive set of principles and consequences of test use". As an initial step towards addressing this void, to providing a more balanced means of establishing clear links from test performance to interpretations and also to utilization under a single meta-structure, Bachman (2005) presented the Assessment Use Argument.

2.6.2 Bachman's Assessment Use Argument

Bachman (2005), like Mislevy et al. (2003) and Kane (2004), implements Toulmin's (2003) logical argument structure. Also similar to Kane, Bachman proposes two parts to the validity argument, the descriptive interpretation based on the results derived from the instrument in question, and the "decision-based interpretation" (Kane, 2002, p. 32), in which a conclusion is reached about the examinee based on the descriptive interpretation. Bachman (2005, 2007), however, provides a more extensive, structured approach to test utilization than has been previously offered by other frameworks. In his original iteration of the model, presented in Figure 2.6, the Assessment Use Argument (Bachman, 2005) is a two-part structure with a validity argument establishing a link between test performance and interpretation, and a second argument connecting test score interpretation to test utilization. In doing so, Bachman essentially extended Kane's decision inference into an entire, second argument attending to utilization issues.

Bachman suggested an initial four areas, discussed below, requiring justification for the utilization argument. The first three come from the third and fourth cells of Messick's (1989) progressive matrix: relevance, utility and consequences. The fourth attends to the sufficiency of the instrument for informing the decisions made about test takers.

Figure 2.6: Bachman’s Assessment Use Argument (adapted from Bachman, 2005 p. 25):



Type 1 Warrants — Relevance:

Relevance warrants justify the position that the score-based interpretation is relevant to the decision being made. In the case of placement tests, for example, the construct defined by the publishers may be well established and the evidence offered as supporting claims that the assessment does measure this ability may be convincing (e.g., College Board, 2003). However, this does not guarantee that such a construct is relevant to the TLU associated with the curricula and assessments students will encounter once enrolled in the program in question (Alderson, Clapham, & Wall, 1995; Burt & Keenan, 1995; Heilenman, 1983; Lytle & Wolfe, 1989; Wrigley, 1992). For a placement test, then, it is important that the tasks in the assessment engage the competencies and processes students will be required to implement in the instructional domain.

Type 2 Warrants — Utility:

Utility warrants establish confidence that the score-based interpretation is actually useful for informing the intended decisions. Essentially, to what extent does the score-based interpretation provide information that increases the likelihood of making the right decision or avoiding decision errors?

Bachman (2005), for example, describes the use of a multiple-choice test of general English ability which might not be considered authentic, and perhaps even only marginally relevant to a particular TLU domain, but which may turn out to quite accurately predict future performance at a particular job or within a specific level of a language program. What has been traditionally referred to as predictive validity (Bachman, 1991; Raatz, 1985; Shepard, 1993), then, can offer valuable insights and evidence regarding this area of a validity argument.

Type 3 Warrants — Intended Consequences:

These warrants refer to the consequences of using the assessment under investigation, and whether or not resulting decisions are of benefit to the test-taker, program, institution, and other stakeholders (Bachman, 2005). This area of enquiry could incorporate such areas of test use discussion often outside the direct realm of validity arguments as Critical Language Testing (Shohamy, 2001), fairness (Kunnan, 1998, 2000, 2003; Shohamy, 1998, 2001), testing ethics (Lynch, 2001; Shohamy, 1997), washback (Alderson & Banerjee, 2002; Alderson & Hamp-Lyons, 1996; Bachman, 1991; Shohamy, et al., 1996) and even practicality (Bachman & Palmer, 1996).

Type 4 Warrants — Sufficiency:

Because an assessment is indicated to be relevant and useful does not mean it, alone, adequately accounts for the full breadth and scope of the competencies and characteristics of the test taker that are relevant to the decision(s) being made (Bachman, 2005). As such, a sufficiency warrant is necessary to establish whether or not the instrument(s) involved provide adequate representation of the evidence necessary in order to make appropriate decisions about the individuals involved. When addressing such a complex, composite construct, such as factors which influence performance in a language course or program, "[t]his is, in essence, the argument in favor of multiple indicators" (Bachman, 2005, p. 19).

Justifications for test use relating to these issues, particularly those relating to consequences and sufficiency, offered a much-needed means of integrating essential elements in the testing discussion such as washback, critical language testing theory, ethics and professionalism, and fairness in testing (Xi, 2008).

2.6.2.1 Advantages of the AUA

By dedicating an entire second argument focusing on various aspects of test utilization, and the assumptions and associated evidence regarding test use and consequences, Bachman offers a validation model that more comprehensively articulates the decision-based interpretation component of Messick's unitary validity. It also allows for the direct incorporation of numerous, divergent issues in test use which, until now, have not traditionally been directly incorporated into validity investigations, such as ethics, fairness, and washback, for example. This is particularly important in a time with increasing use of (and costs associated with) assessments, particularly high-stakes, standardised instruments used to make important decisions affecting test-takers, programs, institutions and a host of other stakeholders (Bachman, 2005, 2006, 2007). Further, it offers an argument meta-structure which assists articulating and investigating the links between test score interpretation and utilization, or descriptive and decision-based interpretation as Kane refers to them. This allows test designers and researchers to consolidate design, development, score interpretation and intended use within a single comprehensive structure.

2.6.3 Bachman & Palmer's (2010) current iteration of the AUA

In their recent book, Bachman and Palmer (2010) present a remodeled AUA in which they have merged the test score interpretation and utilization arguments into a single, overarching validity argument comprised of 4 claims: assessment records, interpretations, decisions, and consequences. For each claim, Bachman and Palmer offer one or more assumption requiring theoretical and/or empirical support in the establishment of a compelling validity argument, presented below:

1. Consequences are to be beneficial
2. Decisions are to be: values sensitive; equitable
3. Interpretations are to be: meaningful; impartial; generalizable; relevant; sufficient
4. Assessment records are to be: consistent

They further offer suggestions for numerous warrants/rebuttals and relevant supporting/refuting evidence, that could prove valuable for test users and researchers alike in the evaluation and/or design of assessment instruments.

While the 2010 AUA may well prove useful for many test developers and users, none would appear to have attempted to implement it as yet. This is not unsurprising, perhaps, given its recent publication and the complicated and lengthy processes that argument-based validation studies require.

2.7 Theoretical framework for the current study

As a more detailed presentation and discussion of the logical and theoretical models adopted will be presented in Chapter Four, only a brief summary will be provided here. For the current investigation it was decided that a hybrid of both Kane’s interpretive model and Bachman’s original (2005) AUA would form the validation framework. This decision was made largely for two reasons. First, because it has been so widely used, discussed and influential, particularly in assessment endeavours outside of language education, Kane’s model was felt to allow for greater ease of communication with stakeholders at the institution, the vast majority of whom have only recent and developing experience and expertise in the area of educational assessment. Warrants relating to evaluation, generalizability, and extrapolation inferences, for example, would seem to be readily parceled and explainable as *scoring, reliability* (or consistency), and *relevance* to student learning in CMI courses; all issues and terminology already in use and under discussion in the college as a result of recent, ongoing quality assurance endeavours.

However, Bachman's (2005) AUA utilization argument was used to replace Kane's decisions inference, in the hopes of benefiting from the greater focus and structure relating to test use issues provided by the AUA model.

2.8 Closing comments

As we have seen in this chapter, considerable advances in the study of validity, and in the development of logical and conceptual frameworks which can help test designers and users alike incorporate these advances in validation studies, have developed rapidly in recent decades. Yet, despite widespread acceptance that such assessment validation is necessary for meaningful and ethical test use, and despite increasing use of standardized instruments to make decisions which significantly impact individuals, institutions, and educational systems throughout the world, many in the literature lament a chronic lack of these essential validation efforts (Bachman, 2005, Kunnan, 2003, Xi, 2008). The following chapter, which reviews recent empirical investigations, will start with a review of some examples of studies which have made contributions to filling this longstanding void of the practical application of validity frameworks. Subsequently, the chapter will present existing discussions and empirical research relevant to the various elements included in this study's validation argument – evaluation, generalizability, extrapolation, sufficiency, and consequences – as relates to Accuplacer Companion (and other standardised tests) and locally marked placement essays.

CHAPTER THREE

Literature Review: Empirical Investigations Relevant to the Current Study

3.0 Introduction

This chapter will review empirical studies relevant to the current inquiry, presented in two sections. The first will focus on recent investigations utilizing one of the two modern validity frameworks – Kane's interpretive model and Bachman's Assessment Use Argument (AUA) – which form the basis of the validation structure for this study. The second section of this chapter will review previous empirical research relevant to various elements included in this investigation's framework – evaluation, generalizability, extrapolation, sufficiency (part of the decisions claim), and consequences – as relates to standardised tests, such as Accuplacer, and locally marked writing samples used for the purposes of placing incoming students in English language programs in higher education contexts.

3.1 Empirical applications of modern validation frameworks

As detailed in the previous chapter, considerable advances in the study of validity, and in the development of logical and conceptual frameworks to help test developers and users alike apply these advances in validation studies, have occurred rapidly in recent decades. This section of the chapter will review some examples of recent studies that have made contributions to filling this longstanding void of the practical application of validity frameworks.

3.1.1 IELTS validity argument

Aryadoust (2011), offers his efforts at establishing and evaluating a validity argument for the Speaking and Listening modules of the International English Language Testing System (IELTS). Using Kane's model, he endeavoured to establish one or more assumptions relating to each of the inferences in the interpretive argument, for a total of seven claims requiring warrants and backing evidence. For example, relating to the inference of generalization, Aryadoust articulated the assumption that observed scores mirror expected scores across parallel test/task versions, established the warrant that G-theory and reliability investigations establish the generalizability of the particular module, and then went about searching the literature for relevant evidence supporting or refuting the warrant.

By collecting and analysing a total of 28 independent research studies and publications detailing revisions to the instruments, each relevant to one or more of these seven assumptions, Aryadoust established an interpretive argument with warrants, rebuttals and supporting/refuting evidence for the Speaking Module of the IELTS.

Through utilizing an argument-based validation framework to guide the collection and analysis of evidence, Aryadoust is able to establish a rather comprehensive, integrated, web of evidence from which he is able to reach important conclusions regarding:

i) strengths and weaknesses in the test score interpretation element of the validity argument for the instrument modules, including apparent theoretical issues with the explanation inference and empirical evidence largely refuting the extrapolation inference for the speaking module;

ii) issues regarding utilization, such as research suggesting decision-makers often have a limited understanding of the test or what its scores imply about the test taker; and

iii) the current state of research regarding various areas of test score interpretation and utilization for each module, including a dearth of evidence regarding the listening component substantial enough that he gave up on this aspect of his study.

While Aryadoust's study is different from the current investigation in many ways – it does not look at the validity of the assessment(s) in a specific context, for example – it does provide important insights for the current study. It demonstrates the value of argument-based validity frameworks for guiding comprehensive studies which can result in important insights into areas of strength of an instrument, areas in which action needs to be taken to improve functioning and usefulness, and provides an initial framework to which others may make contributions through further study and initiatives for improvement, all of which are hoped for with the current work at CMI. It also, however, demonstrates the limitations of Kane's model, in that utilization is a single inference, leaving an imbalance between consideration of the meaning of test scores and their utilization. It is in the hopes of avoiding such an imbalance that the current study opted for a hybrid framework combining Kane's model and Bachman's (2005) AUA (to be described in Chapter Four).

3.1.2 Validation of a Spanish listening test

Pardo-Ballester (2007, 2010) looked to utilize validity frameworks to guide the design and evaluation of a listening test used to inform the placement of students in a university Spanish language program. Pardo-Ballester did this in two parts. First, she used Bachman and Palmer's (1996) test usefulness framework to establish and investigate a 'validity' (i.e., test score interpretation) argument for the instrument. Pardo-Ballester established a warrant and potential rebuttal for each of the six qualities of test usefulness in the model – reliability, construct validity, interactiveness, authenticity, impact, and practicality – and identified evidence to be collected and analysed to inform the evaluation of each warrant/rebuttal. Subsequently, she used Bachman's (2005) AUA to guide the establishment and evaluation of warrants/rebuttals, and backing/rebuttal evidence relating to four areas of test utilization: relevance, utility, intended consequences, and sufficiency.

The established warrants and rebuttals were investigated utilizing a variety of types and sources of evidence, including: the solicitation of test-takers'/students' and raters' opinions; an analysis of the scoring procedure, including descriptors for the rubric; and a review of the procedures used to establish cut scores.

The established claims, warrants, rebuttals and evidence outlined helped guide a comprehensive and transparent validation of the instrument, which included numerous concerns often expressed but not always investigated regarding placement instruments, such as consequences on teaching and learning, and stakeholder perceptions of the instrument.

Pardo-Ballester's study informed the current investigation in a number of ways. First, it served as an example of the possibility and benefits of utilizing the AUA as a basis for developing a framework attending to elements of test utilization long-overlooked in validation studies. Further, as with Aryadoust's study, use of an argument-based validation framework was demonstrated to help ensure the evidence considered and methodologies used were varied and comprehensive, and the established AUA for the instrument can now serve as the basis for future, ongoing investigations into the validity of this assessment, and perhaps others used by the program.

However, the selection of Bachman and Palmer's test usefulness model as a framework for the score interpretation portion of the study could be argued to have resulted in a somewhat disjointed overall validation argument. While the qualities of reliability, construct validity, authenticity and interaction certainly address test score interpretation, the two remaining elements of the model, impact and practicality, attend to utilization issues. Yet, Pardo-Ballester has opted to keep these as warrants in the test score interpretation section of the validity argument, even though both would seem readily addressable in the consequences warrant of the utilization argument. Perhaps this has been done in order to maintain focus on impact and practicality during the design phase of the listening assessment involved in the study. However, for the current investigation, it was felt such overlapping foci of investigation in the score interpretation and utilization sections of the validity framework would be unnecessary, and possibly confusing. In order to avoid this, Kane's interpretive inferences were adopted as the warrants for the score interpretation half of the validation framework and a modified version of Bachman's AUA was used for the utilization argument.

3.1.3 Validation of a standards-based classroom assessment of English proficiency

Llosa (2008) reports her investigation regarding the performance and impact of a standards-based classroom assessment of English proficiency based on teacher judgements. Using Bachman's (2005) AUA as her guide, she created a validation framework, including claims, warrants, potential rebuttals and evidence needed to justify (or refute): links between teachers' scores on the English Language Development (ELD) Classroom Assessment and the interpretations made about students' language ability; and, subsequently, links between these interpretations and decisions made about students as a result.

Llosa then uses the findings from two previous studies, one quantitative (comparing results from the ELD Classroom Assessment and the standardised California ELD Test) and one qualitative (verbal report protocol to investigate the process instructors/raters implemented when judging the ELD Classroom Assessment) to inform the evaluation of the warrants and rebuttals regarding teacher judgements.

Llosa's work, then, further demonstrates some important benefits to the use of a validation framework. As she concludes herself (2008, p. 40), the implementation of Bachman's "AUA provides a coherent framework that allows for a comprehensive examination of all warrants and potential rebuttals in order to justify interpretations and decisions". Further, as Llosa's investigation uncovered a number of problematic areas in the performance of the assessment, such as inconsistencies in the comprehension and application of rubric criteria by raters, it also shows the opportunities for identifying and attending to specific issues impeding the performance of the assessment and positive outcomes intended from its use.

3.1.4 Validation of a British Sign Language assessment

Implementing Bachman's (2005) AUA, Mann and Marshall (2010) describe their work to validate an instrument of their own creation, the Nonsense Sign Repetition Test

(NSRT), intended to assess deaf children's skills in British Sign Language and help inform decision on the type of intervention best suited to their language needs. Oddly, Mann and Marshall establish only a single test score interpretation argument, that the assessment "tap[s] into deaf signing children's phonological working memory" (Mann & Marshall, 2010, p. 253), and counterclaim, which considers the possibility hearing children (with no previous contact or training regarding BSL) may do as well as deaf children. The utilization argument is more robust, with warrants (and rebuttals) intended to justify (or not) claims of the assessment's:

relevance – to phonological working memory

utility – as a predictor of wider language ability

sufficiency – as an evaluation of children's BSL skills; and

consequences – of the support services selected as a result of the instrument scores.

The authors then drew upon data from their previous research regarding the NSRT to inform the AUA. In the pre-existing study, the authors found that deaf children outperformed hearing counterparts on the NSRT. For the AUA, they suggest this data supports the lone test score interpretation warrant and the first claim of the utilization argument, that the NSRT task is relevant to phonological working memory. They again use this data to support the utility claim. With regard to the consequences claim, they suggest that the NSRT provides practitioners information regarding deaf children's sign language proficiency, and, therefore, results in greater likelihood of beneficial interventions being prescribed. The final claim, sufficiency, is also deemed supported by their previous research finding that the NSRT results correlate with those of the BSL receptive skills test and, they suggest, would seem to address the same underlying construct -- phonological working memory.

Mann and Marshall conclude that the AUA offers both a transparent framework for researchers developing language assessments and a means for more beneficial consequences of test use for all stakeholders through the guided focus on both interpretation and utilization aspects of assessment. While the work of others, such as Aryadoust, Pardo-Ballester, and Llosa, has suggested these conclusions for the use of Bachman's AUA, Kane's interpretive model, or modern argument-based validation frameworks in general, may well be accurate, problems with Mann and Marshall's study perhaps leave the veracity of these conclusions based on their work in question.

First, one of the hopes of utilization of a validity framework is that it guides researchers towards comprehensive evaluations of the score-based interpretation and utilization of assessments and, further, safeguards against the selection of confirmatory evidence alone. The authors, however, do not mention the limitations of their work which results from their consideration of only one source of evidence -- their own research -- which they use to confirm every one of their claims for its validity and usefulness. Their work is further limited in that they use the same evidence to support more than one claim. While this is not unfeasible, it does point out the limitations of the range and types of evidence they have reviewed for their validation framework (in its first iteration, at least).

As such, while Mann and Marshall are to be commended for their efforts in establishing a validity argument framework for the NSRT, upon which they and others may expand and eventually attend to the problems mentioned above, it served as a reminder for the current investigation that the use of a validation framework, on its own, does not guard completely against the selection of confirming (or disconfirming) evidence alone, or the articulation and selection of warrants and evidence that lead to a comprehensive validity argument.

3.1.5 Final comments regarding reviewed investigations

The limited number of validation studies investigated here is representative of a number of things. First, the ongoing dearth of validation investigations that would appear to be occurring and/or made available to others either in the literature or elsewhere. Second, that while validation argument frameworks certainly would seem helpful for guiding researchers towards transparent and comprehensive validity investigations, the potential for confirmatory investigations still exists. Finally, they represent the important value that such studies can present to myriad stakeholders in the growing business of testing. By establishing frameworks for the investigation of instruments used to make high-stakes decisions about test-takers, researchers help to establish a common structure facilitating the collective investigation and evaluation of various assessments, highlighting their strengths as well as the areas of weakness which, once identified, can be addressed, either by test developers or test users.

Having reviewed recent examples of the empirical application of the two modern validation frameworks adopted for the current study – the interpretive model and the assessment use argument – the following section of this chapter will discuss current theoretical positions and empirical evidence relating to various aspects of the two types of assessment used at CMI to inform placement decisions: Accuplacer Companion (and other standardised, multiple-choice instruments) and locally judged writing samples.

3.2 Research on standardised tests and writing samples as placement instruments

The purpose of a placement exam is, of course, to assess an incoming student's competencies and match them with the most beneficial course(s). While the concept is straightforward, it would seem that the vast majority of higher education institutions struggle with the process (Armstrong, 2000; Hughes & Scott-Clayton, 2011; Klee & Rogers, 1989; Sullivan, 2008). This segment of the chapter will address previous research on placement testing. More specifically, we will address the research and discussions in the literature regarding the relative strengths and weaknesses of standardised, multiple-choice instruments and locally developed and marked writing samples for placing students in language courses. The inclusion of instruments other than Accuplacer in this review was felt necessary due to the apparent scarcity of investigations into Accuplacer in the literature, or made available by some other means, such as institution websites.

The discussion will be structured along a number of claims, or components of claims, of the current study's validity framework: evaluation, generalizability, extrapolation, sufficiency and consequences. In addition to fitting the claims and warrants of the current study's validation structure, such an organizational arrangement allows us to attend to many of the contentious issues within the ongoing debates surrounding standardized exams and local performance assessments.

3.2.1 Evaluation

For standardised tests, we might expect a considerable amount of resources invested by publishers in assuring such factors as confusing instructions, errors in answer keys, or problematic test items or answer options would have been corrected during extensive piloting and review of the instruments. A review of the literature would seem to suggest little concern regarding errors in test items or problems with standardised instruments' instructions, characteristics, or scoring procedures. However, at least one instance of a standardised placement instrument having answer key issues has occurred in recent past (California Community Colleges Assessment Association, 2008), reminding us that test users need to be vigilant of such issues and thoroughly review all tests used which impact their constituents.

One area of considerable attention of late, particularly since the No Child Left Behind Act mandated high-stakes test use throughout the US public education system, is the use of instruments designed for native speakers of English with examinees who are English Language Learners (ELL's). If the instructions, items or answer options are unclear or misunderstood by the examinee, this is almost certain to influence outcomes, and brings the evaluation inference and overall validity of the instrument (when used with ELL's) into doubt (Crawford, 2004; Solórzano, 2008). This is a particular concern for CMI and its use of Accuplacer Companion, as it is designed for use with students for whom English is a ‘best language’, while the college serves an almost exclusively ELL population.

With regards to placement essays, Kane (2004, p. 156) points out that “scoring of essay questions and performance tasks is more judgmental than that of objective tests and therefore requires additional backing for its dependability (e.g., interrater reliability)” (p.156). Therefore, inter- and intra-rater reliability are particular concerns for the evaluation inference when it comes to judged assessments, and both are certainly aspects of the ongoing debate regarding the use of writing samples for informing placement decisions. However, as inter- and intra-rater reliability also relate to the generalizability inference, discussion of previous research regarding these issues will occur in that section of the chapter.

One thing that should be touched upon before moving on, however, is the necessity for establishing the criteria used in the scoring rubric for judging the writing sample is relevant to the construct we are attempting to assess. A criterion focusing the rater on the writer’s use of contractions, for example, when the essay is meant to be an indication of the test taker’s formal writing skills (and therefore not to contain contractions) is introducing variance in the instrument results irrelevant to the target construct. While many in the literature discuss the issue of scoring criteria, most offer warnings of the dangers of inappropriate criteria or include caveats such as “assuming the… criteria were valid indicators of writing skills” (James, 2006, p. 6) before presenting their results. It would appear few studies have actually attended to scoring rubric criteria relevance (or, if they have, they have not been made available in the literature or via the web).

3.2.2 Generalizability

With regard to generalizability, the evidence reported in the literature suggests very high estimates of consistency for standardised tests such as Accuplacer (College Board, 2003; Mattern & Packman, 2009) or IELTS (Cambridge ESOL, 2007); levels which writing sample results may be unlikely to approximate simply because of their open-ended nature and the subjective evaluation of individual raters (East, 2009; Haswell, 2005; Jonsson & Svingby, 2007; Lumley, 2002; Rezaei & Lovorn, 2010; Stemler, 2004). The question, then, is whether intra-rater and inter-rater consistency are of such concern that we must question the validity and usefulness of writing sample results.

While at least some studies suggest internal consistency can be problematic, with a single rater's opinion potentially differing from one time and context to another (Greenberg, 1992), most intra-rater reliability studies conclude that consistency levels were acceptably high (Brown, Glasswell, & Harland, 2004; Jonsson & Svingby, 2007; Park, 2004). With regard to consistency between raters, reports range from unacceptably low (Brown, et al., 2004; Jonsson & Svingby, 2007) to reassuringly high (East, 2009; James & Templeman, 2009; Jonsson & Svingby, 2007; Matzen & Hoyt, 2004). Park (2004) conveys the results of his own research and a review of others' that also used multi-facet Rasch measurement (MFRM) to investigate rater reliability, and concludes that trained raters using a common rubric may differ in severity, but they do so consistently, and rater reliability need not be a validity-threatening issue when such best practices are followed.

As Hamp-Lyons (2007) points out, the concerns about the reliability of rating in writing assessment are usually concerns about the reliability of raters, and it would seem that when steps are taken to address this, such as rater training, the use of rubrics, common benchmarks of performance, and so on (East, 2009; Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010; Weigle, 2002), the reliability of writing sample assessment can be maintained at levels demanded by high-stakes assessment contexts.

3.2.3 Extrapolation

While the publishers of various instruments used by higher education institutions to place students invest considerable resources into ensuring their instruments assess the construct intended (Bejar & Jamieson, 2000; Butler, Eignor, Jones, McNamara, & Suomi, 2000; Cambridge ESOL, 2007; College Board, 2003; Cumming, Kantor, Powers, Santos, &

Taylor, 2000; Sawaki, Stricker, & Oranje, 2008; Stoynoff, 2009), this alone is no guarantee that the competencies assessed are the same as those required in order to achieve success in a particular language course or program (Bachman, 1996, 2005; Bachman & Palmer, 1996; Behrman, 2000; Sawyer, 2007). As Hughes and Scott-Clayton (2011) point out, this is an area of fundamental importance for placement decisions, and of far too little attention from test publishers, researchers, and particularly test users, to date.

Thorough matching of standardised instrument tasks and items to course activities and outcomes would seem to have been rarely, if ever, attempted (and made available) by institutions or researchers. Many have reported on stakeholder, particularly faculty, opinion, however, and the overall perspective would seem to clearly regard authenticity as a weakness of standardised tests used for placement.

Some point to the derived (Williams, 1990) and necessarily general (so as to be usable at a wide range of programs) (Behrman, 2000) nature of the test items, and a lack of reflection of what will be expected of students once they start their classes (Armstrong, 2000; CCCAA, 2007a, 2007b, 2008). Of particular concern, with regard to estimating students' writing ability, is the widely held perception of construct underrepresentation (Engemann & Gallagher, 2006; Hebel, 2001; Hillocks Jr, 2002, 2003; Moss, 1994; Williams, 1990), that limited-response instruments fail to address, for example: "cognitive and reflective processes involved in creating a text — such as making plans for writing, generating and developing ideas, and making claims and providing evidence" (Murphy & Yancey, 2008, p. 450).

Murphy and Yancey further argue that the artificiality of multiple-choice instruments as indices of writing skills is significant enough to impact the validity of their results, as construct-extraneous factors such as familiarity with and obedience to testing rules and formats, test-taking strategies, and so on can considerably influence outcomes.

These and similar concerns related to construct and relevance issues were named as some of the reasons 28 out of 31 community colleges surveyed by the California Community College Assessment Association (CCCAA) reported dissatisfaction with their placement test of choice — typically Accuplacer or Compass (Hughes & Scott-Clayton, 2011; Sullivan, 2008) — and all 60 representatives from member institutions voted unanimously to pursue the creation and adoption of an alternative instrument for California community colleges (CCCAA, 2007a, 2007b).

Most faculty members, and researchers as well, would seem to agree with Hughes (2003, p. 32), that "when in doubt, where it is possible, direct testing of abilities is recommended". Performance assessments, with their open-ended nature and direct assessment of abilities are argued to better reflect the complex cognitive processes and competencies required of students in the courses into which they are being placed (Cohen & Brawer, 1987; East, 2009; Jonsson & Svingby, 2007; Matzen & Hoyt, 2004; Messick, 1989; Stoynoff, 2009). That the writing performance of students is assessed by faculty members who teach in the program, those in best position to evaluate examinees' writing in context of the use and purpose of the performance, is argued by some to further enhance the authenticity and value of the process (Jonsson & Svingby, 2007; Moss, 1994; Swain, 1993).

The prevailing opinion, then, would seem to be that communicative, performance assessments, like writing samples, are more representative of real-world and/or instructional domain demands, and this has lead to a substantial push for such instruments in L2 and college placement testing (Kim, 2008; McNamara, 1996). However, as Murphy and Yancey (2008) remind us, it is important to remember that in order to ensure writing samples are as authentic as possible, they must be designed to reflect the type of writing task students will be required to engage in once in their courses in the program in question. Further, we cannot discount that students will likely encounter multiple-choice, limited response, indirect assessments of their skills and knowledge in their courses, and that their performance on such instruments will influence their course results.

In order to address this important issue further, results from a number of investigations into the performance of standardised tests and/or writing samples were reviewed, and will be presented in two categories: i) those addressing the convergence of instrument results with expert opinions as to candidates' abilities relevant to the courses into which they are being placed; and ii) the predictive capacity of the instruments in anticipating student performance.

3.2.3.1 Convergence with expert opinion

While concerns about the reliability of subjective assessments of student abilities and performance are well warranted (Armstrong, 2000; Behrman, 2000; Sawyer, 1989), we must also consider the argument that "the most credible judges are those who are most knowledgeable about the context in which a performance occurred, and about the nature [and purpose] of the performance itself" (McLeod et al., 2005, p. 462; Moss, 1994; Murphy & Yancey, 2008; Swain, 1993). Evidence from James and Templeman (2009) would seem to support this position, as estimates of accurate student placement improved from 44% to 66% (depending on the course and subtest(s) involved) if based entirely upon Accuplacer results, to 81% to 84% if the "faculty factor" (p. 82) — locally marked writing samples, oral interviews, and instructor interpretations of students' results on all assessments — was involved.

If we accept, then, that instructors are experts in the competencies required of students in the courses they teach, and recognize the considerable influence they have over students' grades (Armstrong, 2000), then agreement between their opinions of students' abilities, in-class performance, and results of placement instruments is an important source of evidence in support of the utility of the procedure. However, few studies would seem to incorporate this source of evidence.

Only two studies (Cabrillo College Office of Institutional Research, 1999; College of the Canyons, 1993) were found comparing placement test results and instructors' opinions and both had either methodological limitations or extremely low participant numbers, leaving the findings in doubt. Only two English courses involved in the College of the Canyons' (1993) study, for example, had more than 10 participants. Results were mixed, with the Assessment and Placement Services for Community Colleges (APS) Reading subtest demonstrating low correlation (r=.31, corrected for restriction of range, n=61) with teachers' opinions of students' abilities, and the multiple-choice Writing subtest showing moderate correlation (r=.57, corrected, n=22).

In the only two studies found to report agreement between writing sample results and later instructors' opinions of student placement, Wall, Clapham and Alderson (1994) found moderate correlation (r=.47) between the two assessments and May (2007) found essay results were the biggest contributor in a model predicting instructor opinion of student placement.

While many (Clapham, 2000; College Board, 2003; Hillocks Jr, 2002, 2003; Hughes & Scott-Clayton, 2011; Spolsky, 1997) advocate the inclusion of writing samples on rational and empirical grounds, it would seem more evidence is needed in order to establish whether their use results in greater agreement with expert opinions of students' abilities or appropriacy of their placement in English programs.

3.2.3.2 Convergence with course results

While occasional instances of moderate predictive capacity for course results are reported for standardised tests (Hill, Storch, & Lynch, 1999; Kerstjens & Nery, 2000), the vast majority report low predictive validity (Armstrong, 2000; Behrman, 2000; Cohen & Brawer, 1987; College Board, 2003; Denham & Oner, 1992; Gabe, 1989; Goodman, Freed, & McManus, 1990; Hill, et al., 1999; Holderer, 1992; Hughes & Nelson, 1991; Isonio, 1991, 1992; James & Templeman, 2009; Rasor & Barr, 1995; Sullivan & Nielsen, 2009).

Looking only at those instruments specifically designed for placing college students in English programs, findings from a number of studies suggest tests like Accuplacer, Compass, and their various versions and subtests, account for 1% to 16% (r^2=.01 to .16) of variance seen in students' final results in credit-level, remedial, and ESL English courses (Armstrong, 2000; College Board, 2003; Sullivan & Nielsen, 2009). Even when statistically correcting for problem areas in predictive validity studies, like sampling error, restricted range of course results, and limited reliability of factors such as course grades, meta-analysis results still suggest a limited range of 5% to 22% of course result variance accounted for (Mattern & Packman, 2009).

With regard to the performance of writing samples as placement instruments, while May (2007) found writing sample scores did not significantly predict course outcomes, and Mathay (1992) found mixed results, most research studies reviewed would seem to suggest locally marked essays produce better placement accuracy than the various standardised instruments involved (Garrow, 1989; Holderer, 1992; Matzen & Hoyt, 2004; Wolcott & Legg, 1998; Zinn, 1998).

Others have found that the best predictive model for student course outcomes occurred when a combination of standardised test and writing sample scores was used (Breland, Camp, Jones, Morris, & Rock, 1987; Galbraith, 1986; Garrow, 1989; Isonio, 1991, 1992; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998).

However, if we were to consider recent research alone, only the two conflicting accounts of Matzen and Hoyt (2004) and May (2007) on the predictive capacities of writing samples remain. The question could be asked, then, if writing samples still hold the possible edge in, or still contribute significantly to, predicting student performance relative to modern versions of standardised tests used for placement.

3.2.4 Sufficiency

Many, if not most, US 2-year colleges would seem to place students based on the results of a single, standardised, multiple-choice test (Hughes & Scott-Clayton, 2011; Sullivan, 2008). As Haertel (1999, p. 6) notes, "wide-ranging inquiry… will run afoul of short timelines and tight budgets”. However, given the broad spectrum of factors known or believed to influence student success in higher education, including English or other language programs, it is difficult to imagine being able to argue for the sufficiency of a single instrument, particularly given the high-stakes nature of the decisions being made.

This would seem to be the conclusion drawn by nearly all educational assessment researchers, organizations, and test publishers as well, advocating the consideration of a variety of types and sources of evidence for placement and other high-stakes decisions (AERA, APA & NCME, 1999; Bachman & Palmer, 1996; Board of Governors of the California Community Colleges, 2008; Camara & Lane, 2006; College Board, 2003; Lynch, 2001; Solórzano, 2008). While well-designed and appropriately selected standardised instruments can offer valuable information for differentiating students based on language and/or academic abilities, several accounts report problems associated with their use alone when placing students (Armstrong, 2000; Belcher, 1993; College of the Canyons, 1996; Garrow, 1989; Jones & Jackson, 1991; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998). With regard to the potential contribution towards sufficiency that including a writing sample might offer, a number of studies would seem to indicate the best predictive model for student course outcomes occurred when a combination of standardised test and writing sample scores was used (Breland et al., 1987; Galbraith, 1986; Garrow, 1989; Isonio, 1994; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998).

However, we must make note of one study, of particular interest for this investigation as it addresses Accuplacer results, which came to the opposite conclusion. Sullivan and Nielsen (2009, p. 4) “suggest that we do not need writing samples to place students” as they believe the instruments assess largely the same construct as standardised placement tests. They base their conclusion on results from a large-scale study conducted at a 2-year college in the US, in which they found significant, positive correlations between local placement essays and the two English subtests of Accuplacer placement instrument, Reading Comprehension (r=.62) and Sentence Skills (r=.69).

However, Sullivan and Nielsen's conclusions would not seem to be shared by other researchers or corroborated by the findings of other studies. Similar investigations have typically reported far less correlation between writing sample and standardised test results (Fulcher, 1997; Lee & Greene, 2007; Matzen & Hoyt, 2004). In addition, in the one other case found in the literature that included a similarly strong correlation between result from a standardised instrument and a writing sample, Wall, Clapham and Alderson (1994) concluded the estimate was not strong enough to indicate the same construct was being assessed. Further, comparisons suggest placement recommendations based on standardised tests like Accuplacer and written essays are frequently divergent (Murphy & Yancey, 2008; Mathay, 1992; James & Templeman, 2009) and construct investigations (Carlson, Bridgeman, Camp, & Waanders, 1985; Park, 2004) would also seem to bolster the position that writing samples address a construct largely different from that assessed by standardised instruments.

Test publishers themselves do not suggest that their instruments or any of their multiple-choice components address the same skills as direct writing performance and, quite the opposite, have invested significant resources into incorporating writing samples into their instruments, including the SAT, ACT, and Accuplacer, for this reason. It is worth noting here, also, that while scores from automatically assessed samples of student writing would seem to correlate well with faculty members' results (r=.70), they would still seem to be attending to different aspects of writing than human experts do (James & Templeman, 2009).

Overall, then, the evidence would seem to suggest that, while there is overlap between the skills assessed by locally marked writing samples and standardised instruments often used for placement, they would appear to be assessing different abilities. However, the studies mentioned above do not specifically address Accuplacer, and therefore we cannot state for certain whether their findings would similarly contrast those of Sullivan and Nielsen's if they did. As such, the possibility that writing samples and Accuplacer results address the same construct was included as a potential rebuttal to the sufficiency warrant (part of the decisions claim) of the validity argument.

3.2.5 Consequences

The consequences of placement instruments for individuals, instruction, and education programs can be substantial. The matriculation of students into the courses best suited to their current abilities is obviously of benefit to them in their educational pursuits and success. Being placed in a level that does not adequately challenge students can lead to motivation issues, dissatisfaction, and also waste valuable time and financial resources. Being placed in a level too challenging can also lead to frustration, and put the student in a position in which they are unlikely to succeed, regardless of effort.

The use of a particular placement instrument (or instruments) can also impact instruction in courses, as a well-functioning placement system can help create classrooms with learners of similar abilities and with similar needs, which is of benefit to both instructors and learners. Accurate initial evidence regarding students' competencies can also result in more beneficial decisions regarding assigning students to particular services which may assist them in succeeding in college, such as tutoring, counseling, academic advising, academic skills programs, and so on (Hughes & Scott-Clayton, 2011; Offenstein & Shulock, 2011).

Research predominantly indicates that standardised multiple-choice instruments, including Accuplacer, demonstrate low predictive correlations with students' course results, even if corrected for restricted range and other factors potentially limiting estimates of the strength of correlations (College Board, 2003; Mattern & Packman, 2009). Hughes & Scott-Clayton (2011) suggest the sole reliance on such instruments, and the general neglect of addressing the relevance of the placement test tasks to that of the instructional domain may be a significant contributing cause of the disappointing reports of limited student (and therefore program) success — whether defined as progress within a particular program, program completion or retention rates, or demonstrated skill improvements — in basic skills, remedial/developmental English, and English as a Second Language courses (Bailey, 2009; Bailey, Jeong, & Cho, 2010a, 2010b; California Community College Chancellor's Office, 2009; Martorell & McFarlin, 2010; Offenstein & Shulock, 2011).

The question, then, becomes whether or not the inclusion of a writing sample improves the situation, and leads to more beneficial consequences. As mentioned earlier, when discussing the issue of relevance, a number of investigations have found writing samples result in greater placement accuracy than a variety of standardised tests (Garrow, 1989; Holderer, 1992; Matzen & Hoyt, 2004; Wolcott & Legg, 1998; Zinn, 1998) and the combination of a writing sample and standardised test result in the best predictive model for student course outcomes (Breland et al., 1987; Galbraith, 1986; Garrow; 1989; Isonio, 1994; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998). All of which suggests writing samples may help lead to more beneficial placement decisions, but few, if any, studies would seem to have actually attempted to address the relative consequences of standardised test use and writing sample use for placement decisions.

The use of placement writing samples may also offer non-placement related benefits for stakeholders. For students and instructors, they can be an initial component of writing or learning portfolios that can be referred to throughout a course or program. They would also seem to be a potential source of valuable data for estimates of incoming student ability and for comparisons with continuing and exiting students' abilities for course and program review efforts. While standardised instruments could be argued to offer similar opportunities, the richness of the information, the greater match between the communicative approach to language learning espoused by most language programs, and the greater likelihood of relevancy to course and program outcomes, may make performance-based assessments like writing samples of more use and benefit for such purposes.

Here again, however, research is needed to substantiate or refute such potential benefits and consequences, both for writing samples and placement exams.

3.3 Closing comments

This chapter has provided a detailed review of two important areas of the literature directly related to the current study: i) recent empirical investigations utilizing one of the two validity frameworks informing the current investigation’s validation structure; and ii) the theoretical and empirical arguments relating to the relative strengths and weaknesses of the use of standardised tests (such as Accuplacer) and locally developed and judged writing samples used for the purposes of placement at higher education institutions. Having thus completed the literature review, the subsequent chapter will address the various materials and methods used to collect and analyze data relevant to the validation framework and its warrants and rebuttals as used in the current study.

3.4 Revisiting the research questions in light of the literature review

As stated in Chapter 1, the primary purpose of this study was to construct and investigate an initial validity argument for the current instruments and overall placement assessment system employed by CMI to exempt, exclude or place students within their Developmental English program.

In doing so, it is hoped that this investigation could provide insights into a number of areas, not only important to stakeholders at the College of the Marshall Islands, but which also may prove valuable at other institutions and in the study of validity and practice of test validation.

Throughout the literature review, particularly in Chapter 3, we have discussed the ongoing debates regarding the relative advantages and disadvantages of standardised, objective tests compared with those of more direct, subjective skills assessments like locally-marked writing samples, with regards to what interpretations can be drawn from the results. With an eye to shedding light on this issue, the first research question (RQ1), asked: With regard to test score interpretation, what do the results of the study indicate relating to: a) the relative advantages of AC and the WS in the CMI context; and b) current debates in the literature regarding standardised instruments and WS's for placement purposes?

Similarly, substantial differences of opinion exist regarding the relative impact of using standardised tests for the purpose of placement, as opposed to more direct skills assessments such as locally marked writing samples. Research question 2 (RQ2), therefore, states: With regard to issues of assessment utilization, what do the results of the study indicate relating to: a) the relative advantages of AC and WS in the CMI context; and b) current debates in the literature regarding standardised instruments and WS's for placement purposes?

Finally, the third research question (RQ 3), looked to draw implications from the current study regarding the usefulness of argument-based validation endeavours, both locally – i.e., for the College of the Marshall Islands – and elsewhere – at other institutions and the field of validity and validation practices in general. As such, RQ 3 asks: Did the articulation and investigation of the validity argument produce insights identifying opportunities for improving the placement assessment system for the benefit of stakeholders at CMI?

CHAPTER FOUR

Materials and Methods

4.0 Introduction

This chapter will outline the various materials and methods employed in the current study. As the context of the investigation was already presented in Chapter 1 in some detail, the chapter begins with a brief description of the participants and ethical consent procedures. Next, the logical model adopted and the theoretical/validation framework developed fro the study will be presented and discussed. Following this, the various materials implemented to collect relevant evidence, and the methods used to analyse the data will be reviewed.

4.1 Participants

CMI serves approximately 850 students, the vast majority of whom are, according to CMI's latest self-study (2009), Marshallese (96%), English Language Learners (98%), reliant on financial aid in the form of US Pell Grants (99.5%), and academically underprepared (35% of those who apply to the college do not meet the minimum test scores to place in any of CMI's English courses and 92% of those that do place into developmental English courses). Additionally, approximately 50% of all students are first-generation college students. Female students comprise 52% of the student population and males 48%. In terms of age, 2.8% of students are under 18, 74.7% are 18-24, 9.5% are 25-29, 8.8% are 30-39, 3.7% are 40-49, and 0.1% are 50 or older. Approximately 8.8% of students at CMI have completed the General Education Development program (also frequently referred to as the General Equivalency Diploma, though best known simply as GED), and the remainder come with high school diplomas. Some 36.4% of students come from the 'outer islands' where exposure and access to the English language, English speakers, and primary and secondary teachers who speak the language of instruction at CMI is far less likely than in the urban centres of Majuro and Ebeye, where 59.6% of CMI students come from. The remaining 4% come from outside of the Marshall Islands, mostly from the Federated States of Micronesia.

Participants in the current study included two groups who completed the placement test instruments and one group of CMI English instructors:

i) candidates — high school or GED graduates (or those about to graduate) who had completed both placement instruments, but may or may not have gone on to study at CMI

ii) first-semester students, who had been accepted and placed (based on their placement results) into English courses at CMI, and were either in the first few or last few weeks of their study, depending on the time and type of data collected

iii) English course instructors at the college

Further details regarding the participants for each specific data collection method will be presented later in the chapter, when discussing the collection methods to which they contributed.

4.1.1 Ethical issues

All instructors who participated in the focus group interview, and all first-year student participants who gave permission to use their placement test and course results, signed an information and consent form (attached as Appendices C and D). All participants were informed of the purpose and nature of the study, and of their rights as participants, including the right to not participate or to end their participation at any time. For first-semester students, the nature of the research and the contents of the information and consent form, where applicable, were explained in spoken Marshallese. At least one person fluent in both English and Marshallese was present every time information and consent forms were presented to potential student participants. For participants asked to complete questionnaires, the completion of the forms is considered demonstrated consent to participate. These participants, however, were also carefully informed of their rights as participants in the study, and this was done with a Marshallese speaker present to translate and facilitate questions and answers.

Approval for the study was granted by both the College of the Marshall Islands and the Human Research Ethics Committee of Macquarie University (attached as Appendix E).

4.2 Logical and theoretical frameworks

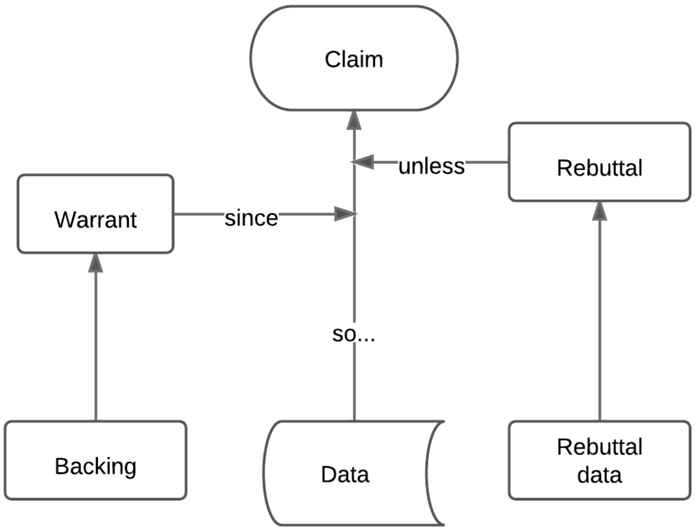
This section of the chapter will present the argument model which would function as the logical mechanism for the investigation, and the theoretical/validation framework used to guide the nature and order of the claims, justifications, and evidence which comprise the argument, and serve to guide the subsequent inquiry.

4.2.1 Toulmin's argument model

As the Toulmin model (Figure 4.1) has already been presented in Chapter One, only a brief summary will be presented here. While the past three decades have seen the widespread acceptance of validation as hypothesis testing (Bachman, 2004, 2005; Chapelle et al., 2004, 2008; Kane, 1992, 2004, 2005; Messick, 1989, 1996; Shepard, 1993), a relatively recent development in this process has been the widespread adoption of a common logical mechanism for validity arguments. First used in educational assessment by Mislevy, Steinberg, and Almond (2002, 2003), Toulmin's (2003) argument model has since been incorporated into the frameworks of many other influential researchers (e.g., Bachman, 2005; Bachman & Palmer, 2010; Chapelle et al., 2004, 2008; Kane, 2004) and would appear to have become the standard in validation studies.

Using one of Bachman's (2005) examples to illustrate the process, we might make the claim that Mark is a US citizen. This claim is based on the data that Mark was born in the USA. A warrant supporting this claim is that all persons born in the US are automatically US citizens. Backing for this claim could come from relevant statements supporting this position from the US constitution, for example. However, we might encounter information that brings our claim into question. A possible rebuttal to our claim could be that Mark has renounced his US citizenship, and evidence supporting this (rebuttal data) could come in the form of an affidavit signed by Mark stating this is the case.

Figure 4.1: Toulmin's Argument Model (adapted from Bachman, 2005, p. 9)



The use of Toulmin's structure was considered essential for the current study, as it assists in articulating an argument consisting of a chain of logical inferences — from performance on a test through to an interpretation of examinee ability, for example, or all the way through to decisions made about examinees and the consequences — explicitly laying out the assumptions upon which these inferences are based, and outlining both supporting and potentially refuting theory and evidence. As such, it is extremely useful as the basic structure for articulating and evaluating test validity arguments.

4.2.2 Conceptual (validation) framework

While researchers have long pointed out the importance of test use and test use consequences in validation (see, e.g., Messick, 1989; Spolsky, 1981), discussions of validity and utilization have typically occurred in isolation to one another (Bachman, 2004, 2005; Chapelle, 1999; Chapelle et al., 2010), and research of the consequences of test use have been comparatively sparse (Bachman, 2005; Kunnan, 2003). Further, these discussions have largely "failed to provide an explicit link between these two essential considerations" (Bachman, 2005, p. 7). Both Kane (2004) and Bachman (2005; Bachman & Palmer, 2010) have attempted to fill this void by addressing both test interpretation and utilization within their validation structures. As detailed accounts of both approaches have been presented earlier, in Chapter Two, only a brief synopsis will occur here with the necessary details to explain the reasons the validation framework implemented for this study — somewhat of a hybrid of the two approaches — was decided upon.

Kane's interpretive model focuses on three inferences made when moving from a candidate's performance on an assessment through to a resulting interpretation of ability. The first inference, evaluation (originally referred to as scoring), refers to the transformation of a performance on an assessment (the observation) into an observed score. The second is the generalization of that observed score to a 'universe score', representative of what the examinee could be expected to obtain on multiple, similar tasks completed in various settings. The third inference is the extrapolation of this indication of examinee ability obtained from the rather narrow realm of the test, to a 'target score', such as what the candidate might be expected to be able to do in another domain, such as the language classroom.

Kane's model was chosen as the basis for the interpretive part of the current study's validation framework because it is likely the most widely known and commonly used framework in educational assessment, serving as the basis for most other influential frameworks, including Chapelle et al's (2004, 2008, 2010) investigations into the validity of the TOEFL, and Bachman's (2005, Bachman & Palmer, 2010) Assessment Use Argument as well. As such, it offers an approach and terminology that may be commonly understood and facilitate the communication and comparability of the results of the investigation.

Kane (2004) later made an important addition to his model, a fourth step which he called 'decisions', in order to incorporate test use within the approach. However, Bachman's (2005; Bachman & Palmer, 2010) Assessment Use Argument (AUA) was felt to offer greater detail and structure in its consideration of utilization, including issues such as sufficiency, equitability, values, and consequences, which were believed to be of particular importance for the investigation of the placement instruments and placement assessment system at CMI. For this reason, a hybridized utilization argument, with a claim for decisions, within which sufficiency, equitability, and other issues were subsumed as warrants, and a claim for consequences. Bachman's publications regarding the AUA (2005, Bachman & Palmer, 2010) were also rich sources of ideas for establishing warrants, rebuttals, and evidence for all segments of the framework developed.

The resulting validation argument, then, consists of both a test use interpretation argument and a complimentary utilization argument. It was hoped that this approach would combine the strengths of both Kane's and Bachman's approaches, and result in a framework that comprehensively attended equally to aspects of both score meaning and instrument use, long argued for by in the literature (e.g., Messick, 1989).

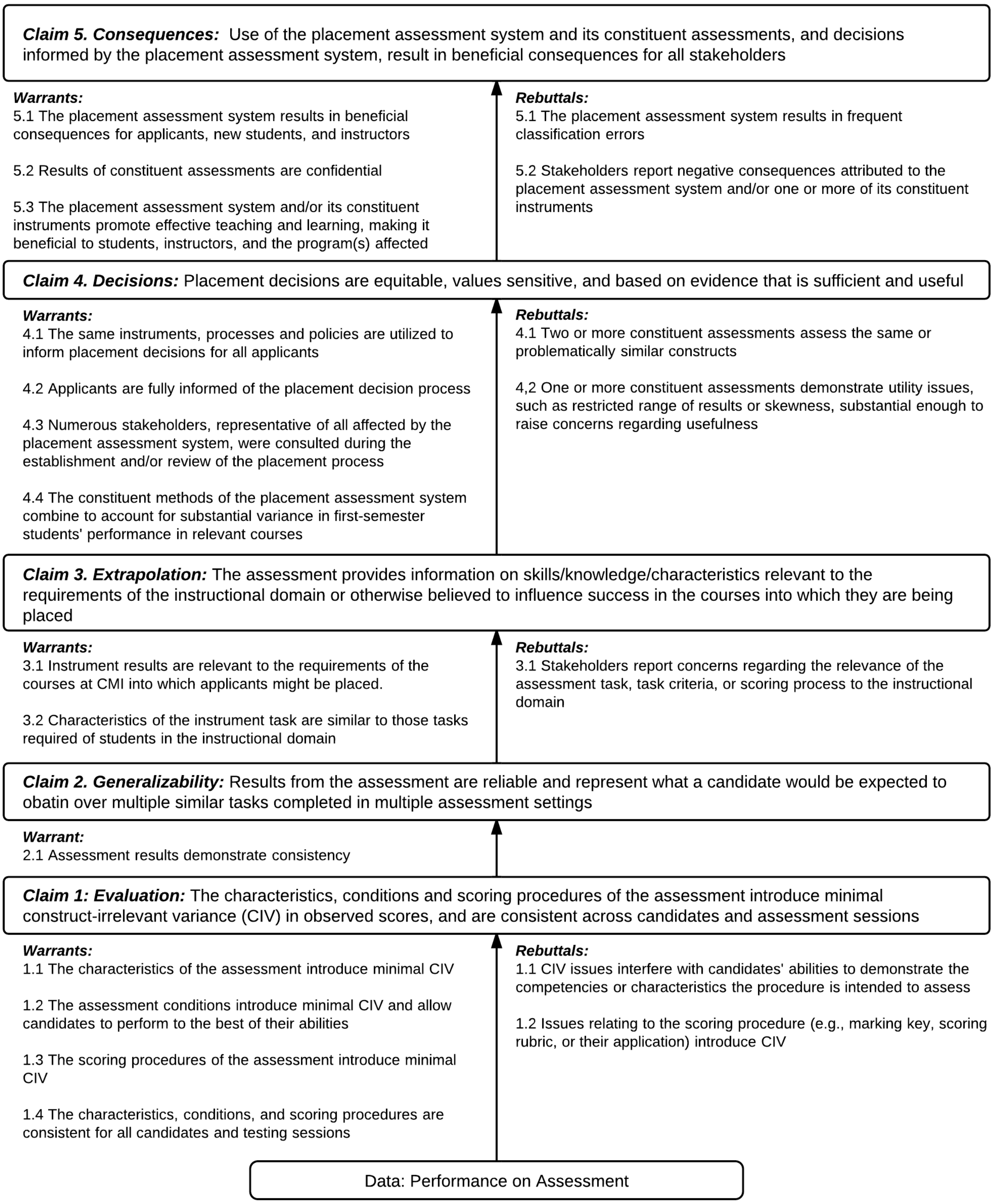
Before presenting the validation framework for this study, however, it should be noted that this model and its constituent claims, warrants, rebuttals, as well as the types and sources of evidence considered and the instruments developed to collect relevant data, are not purely the design of the author, but also the outcome of substantial input and negotiations with several constituents at the college, including instructors, departments chairs, and academic administrators. The design is certainly neither complete nor exhaustive. In future, it is hoped the framework will be expanded, to include further data, warrants, and claims as well to make the investigation more comprehensive and fruitful.

Figure 4.2 presents a summary of the validation framework for the study. The first claim of the argument structure attends to the evaluation inference, involving the transformation of examinees' performances on an assessment into an observed score. What is essentially of concern here is that issues related to the: i) characteristics of the assessment; ii) conditions of the assessment situation; iii) scoring procedures; and iv) the consistency of these factors for all candidates and assessment sessions, introduce minimal construct irrelevant variance (CIV) which might influence results in ways that compromise the scores' representativeness of the candidate's ability (Kane et al., 1999). If we were to find out, for example, that numerous questions on an instrument had no correct or ‘best' answer, or that applicants were not given enough time to possibly complete the assessment, or that the instructions or content were confusing to candidates, it would cause us to question the reliability of the resultant scores as an indication of candidates' performance, or what they might have been capable of performing if these issues had not existed.

Evidence informing the warrants and rebuttals of the evaluation claim were sought through the solicitation (via questionnaires and focus group interview) of the insights of various stakeholders: English instructors (also the placement essay raters), candidates who had completed the placement assessment instruments (and may or may not have subsequently become students at CMI), and first-semester students (who had completed the PAS instruments and been placed accordingly). Additionally, consideration of the characteristics of the instruments, information provided by the publisher (in the case of Accuplacer Companion), and testing conditions and procedures outlined by CMI policies were reviewed for evidence relating to these warrants. Materials, methods and participants related to each source of evidence mentioned in this section will be discussed in more detail in the relevant sections later in this chapter.

Kane (2004) points out that “scoring of essay questions and performance tasks is more judgmental than that of objective tests and therefore requires additional backing for its dependability” (p.156). To this end, MFRM was used to analyse the functioning of a number of facets of the WS scoring procedure, including rater behaviour – both intra- and inter-rater consistency – and the scoring rubric criteria and rating scale.

Figure 4.2 Validation Framework



Claim 5. Consequences: Use of the placement assessment system and its constituent assessments, and decisions informed by the placement assessment system, result in beneficial consequences for all stakeholders

Warrants:

5.1 The placement assessment system results in beneficial consequences for applicants, new students, and instructors

5.2 Results of constituent assessments are confidential

5.3 The placement assessment system and/or its constituent instruments promote effective teaching and learning, making it beneficial to students, instructors, and the program(s) affected

Rebuttals:

5.1 The placement assessment system results in frequent classification errors

5.2 Stakeholders report negative consequences attributed to the placement assessment system and/or one or more of its constituent instruments

Claim 4. Decisions: Placement decisions are equitable, values sensitive, and based on evidence that is sufficient and useful

Warrants:

4.1 The same instruments, processes and policies are utilized to inform placement decisions for all applicants

4.2 Applicants are fully informed of the placement decision process

4.3 Numerous stakeholders, representative of all affected by the placement assessment system, were consulted during the establishment and/or review of the placement process

4.4 The constituent methods of the placement assessment system combine to account for substantial variance in first-semester students' performance in relevant courses

Rebuttals:

4.1 Two or more constituent assessments assess the same or problematically similar constructs

4,2 One or more constituent assessments demonstrate utility issues, such as restricted range of results or skewness, substantial enough to raise concerns regarding usefulness

Claim 3. Extrapolation: The assessment provides information on skills/knowledge/characteristics relevant to the requirements of the instructional domain or otherwise believed to influence success in the courses into which they are being placed

Warrants:

3.1 Instrument results are relevant to the requirements of the courses at CMI into which applicants might be placed.

3.2 Characteristics of the instrument task are similar to those tasks required of students in the instructional domain

Rebuttals:

3.1 Stakeholders report concerns regarding the relevance of the assessment task, task criteria, or scoring process to the instructional domain

Claim 2. Generalizability: Results from the assessment are reliable and represent what a candidate would be expected to obatin over multiple similar tasks completed in multiple assessment settings

Warrant:

2.1 Assessment results demonstrate consistency

Claim 1: Evaluation: The characteristics, conditions and scoring procedures of the assessment introduce minimal construct-irrelevant variance (CIV) in observed scores, and are consistent across candidates and assessment sessions

Warrants:

1.1 The characteristics of the assessment introduce minimal CIV

1.2 The assessment conditions introduce minimal CIV and allow candidates to perform to the best of their abilities

1.3 The scoring procedures of the assessment introduce minimal CIV

1.4 The characteristics, conditions, and scoring procedures are consistent for all candidates and testing sessions

Rebuttals:

1.1 CIV issues interfere with candidates' abilities to demonstrate the competencies or characteristics the procedure is intended to assess

1.2 Issues relating to the scoring procedure (e.g., marking key, scoring rubric, or their application) introduce CIV

Data: Performance on Assessment

As the generalizability claim addresses the inference that the observed score is representative of the ‘universe score' — the result the examinee might be expected to have attained had they completed multiple tasks similar to the assessment, over a variety of test settings (Kane et al., 1999) – indices of reliability can serve to justify or rebut the claim. For

Accuplacer Companion , an estimate of internal consistency of the items was used. For the writing sample, estimates of inter- and intra-rater reliability, and indices of reliability for the facet of candidate ability were consulted.

For the extrapolation claim of a placement instrument, we are particularly interested in what instrument results might be able to tell us about whether the instrument elicits evidence of candidate competencies believed to be relevant to student success in the courses into which students are being placed.

Additionally, the extrapolation claim includes mention of "other attributes or relevant information" in order to be applicable to other possible sources of evidence which may, in future, be considered for the placement assessment system. These could include, for example, candidate self-reports or the collection of background information relating to issues like educational history, motivation for pursuing a tertiary education, time management skills, and other personal and situational factors indicated as significant predictors of student success in college courses (Armstrong, 2000)

Evidence for the extrapolation warrants/rebuttals was sought primarily through expert (i.e., CMI English course instructors') opinion regarding the relevance of the instrument tasks to those of the instructional domain, and estimates of convergence between instrument results and first-semester students' final course outcomes.

The decisions claim asserts the equitability and values sensitivity of the PAS and its constituent instruments, and that the information considered is sufficient and useful for the decisions being made about candidates. The equitability warrant was informed via review of current CMI placement policies and an investigation into how many, if any, candidates have been placed via means other than the PAS. Values sensitivity was assessed through a consideration of the parties and methods used to select and implement the current PAS. With regard to sufficiency, the predictive capacity of the combined outcomes of both PAS instruments for first-semester students' final course outcomes was investigated. With respect to potential utility issues for the constituent placement instruments, this was addressed through an analysis of the descriptive statistics and/or relevant MFRM evidence. Finally, in order to address the rebuttal that writing samples and Accuplacer the same construct (as asserted by Sullivan and Nielsen, 2009) and, as such, utilizing both does not expand the collective competencies or other attributes assessed by the PAS, estimates of correlation between AC and WS results were established.

Perhaps the most important aspect of the framework, the consequences claim asserts that the PAS and its constituent instruments result in beneficial consequences for all stakeholders. Evidence sought to inform the warrants/rebuttals of this claim included stakeholder opinion, solicited via questionnaire items completed by candidates who had just completed AC and the WS, and first-semester students who had been placed via the current PAS. Additionally, insights from English instructors were gathered via questionnaire and focus group interview.

The following sections will provide further details regarding the materials and data collection, participants in the various components of the investigation, and the methods of analysis implemented.

4.3 Materials and data collection

Table 4.1 provides a summary of the various types and sources of data collected to inform the evaluation of the validity argument for Accuplacer Companion and the local writing sample. Aggregate data, such as anonymous placement instrument results, considered for the current study ranged from the Fall 2008 through Fall 2011 semester. Individual participant data, such as placement instrument scores and final course outcomes, were collected from Fall 2009 through Fall 2011 (i.e., 3 full academic semesters).

4.3.1 Placement test results

As both AC and the WS have been described in some detail in Chapter One, discussions here will focus on how the results were used to inform the validity arguments for the two instruments.

Table 4.1: Data sources utilized

Data Source			Description
Placement Test Results	1	Aggregate AC results	Automatically scanned, scores and compiled by computer, for all candidates over the duration of the study
	2	Aggregate WS results	Rated by CMI English language instructors, for all placement test takers over the duration of the study
	3	AC and WS results for all participating first-semester students	Both scored as described above
Course Results	4	Final English course results	For all participating first-semester students, provided by instructors as a score out of 100 (as opposed to a letter grade)
Questionnaires	5	Examinee questionnaire	Conducted post-exam, regarding placement instruments
	6	First-semester students questionnaire	Regarding the appropriacy and impact of their placement
	7	English instructor questionnaire regarding placement instruments	Soliciting opinions as to their functioning and consequences
	8	English instructor questionnaire regarding student placement	Soliciting opinions as to where first-semester students should ideally have been placed
Interview	9	English instructor focus group interview	Discussing various aspects of the performance and impact of the placement instruments and overall PAS

Results from the tests were collected for two groups. First, aggregate results for all candidates from the Fall 2008 through Fall 2011 semester were provided by the Registrar's Office, at the request and approval of the Institutional Research department. Results for 2118 total examinees were provided. For all candidates (whether admitted to CMI or not), total scores for the AC English test and both subtests – Reading Comprehension and Sentence Skills – were provided. With regard to the WS, only the final score was provided. The final score was either the average of two raters' scores if both resulted in the same placement recommendation (see Fig. 1.1), the lower score if raters' opinions resulted in adjacent placement recommendations, or the median score if raters' judgments resulted in three sequential placement recommendations. Oddly, for all 2118 candidates, the recommended placement decision as a result of the WS score was provided, but for 613 examinees the actual final WS score was missing, resulting in 1505 applicants for whom a final WS score was available. These results were used to establish descriptive statistics and inform various analyses with regard to the functioning the PAS tests at CMI.

Second, placement instrument results for first-semester students were used to inform the evaluation claim of both individual instruments and the sufficiency warrant of the decisions claim for the overall PAS. For the individual instruments, estimates of predictive capacity were established through correlating results of AC and the WS, separately, with first-semester students' final course outcomes. For the sufficiency warrant, the combined results of the two instruments were used in the predictive model, in order to establish the predictive capacity of the overall PAS.

Table 4.2 shows the numbers of first-year students participating in this part of the study, broken down by English course and level.

Table 4.2: First-year student participant numbers, by English course, level, and placement assessment

	Course	Accuplacer Companion	Writing Sample
Listening & Speaking	Level 1	93	92
	Level 2	43	43
	Level 3	23	23
	Credit (Speech)	1	1
	Total	160	159
Reading & Writing	Level 1	100	99
	Level 2	20	20
	Level 3	24	24
	Credit (Composition)	0	0
	Total	144	143

Aggregate AC results were used to establish internal consistency estimates for the instrument. As insufficient detail was available to utilize MFRM with the WS results, these outcomes could not be used to inform scoring consistency concerns for the instrument.

4.3.2 Writing sample rater results

In order to inform the appraisal of the generalizability and evaluation claims for the WS, it was important to be able to investigate the functioning of the scoring rubric criteria and scale, and the influence of rater behaviour on instrument results. In order to do this, through MFRM, more detailed information than final rater scores was necessary. To this end, the Chair of the Developmental Education Department and the faculty member in charge of organizing writing sample rating efforts provided anonymous results from 358 writing samples, produced by candidates in the 2009 academic year, each judged by at least two raters, involving contributions from 15 unique raters. These results came from essay ratings used to place applicants. However, unlike the writing sample scores used in other parts of the study (such as the aggregate results used to establish descriptive statistics or the writing sample scores of participating first-semester students used to establish estimates of the predictive validity) these results indicated individual raters (as an anonymous number) and included the scores ascribed for each criteria of the marking rubric, making insightful Rasch analysis of the functioning of the scoring rubric criteria and scale, and insights into rater behaviour, possible. Rasch methodology is explored in further detail in the methods section later in the chapter.

4.3.3 Final course results

Participants' course results was an important source of information used in this study to investigate the predictive capacities of the placement instruments for student performance in the courses into which they were placed. These results were used in the evaluation of the extrapolation claim for both instruments and the sufficiency warrant of the decisions claim for the overall PAS.

However, a number of concerns regarding the use of course results in such estimates, particularly letter grades, have been raised in the literature. First, letter grades provide a restricted range of potential outcomes, which, as a result, can lead to overly conservative correlation estimates and under-represent the predictive capacity of the instrument(s) in question (Armstrong, 2000; College Board, 2003). In an attempt to minimize this problem, all instructors were requested to confer final grades as a percentage rather than a letter grade.

Additionally, grades, whether reported as a letter or a percentage, cannot separate the influence of competencies which may be assessable via placement tests, and various other factors known to influence grades, such as student course load (Graham, 1987), whether or not students are employed or how many hours they work, their time management skills, punctuality, motivation, and so on (Armstrong, 2000). For CMI Developmental English courses, for example, many faculty members reserve as much as 15% of student grades for attendance and participation, neither of which are likely to be predicted by a placement test. In an attempt to address this issue, where sufficient data was provided by instructors, final course results were adjusted to remove the influence of attendance and participation, and make course grades more reflective of student competencies assessable by placement tests. These adjusted performance estimates were then used to create new estimates of predictive validity for the instruments.

Finally, estimates from Bailey, Jeong, and Cho (2010) suggest that only 30-40% of students in remedial courses actually complete the class. While similar estimates for CMI were not available, retention and completion rates have been identified as significant problems in self-study reports to accreditors (CMI, 2008, 2009). As many students who start the course do not finish, we are left with course results from a sample of students who may not be representative of the general population of test-takers. Or, if students who do not complete the course are included in the data analysis as having been unsuccessful, their actual abilities may be misrepresented. For example, if we imagine a course with four equally weighted in-class assessments determining the final grade, a student who completed the first two assessments and achieved 100% on both, but then stopped attending, will have the same impact on the results in our data analysis as a student who attended the entire semester and achieved 50% on all four assignments. Obviously, there is a difference in performance and likely ability between these two students that is lost in the typical means of data analysis using course results. In the current study, where possible, further efforts were made to also produce final course scores determined only by assessments which were completed by the student, in order to establish estimates more reflective of student abilities, and less influenced by issues such as time management, motivation, and other factors for which the placement instruments were not designed to account.

The downside of making such adjustments was the potential loss of participants for whom sufficient data was available to re-calculate final course results. Table 4.3 shows the numbers of participants for each phase of the final course results used to establish predictive validity estimates for the instruments.

Table 4.3: First-year student participant numbers by course, level, assessment, and estimate of course result

Course	Level	Instrument	Final Grades	Final Grades Adjusted 1*	Final Grades Adjusted 2**
Listening & Speaking	1	Accuplacer Companion	93	92	84
		Writing Sample	89	88	84
	2	Accuplacer Companion	43	43	43
		Writing Sample	43	43	43
	3	Accuplacer Companion	23	22	22
		Writing Sample	23	22	22
	4	Credit Speech	0	0	0
Reading & Writing	1	Accuplacer Companion	100	29	0
		Writing Sample	99	29	0
	2	Accuplacer Companion	20	0	0
		Writing Sample	20	0	0
	3	Accuplacer Companion	24	24	24
		Writing Sample	24	24	24
	4	Credit Speech	0	0	0

* influence of attendance and participation mark removed

** influence of attendance, participation, and incomplete assessments removed

4.3.4 Questionnaires

A number of questionnaires were employed during the study in order to gain insights into the opinions and experiences of various stakeholders affected by the placement instruments and the PAS. Stakeholders completing questionnaires included instructors, placement test examinees, and new students placed by the current PAS.

Before presenting the questionnaires, it should be reiterated that no particular theory or model informed the design of the questionnaires in this first iteration of the validation study. Rather, the instruments are more the product of negotiations amongst a variety of stakeholders – the researcher, administrators, department chairs, and faculty members, for example – as to what information they felt should (or should not) be collected and what uses they had for the information. While the final products, presented here, may not be the same as if they had been designed solely by the researcher and solely for the purposes of the study, this approach was hoped to help facilitate stakeholder investment in the project and expand the insights and evidence considered. Instruments used in future iterations of the study may be more focused, in construct and or model, upon areas of particular interest revealed in the current study. The sections that follow describe each of the individual questionnaires developed and used in the study.

4.3.4.1 Instructors' opinions of student placement

At the end of each of the four semesters across which the study was conducted, developmental and credit English course instructors were asked to give their opinions as to where first-semester students in their courses should, ideally, have been placed based solely on the language skills the students demonstrated throughout the duration of the course. It is important to note that instructors were requested to categorize students on the language skills relevant to the particular course they were teaching. That is, Listening and Speaking (or credit Speech) course instructors were requested to place students in the Listening and Speaking (or credit Speech) level they felt best matched the student’s English listening and speaking abilities.

Opinions were solicited via questionnaire (attached as Appendix F), which had a list of first-semester students in each of their courses, and a Likert scale from 0 to 4 representing:

0 — not prepared for any English course currently available at CMI

1 — Level 1 of the Developmental English Program

2 — Level 2 of the Developmental English Program

3 — Level 3 of the Developmental English Program

4 — credit level English

Results were used to inform the consequences claim, as an indication of the level of satisfaction amongst instructors with the perceived performance of the PAS, and the existence or not of mixed ability level classes, which could negatively impact teaching and learning in the courses.

4.3.4.2 Instructors' opinions of the placement instruments

Instructors' opinions regarding various aspects of both AC and the WS, and their use at the college, were solicited via questionnaire (attached as Appendix G). The questionnaire was completed in the Spring 2009 semester by 14 (82%) of the 17 Developmental English course instructors. As faculty turnover at the College of the Marshall Islands is significant, a substantial variance in instructor experience at the institution, in the Marshall Islands, and working with the student population at CMI is typical. As such, it was deemed important to establish the familiarity of participating instructors with the learning objectives of CMI English courses. In response to the statement "I am familiar with the Student Learning Outcomes (SLOs) for the courses which I am teaching or have recently taught at CMI", 13 of the 14 participants agreed (92.9%, 6 strongly agreed, 7 agreed, 1 did not respond). As such, the instructors, as a group, would appear confident in their knowledge of the SLOs of their courses. This is important if we are to ascribe value to the insights of these instructors when it comes to issues such as the relevance of the test tasks to the instructional domain, opinions as to where students should have been placed based on their language skills, and other aspects of the validation framework.

Table 4.4 reports the number of instructors who reported currently teaching, or who having recently (within the past 2 years) taught, each of the English courses first-semester students might be placed into at CMI.

Table 4.4: Number of instructors with recent experience teaching various English courses at CMI

Listening & Speaking Courses	Instructors Reporting Recent Experience	Reading & Writing Courses	Instructors Reporting Recent Experience
Level 1	5	Level 1	6
Level 2	2	Level 2	2
Level 3	2	Level 3	3
Credit Speech	0	Credit Composition	1

Results indicate that, other than credit-level Speech, there was at least one, and typically 2-6, instructors amongst the participants with recent experience and current familiarity with the learning outcomes for each of the English courses first-semester students might be placed into at CMI. This was felt important, in order to establish the range of courses and breadth of the overall, relevant English courses, participating instructors could refer to with recent experience and expertise.

Instructor questionnaire items and the aspects of the validity arguments they were intended to inform are presented in Table 4.5, below (item 1 is not listed as it solicited the information regarding familiarity with course SLO's).

While the inclusion and use of stakeholder opinion was always regarded as an important aspect of the current study, it was not always clear-cut as to what facets of the validation argument were best informed by particular items and respondents' opinions. Items 2 and 3, for example, were intended to address issues of usefulness; specifically, whether the instruments might be too difficult for the large majority of test takers (something many instructors had suggested since the adoption of Accuplacer Companion), resulting in a restricted range of scores problematic for discriminating amongst candidates. It could be argued, however, that as these items address subjective opinions, as opposed to quantitative data, they may be more telling of consequences of the instrument's use, as an indication of stakeholder (dis)satisfaction, for example. It is entirely possible, after all, for an instrument to be perceived as useful for making a particular decision, while empirical evidence suggests the opposite.

Table 4.5: Instructor questionnaire items, related warrants/rebuttals, and rationale

No.	Item	Warrant/ Rebuttal	Rationale
2	The English subtests of Accuplacer Companion would seem to be of an appropriate difficulty level for applicants to the college.	Decisions claim: usefulness of results	While arguably more informative of the consequences claim, better indicating instructor (dis)satisfaction with the use of AC, these items were intended to help inform the usefulness rebuttal of the decisions claim. Should results of the instrument be too skewed, decisions based on a restricted range may increase the risk of misplacement due to measurement error
3	The English subtests of Accuplacer Companion are probably too difficult for most Marshallese students.		
4	Most students applying to the college will be able to understand the questions in the test.	Evaluation claim: Test characteristics	If instructions, test items, and other texts are confusing to examinees, because of how they are written, the language they use, cultural references, etc., this impacts candidates' ability to demonstrate competencies in which we are interested
5	The Accuplacer Companion test asks students to do the same sorts of things they will be expected to do in their classes at CMI.	Extrapolation claim: Test tasks mirror instructional domain	If instructors feel the instrument engages the same or similar competencies required of students in the courses they teach, this supports the claim of relevance to the instructional domain
6	The Accuplacer Companion test is a good test to choose which applicants are admitted to study at CMI.	Decisions claim: usefulness rebuttal	While, like items 2 and 3, arguably more informative of the consequences claim, these items were intended to help inform the usefulness rebuttal of the decisions claim
7	The Accuplacer Companion test is a good test for placing incoming students into Developmental English or Credit level studies.		
8	The Accuplacer Companion test has a positive impact on students' perceptions of their English language skills.	Consequences claim, Consequences for examinees and instructors	If the instrument negatively impacts test-takers perceptions of their own English language skills and/or motivation to pursue a higher education, these are negative outcomes of test use that would rebut the consequences claim. Further, instructor perception of such negative consequences for test-takers would indicate dissatisfaction with the instrument and its inclusion in the PAS.
9	The Accuplacer Companion test likely has a positive impact on students' desire to pursue postsecondary studies at CMI or another institute.		

However, it was decided that the original intentions of the items would be retained in the study, for three reasons. First, these were the intended uses of the items as argued for by stakeholders contributing to the design of the questionnaire. Second, as instructors, familiar with the requirements of the courses into which new students are being placed, and with (in some cases considerable) experience working the students of the institution, it was felt that the perspective of these participants offered valuable, though certainly not comprehensive, insights into the various aspects of the study addressed by the questionnaire. And third, no warrant or rebuttal is informed entirely by the perception of one group of participants alone. For example, items 6 and 7 solicit instructor opinion as to the usefulness of the instrument. These results did not inform the usefulness rebuttal alone, as descriptive statistics and, in the case of the WS, MFRM results were also employed.

The remaining items on the questionnaire are the same as those listed above, but reworded slightly so as to inquire about respondent opinions regarding the WS.

At the end of the items referring to the AC test, and again at the end of the items addressing the WS, open-ended ‘Comments’ sections were provided for respondents to offer any further insights they wished.

4.3.4.3 Candidate questionnaire

At two separate testing sessions during the Fall 2009 semester, upon completion of AC and the WS, a number of examinees were asked to complete a questionnaire (attached as Appendix H) soliciting opinions regarding the placement instruments. Table 4.6 provides a summary of background information on the candidate respondents.

A total of 175 examinees completed the questionnaire. Respondents came from testing sessions held at two of the largest high schools in the country, both of which are in Majuro, an urban centre where access to, and need for use of, English on a daily basis is far greater than in more rural parts of the nation. As such, it is possible that the participant pool reflects a group which has been exposed to English more and uses English more frequently, both in and out of school, than applicants who come to the college from the ‘outer islands’ where there is very little exposure to English in day-to-day life. It is possible, then, that results from this questionnaire may be more reflective of the perceptions of the roughly 60% of CMI students who come to the school from the urban centres, rather than the approximately 36% who come from more remote locations.

Table 4.6: Self-reported background information of examinee questionnaire respondents

1	Gender	Male	Female		NR			
		77 (44%)	97 (55.4%)		1 (0.6%)			
2	Age	18-24	24-29	30-39	40-49	50+	NR*	Mult.**
		136 (77.7%)	22 (12.6%)	10 (5.7%)	2 (1.1%)	1 (0.6%)	3 (1.7%)	1 (0.6%)
3	First language	Marshallese		English		Other	NR	Mult.
		169 (96.6%)		3 (1.7%)		3 (1.7%)		
4	Years studying/using English	1-3	4-6	7-9	9-11	11-13	14+	NR
		31	40	22	25	46	11	
5	Hours using English in school daily	<1	1-2	2-3	3-4	>4	NR	Mult.
		48	34	24	14	46	9	
6	Hours using English outside school daily	<1	1-2	2-3	3-4	>4	NR	Mult.
		59	33	22	16	20	25	
7	I often speak English with family and friends	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	NR	Mult.
		14	63	49	28	18	2	1
8	I am good at listening and speaking in English	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	NR	Mult.
		22	73	43	20	14	3	
9	I am good at reading and writing in English	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	NR	Mult.
		17	65	59	20	11	3	

*NR = no response
**Mult. = multiple responses

As there are no time limits for the placement instruments, examinees simply leave whenever they have completed both tests. As such, it was not possible to address the entire group of examinees at one time before they began completing the questionnaire. However, at least two Marshallese native speakers were on hand at all times to distribute, explain, help with, and collect the questionnaires.

Questionnaire items and the aspects of the validity arguments they were intended to help evaluate are presented in Table 4.7 (items 1-9 informed the background information reported in Table 4.6).

Items 18 through 25 on the questionnaire were the same as items 11 through 17, but reworded slightly to inquire about the writing sample.

At the end of the questionnaire, an open-ended ‘Comments’ section was provided for respondents to offer any further insights they wished to pass along.

Table 4.7: Examinee questionnaire items, related warrants/rebuttals, and rationale

No.	Item	Warrant/ Rebuttal	Rationale
10	I understood the Accuplacer Companion English test questions.	Evaluation claim: Test characteristics	If examinees had trouble understanding the exam instructions, questions, or other components, it raises concerns about their ability to demonstrate the competencies in which we are interested, and introduces CIV into instrument outcomes.
11	The Accuplacer Companion English test was easy for me.	Decisions Claim: Usefulness	Arguably more relevant to the consequences claim, as an indicators of test-taker satisfaction with the use of the instrument, these items were intended to address the utility rebuttal of the decisions claim.
12	The Accuplacer Companion English test was too difficult for me.		
13	The Accuplacer Companion English test is a good test to choose which students can study at CMI.		
14	I had enough time to carefully read and answer all of the questions.	Evaluation claim: Test conditions	If examinees feel they had insufficient time to complete the assessment, this raises concerns as to whether the assessment conditions allowed them to demonstrate the best of their abilities.
15	Taking the Accuplacer Companion English test made me think I can be a successful student at CMI.	Consequences claim: Consequences for test-takers	These items relate to the consequences of test use, specifically the impact of the instrument on students' perceptions of their academic and linguistic competencies, and their motivation to pursue a tertiary education at CMI
16	Taking the Accuplacer Companion English test made me feel good about my English abilities.		
17	Taking the Accuplacer Companion English test made me want to study at CMI.		

4.3.4.4 First-semester student questionnaire

Near the end of each semester, first-semester students participating in the study completed a questionnaire (attached as Appendix I) about their experiences in their English classes and their opinions as to where they should have been placed. Table 4.8 presents their self-reported background information.

Table 4.8: First-semester student questionnaire respondents' background information

Gender	Male	Female		Blank			
	43	41		6			
Age	18-24	24-29	30-39	40-49	50+	NR*	Mult.**
	74	5	5	3	0	3	0
First language	Marshallese		English	Korean	Other	NR	Mult.
	79		3	1	5	1	1
Years studying/using English	1-3	4-6	7-9	10-11	11-13	14+	NR
	9	7	6	17	46	5	0
Hours using English in school daily	<1	1-2	2-3	3-4	>4	NR	Mult.
	16	19	7	19	29	0	0
Hours using English outside school daily	<1	1-2	2-3	3-4	>4	NR	Mult.
	32	17	7	12	18	4	0
I often speak English with family and friends	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	NR	Mult.
	14	63	49	28	18	2	1
I would rate my English listening and speaking as	Beginner	Beg./Int.	Intermediate	Int./Fluent	Fluent	NR	Mult.
	0	4	24	42	20	0	0
I would rate my English reading and writing as	Beginner	Beg./Int.	Intermediate	Int./Fluent	Fluent	NR	Mult.
	1	3	20	44	19	3	0

*NR = no response
**Mult. = multiple responses

Respondents were asked to rate the extent to which they agree with the following statements, based on a 5-point Likert scale with the following options: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree. (Items 1-11 obtained the background information described above.)

12. I am doing well in these classes.
13. I think I am in the right Listening and Speaking class for my ability.
14. I think my Listening and Speaking class is too difficult for me.
15. I think my Listening and Speaking class is too easy for me.
16. If I do my best I can pass my Listening and Speaking class.
17. If I do my best I can get an A or B in my Listening and Speaking class.
18. I think I am in the right Reading and Writing class for my ability.
19. I think my Reading and Writing class is too difficult for me.
20. I think my Reading and Writing class is too easy for me.
21. If I do my best I can pass my Reading and Writing class.
22. If I do my best I can get an A or B in my Reading and Writing class.
23. I have the ability to do well in these classes.
24. I understand what the teacher asks me to do in these classes.
25. I can do what the teacher asks me to do in these classes.
26. I have the ability to pass these classes.

All of the items in the questionnaire were intended to inform the consequences claim, specifically with regard to consequences for first-semester students.

Additionally, the questionnaire recorded which English courses the students were currently in, and which courses they thought they should have been placed in at the beginning of the semester. This was intended to give a students' perspective as to the performance of the current placement assessment system in matching applicants with the courses best matching their current language skills.

At the end of the questionnaire, an open-ended 'Comments' section was provided so that respondents might offer any further insights they wished to, regarding their placement or experiences in their first-semester English courses.

4.3.4.5 Faculty focus group interview

The English instructors who completed the instructor questionnaire also participated in an hour-long focus group interview, led by the researcher, discussing various aspects of the placement instruments and the impact of their use and the decisions made based upon their results. Each participant was provided a copy of the Accuplacer Companion English test, student learning outcomes for all CMI English courses into which applicants can be placed, and samples of recent writing prompts and instruction to help inform their decisions.

The focus group interview was semi-structured, in that a small number of questions were prepared ahead of time, and significant flexibility was given for participants to lead the line of discussion and expand upon their thoughts and opinions.

Participants in the focus group were the same as those who completed the English instructor questionnaires outlined earlier. Fourteen (82%) of the 17 Developmental English course took part. As established earlier, as a group, the instructors reported familiarity with the learning outcomes of the English courses at CMI and there was present at least one, though typically 2-6, instructors with recent – within the past 2 years – experience teaching every English courses into which new students might be placed, with the lone exception of the credit Speech course. As such, it is argued the instructors possess important expertise and experience regarding the instructional domain, and resultantly, their insights provide important evidence regarding various aspects of the validation framework.

The questions prepared for the focus group interview were as follows:

1. Would the instruments (Accuplacer and the writing sample, addressed at separate times in the interview) seem to be of an appropriate difficulty level for applicants to the college?

This item was intended to inform the usefulness rebuttal of the decisions claim. However, it was later decided to be more appropriate to the consequences claim, as an indication of the satisfaction of instructors with the use of AC in the PAS.

2. Would the texts of the instruments (such as the instructions, prompts, questions, etc.) be readily understood by most applicants?

This question relates to the evaluation claim. If instructions, test items, and other texts are confusing to examinees, either in the way they are written, their content, or the nature of the language they use, this impacts the examinee's ability to demonstrate the skills the instrument is intended to assess and introduces unwanted CIV into observed scores.

3. Do the instruments ask students to do the same sorts of things they will be expected to do in their English classes at CMI?

This question relates to the extrapolation claim. As instructors of the courses that new students are placed into, insights of these participants regarding the relevance of the instrument tasks to the requirements of the instructional domain is important evidence informing the extrapolation claim..

4. What impact, if any, will the instruments have on test takers, such as their perceptions of their abilities, their desire to pursue a higher education, for example?

This was intended to inform the claim of beneficial consequences of test use, particularly regarding impact on test-takers.

Participants were asked their thoughts regarding each question in relation to both instruments — AC and the WS.

4.4 Data analysis procedures

A number of data analysis procedures were utilised throughout the current study. This section of the chapter will outline each of the procedures, the data it was applied to, and the aspect of the study the results were intended to inform.

4.4.1 Descriptive statistics

Descriptive statistics and raw score distributions for AC and WS results were established as a means of informing the utility rebuttal of the decisions claim for each instrument. Specifically, this evidence was intended to address the possibility of a restricted range of results produced by the assessment.

Kuder-Richardson 21 formula was used as a means of estimating the internal consistency of AC. As only final total scores, for each candidate, were available for the two subtests – Reading Comprehension and Sentence Skills – and their combined score (i.e., the total AC score), score 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. While KR-21 is known to produce an overly conservative estimate of reliability (Brown, 2005), as it does not require knowledge of specific item variance to determine, it was used to inform the generalizability claim for AC.

4.4.2 Coefficients of correlation and determination

Throughout the study, Pearson product-moment correlation coefficients (r) are used to establish estimates of convergence between placement test results and a selection of other variables, such as students' final course grades. Based on Pearson correlation efficient (r), it was possible to calculate coefficient of determination (r^2), which provides an estimate of the amount of variance in one factor apparently accounted for by variance in another. If the correlation between the writing sample results and students' course grades is found to be

r=0.5, for example, then r^2=.25, meaning approximately 25% of the variance in final grades seems to be accounted for by the construct assessed by the writing sample.

4.4.3 Multi-faceted Rasch measurement

Writing samples and other instances of judged performances involve a number of factors which may influence observed scores. These variables may be related to the writing task, such as the prompt (topic), expected mode of response, or number of writing samples provided, or may be related to the scoring process, including the training of raters, rater biases, and the criteria and scales of the scoring rubric (Hamp-Lyons, 1990; Park, 2004).

Many-facet Rasch measurement (MFRM) (Linacre, 1989), an extension of the one-parameter Rasch model, "provides a framework for obtaining fair measurements of examinee ability that are statistically invariant over raters, tasks, and other aspects of performance assessment procedures" (Park, 2004, p. 2). As such, MFRM can be a valuable tool in both the quality control of performance rating, but also in validity investigations regarding performance assessment procedures (see, for example: Kondo-Brown, 2002; Milanovic, Saville, Pollitt, and Cook, 1996; Park, 2004; Weigle, 1998)

For the purposes of this study, the software program FACETS (version 3.68.1, Linacre, 2011) was employed to investigate a variety of aspects potentially influencing variance in the writing sample results, including: i) performance at differentiating amongst candidates based on ability; ii) rater behaviour (i.e., rater severity/leniency and intra- and inter-rater consistency); and iii) scoring rubric criteria and rating scale functioning. Results of the MFRM analysis informed a variety of warrants/rebuttals, including WS scoring procedures, generalizability, and utility.

While the topic for the writing sample can also influence results in writing samples, this facet was not included in the MFRM analysis, as no record of which task individual candidates responded to would seem to have been kept by the institution.

4.4.4 Linear regression

Linear regression is a useful tool for predictive modeling — the estimation of one variable, such as performance in a language class, based upon other, known variables, such as estimates of related competencies (e.g., placement test scores), or other factors known or believed to influence performance in the future variable.

In this study, linear regression was used to estimate the apparent predictive capacities of both placement tests, individually and combined, for student performance in CMI English classes. The greater the amount of variance an instrument seems to account for in student course outcomes, the stronger the case that the sufficiency warrant for that procedure is justified.

4.5 Closing comments

Having presented a detailed account of the materials, logical and theoretical frameworks, and methods of analysis employed by the validation study, the following chapter will present the results of these various lines of enquiry.

CHAPTER FIVE

Results

5.0 Introduction

This chapter will present the results of the various lines of investigation detailed in Chapter 4. The structure of the chapter will follow that of the validity arguments for the placement tests, focusing first on the evaluation claim and its relevant warrants and/or rebuttals, followed by the generalizability, extrapolation, decisions, and consequences claims.

5.1 Evaluation claim

The evaluation claim asserts that the characteristics, conditions and scoring procedures introduce minimal construct-irrelevant variance (CIV), and are consistent for all individuals and assessment sessions. This section will review results of the lines of evidence informing the warrants articulated for this claim. As different sources of evidence and methods of analysis were employed regarding the two placement procedures, results for each instrument will be presented separately, starting first with the writing sample.

5.1.1 Evaluation claim warrants for the writing sample

Evidence gathered to inform the evaluation warrants for the placement essay came from three sources: MFRM, current institute policies and procedures regarding the development of instrument texts, and stakeholder insight.

As much of the evidence for the evaluation warrants of the writing sample are informed by the MFRM, and as this is the first time these findings are presented, a brief overview of the results (from a sample of placement essays produced by 358 candidates, rated by 15 different CMI English instructors, in the 2009 academic year) will be depicted here in (Figure 5.1). Further, as MFRM can offer useful estimates of score variance due to facets incorporated into the model (including estimates of candidate ability) and those not (including error, construct-irrelevant factors, etc.), these will also be presented before moving on to discuss specific potential sources of CIV relevant to each warrant.

Figure 5.1: Vertical Wright map of overview of MFRM results for the writing sample

```
+-----------------------------------------------------------------------------+
|Measr|+Raters |+Examinees  |+Items                                     |Scale|
|-----+--------+------------+-------------------------------------------+-----|
(high) (lenient) (more able) (easy)                                     (high)
|   6 +        +            +                                           + (6) |
|     |        |            |                                           |     |
|     |        |            |                                           |     |
|     |        | .          |                                           |     |
|     |        |            |                                           |     |
|   5 +        +            +                                           +     |
|     |        |            |                                           |     |
|     |        |            |                                           |     |
|     |        |            |                                           |     |
|     |        |            |                                           |     |
|   4 +        +            +                                           +     |
|     |        |            |                                           |     |
|     |        |            |                                           |     |
|     |        | .          |                                           |     |
|     |        |            |                                           | --- |
|   3 +        + .          +                                           +     |
|     |        | .          |                                           |     |
|     |        | *.         |                                           |     |
|     |        | **         |                                           |     |
|     |        | *          |                                           |  5  |
|   2 +        + *          +                                           +     |
|     |        | *.         |                                           |     |
|     |        | ****       |                                           |     |
|     |        | ****       |                                           |     |
|     |        | *****.     |                                           | --- |
|   1 +        + *****.     +                                           +     |
|     |        | ******.    | Grammar                                   |     |
|     | 11  45 | ********** |                                           |  4  |
|     |        | ********.  |                                           |     |
|     |        | ********.  | SentenceVariability                       |     |
*   0 *        * *****.     * Diction                                   * --- *
|     | 23  44 | *****.     |                                           |     |
|     | 22  27 | *********  |                                           |     |
|     | 55     | ****.      | Organization         Support              |     |
|     | 17  42 | *****.     |                                           |  3  |
|  -1 + 10  33 + *****.     +                                           +     |
|     | 30  64 | ****       |                                           |     |
|     | 7      | ***.       |                                           | --- |
|     | 24     | ***.       |                                           |     |
|     |        | **.        |                                           |     |
|  -2 +        + **         +                                           +     |
|     |        | **.        |                                           |  2  |
|     |        | .          |                                           |     |
|     |        | *.         |                                           |     |
|     |        | .          |                                           |     |
|  -3 +        + *.         +                                           + --- |
|     |        |            |                                           |     |
|     |        | *          |                                           |     |
|     |        |            |                                           |     |
|     |        | .          |                                           |     |
|  -4 +        + .          +                                           + (1) |
 (low) (severe)  (less able) (difficult)                                (high)
|-----+--------+------------+-------------------------------------------+-----|
|Measr|+Raters | * = 3      |+Items                                     |Scale|
+-----------------------------------------------------------------------------+
Measr = measure
```

The first column on the left of the figure shows the interval scale used in Rasch analyses, the logit. The second column provides an estimate, in logits, of the leniency/severity of the individual raters, reported according to their assigned numbers, with the more lenient judges towards the top of the scale and the more conservative towards the bottom. The third column, reports the estimated ability of each candidate. Those towards the top of the scale would be expected to do better, across criteria and raters, than those towards the bottom of the scale. The fourth column reports the relative difficulty of the scoring rubric criteria. Grammar would appear to be the easiest (and/or marked the most easily) for candidates, while organization and support were the most difficult (and/or marked most conservatively).

Rasch analysis can offer useful estimates of the ratio of raw score variance due to facets incorporated into the model, including candidate performance, and residual variance not explainable by the dimensions in the model (Baghaei, 2008; Khairani & Nordin, 2011). The variance attributable to the three facets included in the model was 58%. After accounting for apparent interactions between facets, 22% of raw score variance was left unexplained. Part of this 22% is due to a fourth facet, the different tasks (i.e., the writing prompts), which is typically included in such models, but was not possible for the current study as the institution would seem to have no record of the prompt to which individual candidates responded.

Further evidence relating to construct-irrelevant variance comes from the reliability index of the candidate performance facet. Results of the MFRM reliability index (.91) for the facet of candidate ability suggests the extent to which assessments of examinee ability reliably discriminated amongst candidates based on ability was contaminated with only 9% error (Aryadoust, in press).

Estimates of only 9% of variance in the facet of candidate ability, and 22% of the observed variance in the total model, unaccounted for by the dimensions included in the Rasch model could be argued to be reassuringly low, particularly for a judged performance assessment. However, we must bear in mind that the institution does not employ Rasch analysis when considering placement assessment data. As Rasch modeling conveys information about examinee performance relative to the functioning of the other defined traits, in this case rater severity and item difficulty, it allows the determination of test-taker performance, and estimates of error variance in the assessment of test-taker performance, independent of the influence of variance in other facets included in the model (Jackson, Draugalis, Slack, Zachry, & D'Agostino, 2002).

5.1.1.1 Warrant 1.1: Test characteristics

Warrant 1.1 states that the characteristics of the instrument should not introduce substantial construct-irrelevant variance. Of particular concern to a number of stakeholders at the college was whether or not candidates, the vast majority of whom are Marshallese English Language Learners, could comprehend the texts of the instruments. Should the instructions, reading, writing prompts, or other texts, be confusing or incomprehensible to examinees, this could introduce construct-irrelevant variance in the observed scores. While stakeholders expressed this concern primarily for AC, it was also investigated for the WS. Evidence gathered to inform this warrant included a review of institutional policy and procedure for the development of the writing prompts, including instructions and topic selection, and stakeholder opinion as to the comprehensibility of the texts for test-takers.

Current CMI policy and procedure is that all writing prompts (including instructions and topic selection) are to be developed locally by a small group of English instructors, selected by the Chair of Developmental Education. According to the Chair at the time of the study, all members of the English faculty are welcome to participate in the creation of writing sample prompts, but efforts are made to ensure individuals with experience teaching and assessing writing, particularly with Marshallese English Language Learners, are well represented. When new prompts are created, they are then reviewed both internally, by the department chair and other Developmental English faculty members, and externally, by a consultant with experience and expertise in second language writing assessment. While it would not appear that any written guidelines for the development or review (internal or external) of the writing prompts have been established, during a focus group interview (detailed in section 4.3.4.6), faculty members experienced with the process stated that clarity and comprehensibility for Marshallese English Language Learners is considered an important focal point. While the process stated would seem to place due value on the importance of developing unambiguous instructions and tasks, policy and intent alone do not guarantee outcome. As such, stakeholders directly involved in the process were consulted for their insights.

During the course of the focus group interview, fourteen of the seventeen (82%) Developmental English instructors (who are also the pool of writing sample raters and prompt creators) were asked for their opinions regarding the comprehensibility of the writing sample texts. Samples of recently used writing prompts (which included all instructions) were made available to all participants for review prior to and during the discussion. The only perception reported by any faculty members during the interview was that the instrument instructions and prompts were likely not a problem for the majority of test-takers. An item on the anonymous questionnaire that followed the focus group interview, reported in Table 5.1, along with results for chi-square tests for statistical significance, also addressed this topic. While the majority of instructors reported agreeing with the sentiment that the instrument texts are comprehensible to most applicants, the majority was not statistically significant (64% agreed, p = .366).

Table 5.1: Stakeholder opinions regarding comprehensibility of Writing Sample texts

Stakeholders	Questionnaire statement	Group	N	Observed Proportion	Expected Proportion	Sig. (2-tailed)
Instructors	Most applicants to the college will be able to understand the texts (such as the instructions, prompts, questions, etc.) of the Writing Sample.	Agree*	7	.64	.50	.366❖
		Disagree+	4	.36		
		Total	11	1.00		
Test-takers	I understood the Writing Sample instructions.	Agree*	91	.75	.50	.004
		Disagree+	30	.25		
		Total	120	1.00		

* Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses
❖Assumption of minimum 5 participants in each cell not met

A questionnaire completed by test-takers who had just completed the writing assessment (detailed in section 4.2.5.3) included a similar item. The significant majority of examinees (75% agree, p = .00) reported no problems being able to understand the assessment instructions, suggesting the instrument texts were not likely an impediment to the procedure engaging and assessing the intended construct.

Results of the different data sources are somewhat difficult to reconcile. The finding that the significant majority only examinees report no confusion regarding the writing sample texts supports the warrant. On the other hand, only seven of eighteen instructors express the opinion writing sample text comprehensibility is not likely an issue (four disagreed, six responded “neither agree nor disagree”, and one provided no response), and 25% of examinees reported being confused by the instrument texts. The warrant states that characteristics of the test introduce minimal construct-irrelevant variance, however, and 25% of examinees being confused by the instructions and/or tasks of the instrument cannot be said to be minimal. As such, the evidence would seem to support a rebuttal against Warrant 1.1 for the writing sample.

5.1.1.2 Warrant 1.2: Test conditions

Evaluation Warrant 1.2, that the assessment conditions do not introduce construct-irrelevant variance, was evaluated using two sources of information: CMI policy regarding placement testing, and test-taker opinion regarding the sufficiency of time they had to complete the instrument. As mentioned in Chapter Four (section 4.3.4.3), only the single test condition of time sufficiency was addressed in this study for two reasons. First, as test-taker opinion was solicited via a questionnaire to be completed after finishing both the Accuplacer Companion (mathematics and English sections) and the writing sample, it was felt the instrument needed to be as brief as possible. Second, sufficiency of time allotted for the writing sample was identified as a potential issue by faculty and other stakeholders during the negotiation of what questionnaires involved in the study were to address.

Current CMI policy is that all placement assessments are to be paper-based or, put another way, non-technology based. One of the reasons for this (along with lack of reliable power or computers at many testing sites) is to avoid problems with inexperience with technology hindering the performance of many examinees, and introducing construct-irrelevant variance in the results. As such, this practice, and that the only instruments used by the college for placement purposes are, indeed, paper-based, supports Warrant 1.2.

Another institutional policy that would certainly seem to support the warrant is that there are to be no time limits placed upon students for the completion of either Accuplacer Companion or the writing sample. Curiously, however, faculty and other stakeholders identified time sufficiency as a potentially problematic aspect of the testing conditions and suggested its inclusion in the examinee questionnaire. Table 5.2 reports examinee opinions regarding the sufficiency of time allotted for the writing sample.

Table 5.2: Examinee opinions regarding sufficiency of time allotted for the writing sample

Questionnaire statement	Group	N	Observed Proportion	Expected Proportion	Sig. (2-tailed)
I had enough time to carefully read the instructions and finish the Writing Sample.	Agree*	100	.78	.50	.000
	Disagree+	29	.22		
	Total	129	1.00		

* Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses

The significant majority of test-takers do not report concerns regarding time available to complete the writing sample (78% agree, p = .00), which could be argued to support the warrant. However, 22% of examinees expressing the opinion they did not have enough time is at least somewhat perplexing, given there are to be no time constraints on either placement test. Further, while not directed specifically regarding the writing sample, a number of test-takers offered additional comments offered on the questionnaire, such as “I would like more minutes and I would also like the instructor to announced every fives until the test is finished”, “Give more time for students to take the test”, “timing is short”, and “add much more time” give the impression of some form of time limit being imposed, at least to the perception of some of the examinees. These outcomes, in addition to the perceptions reported by faculty members that applicants are not allotted sufficient time for the instrument leave the conclusion here unclear. Unclear, however, is not evidence in support of the warrant.

5.1.1.3 Warrant 1.3: Scoring procedures

Warrant 1.3 assures against construct-irrelevant variance introduced by the scoring procedures of the assessment. As Kane (2004, p. 156) points out, “scoring of essay questions and performance tasks is more judgmental than that of objective tests and therefore requires additional backing for its dependability”. A variety of evidence produced by MFRM, relating to rater behaviour (i.e., inter- and intra-rater consistency), the performance of the scoring rubric items, and the functioning of the scoring rating scale employed, were used to inform this warrant.

5.1.1.3.1 Rater consistency

The first aspect of rater consistency reviewed to inform Warrant 1.3 was intra-rater consistency. Should raters be found to be inconsistent from one essay to the next, this is particularly problematic. While inter-rater discrepancies also introduce CIV, if raters themselves are consistent, these discrepancies can at least be estimated and attenuated via methods such as the polytomous Rasch analysis employed in the current study. Results of the analysis for rater severity/leniency are presented in Figure 5.2.

Evidence regarding the consistency of individual raters can be found in the infit and outfit mean square results in Figure 5.2 (Infit MnSq in column 7, and Outfit MnSq in column 9). Using the traditional .5 to 1.5 acceptability range (Weigle, 1998), raters with mean square values above this array are likely to be problematically unpredictable, while those below are overly predictable. None of the raters included in the sample were found to be outside this range, suggesting no evidence of intra-rater reliability concerns for the writing sample.

Looking to inter-rater reliability, results in column 5, 'measure', in Figure 5.2, report the estimate of rater leniency/severity. From the most lenient rater (45, at +.57 logits, to most severe (24), at −1.70, there is a 2.27 logit discrepancy in leniency amongst the 15 judges. Other studies reviewed, which involved from 3 to 34 raters, report leniency variance from .54 to 5.24 logits (Haiyang, 2010; Kassim, 2011; Park, 2004; Schumaker & Smith Jr., 2007; Aryadoust, in press). While relatively moderate, the discrepancy in rater severity was found to be statistically significant (fixed, all-same, X^2=596.3, df = 14, p=.00), suggesting it is introducing CIV in candidate performance estimates. The relatively low model standard error of measurement (Model S.E.) for each rater (from .05 to .15) and mean for all raters (.09) indicate the data available was sufficient for the estimates generated.

Figure 5.2 Writing Sample MFRM Rater Measurement Report

```
+-----------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit     |Estim.| Correlation |                      |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Raters            |
|-------------------------------+--------------+---------------------+------+-------------+----------------------|
|   155      65       2.4   4.01|    .57   .15 |  .86  -.8   .83 -1.0| 1.22 |   .76   .64 | 45 45                |
|   670     200       3.4   3.98|    .53   .09 |  .72 -3.1   .70 -3.4| 1.32 |   .83   .79 | 11 11                |
|   642     183       3.5   3.34|   -.28   .09 |  .82 -1.8   .82 -1.8| 1.23 |   .83   .76 | 23 23                |
|   535     180       3.0   3.34|   -.29   .09 | 1.31  2.7  1.27  2.4|  .67 |   .65   .70 | 44 44                |
|  2122     640       3.3   3.28|   -.37   .05 | 1.23  3.9  1.24  4.1|  .74 |   .63   .70 | 27 27                |
|   261      80       3.3   3.18|   -.49   .13 | 1.02   .1  1.00   .0|  .97 |   .71   .70 | 22 22                |
|  1184     379       3.2   3.15|   -.52   .06 | 1.16  2.2  1.14  1.9|  .83 |   .73   .70 | 55 55                |
|  1548     460       3.4   2.91|   -.84   .05 |  .83 -2.8   .82 -2.9| 1.15 |   .70   .70 | 17 17                |
|   336     160       2.1   2.87|   -.89   .11 |  .90  -.8  1.02   .2|  .96 |   .69   .73 | 42 42                |
|   297     110       2.7   2.81|   -.98   .12 | 1.16  1.1  1.15  1.0|  .84 |   .72   .77 | 33 33                |
|  1393     430       3.2   2.79|  -1.00   .06 |  .75 -4.1   .77 -3.7| 1.27 |   .70   .70 | 10 10                |
|   768     265       2.9   2.68|  -1.15   .07 | 1.04   .5  1.02   .3|  .99 |   .75   .69 | 30 30                |
|   338     120       2.8   2.62|  -1.23   .12 |  .67 -2.8   .76 -1.8| 1.33 |   .86   .81 | 64 64                |
|   403     160       2.5   2.48|  -1.42   .10 |  .93  -.5  1.06   .4|  .95 |   .76   .76 |  7 7                 |
|   342     160       2.2   2.28|  -1.70   .11 | 1.05   .5  1.02   .2|  .97 |   .64   .65 | 24 24                |
|-------------------------------+--------------+---------------------+------+-------------+----------------------|
|   732.9   239.5     2.9   3.05|   -.67   .09 |  .96  -.4   .98  -.3|      |   .73       | Mean (Count: 15)     |
|   554.2   159.1      .4    .48|    .63   .03 |  .19  2.3   .18  2.2|      |   .07       | S.D. (Population)    |
|   573.7   164.7      .5    .50|    .65   .03 |  .19  2.3   .18  2.2|      |   .07       | S.D. (Sample)        |
+-----------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .10  Adj (True) S.D. .62  Separation 6.34  Strata 8.79  Reliability .98
Model, Sample: RMSE .10  Adj (True) S.D. .64  Separation 6.57  Strata 9.10  Reliability .98
Model, Fixed (all same) chi-square: 615.0  d.f.: 14  significance (probability): .00
Model,  Random (normal) chi-square: 13.6  d.f.: 13  significance (probability): .40
```

Contrasting the raw scores of the two (or, in some instances three) raters who judged the same essay, the average difference was 1.93, on an assessment out of 12 possible points (i.e., 16%), and the range of differences was 0 to 7.6 (a 63% difference). Table 5.3 reports the differences found when contrasting the placement recommendations (using current CMI cut scores) resulting from individual rater judgments. On average, the placement recommendations resulting from the two raters were .933 levels apart. While most placement recommendations resulting from the two different scores ascribed by raters were into the same or adjacent levels (76%), a substantial number (24%) recommended placement into courses that were two or even more levels apart.

As CMI does not currently utilize Rasch modeling or another means of attenuating inter-rater variance, instead basing placement recommendations on raw scores alone, the variance in rater behaviour is not only significant, but demonstrated to have a substantial impact on observed scores and resulting placement recommendations, rebutting Warrant 1.2.

Table 5.3: Comparison of placement recommendations resulting from different raters' scores

Scores from two raters result in placement recommendations that are:	Instances	% Instances
0 levels apart (the same)	129	36%
1 level apart (adjacent)	145	40%
2 levels apart	71	20%
3 levels apart	14	4%
4 levels apart	2	0.6%
Total	361	

5.1.1.3.2. Scoring rubric criteria

Three sources of evidence relating to the functioning of the scoring rubric were used to inform this part of the study: i) instructors/raters' opinions solicited during the focus group interview and expressed via questionnaire comments; ii) analyses produced by the Rasch model relating to the unidimensionality of the rubric criteria, and iii) the instances of unexpected responses produced for a particular criteria.

During the focus group interview with English course instructors, some expressed concerns regarding the scoring rubric, describing it as "complicated" and "a bit confusing". Comments offered by instructors on follow-up questionnaires also included concerns about the writing sample rubric, such as "I believe the rubric itself needs to change" and "Currently, the rubric contains overlapping and nebulous categories and descriptions". No individual expressed an opinion in support of the functioning or clarity of the rubric or its criteria. Rater dissatisfaction, however, does not provide direct evidence as to whether the scoring procedure is functionally problematic. For that, evidence from the Rasch model was sought.

Analysis of the item facet, specifically item fit statistics (Figure 5.3), provides valuable evidence regarding the likely extent of the construct-irrelevant variance contaminating an instruments' assessment of the target skill(s) (Baghaei, 2008; Goh & Aryadoust, 2010; Khairani & Nordin, 2011).

Figure 5.3: Writing Sample Items Measurement Report

```
+------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|         Model | Infit      Outfit    |Estim.| Correlation |                       |
|  Score   Count  Average Avrage|Measure   S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Items               |
|-------------------------------+---------------+---------------------+------+-------------+-----------------------|
|  1947    719        2.7   2.66|   -.51    .05 | 1.26  4.6  1.23  4.0|  .74 |   .70   .71 | 1 Support             |
|  1939    718        2.7   2.65|   -.52    .05 |  .99  -.2  1.01   .1|  .96 |   .70   .71 | 2 Organization        |
|  2309    719        3.2   3.21|    .22    .04 | 1.01   .3  1.01   .1| 1.00 |   .74   .73 | 3 SentenceVariability |
|  2191    718        3.1   3.04|   -.01    .04 |  .86 -2.8   .90 -1.8| 1.12 |   .71   .72 | 4 Diction             |
|  2608    718        3.6   3.68|    .82    .04 |  .84 -3.3   .83 -3.5| 1.18 |   .77   .74 | 5 Grammar             |
|-------------------------------+---------------+---------------------+------+-------------+-----------------------|
|  2198.8  718.4      3.1   3.05|    .00    .05 |  .99  -.3  1.00  -.2|      |   .72       | Mean (Count: 5)       |
|   249.2     .5       .3    .38|    .50    .00 |  .15  2.8   .14  2.5|      |   .03       | S.D. (Population)     |
|   278.6     .5       .4    .43|    .56    .00 |  .17  3.2   .15  2.8|      |   .03       | S.D. (Sample)         |
+------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .05  Adj (True) S.D. .50  Separation 11.07  Strata 15.09  Reliability .99
Model, Sample: RMSE .05  Adj (True) S.D. .56  Separation 12.38  Strata 16.85  Reliability .99
Model, Fixed (all same) chi-square: 616.2  d.f.: 4  significance (probability): .00
Model, Random (normal) chi-square: 4.0  d.f.: 3  significance (probability): .26
```

Baghaei (2008) asserts that items that fit the Rasch model – i.e., those within the .5 to 1.5 infit and/or outfit mean square range – are likely contributing to the assessment of the single dimension intended by the instrument; in this case, writing ability. Misfitting items, meanwhile, are possible threats to unidimensionality, as they are potentially assessing something other than the construct intended and, thusly, contributing to construct-irrelevant variance (Baghaei, 2008; Kassim, 2011; Park, 2004). From figure 5.3 we see that none of the scoring criteria items for the writing sample fall outside the acceptable range, thus suggesting none of the criteria utilized are irrelevant to the construct assessed.

Additionally, results of the candidate ability facet analysis (Appendix J) indicate that the procedure is reliably (reliability = .91) separating examinees into three distinct ability groups (separation = 3.18), or perhaps four groups (strata = 4.58) if we trust that construct-irrelevant factors (such as differences in background knowledge, for example) not included in the Rasch model are unlikely to be significantly influencing results (Linacre, 1999).

Additional insight regarding the functioning of the individual criteria included in the scoring rubric was sought from MFRM reports of instances in which raters provided a score considered inconsistent with their own and others' systematic rating behaviour. Should a particular criteria be involved with a substantial number of unexpected responses from judges, this could suggest disagreement between raters in what the criteria is or how it is to be evaluated in the performance. Results, reported in Table 5.4, indicate only a miniscule percentage of the responses provided by raters for each criteria did not fit the expectations of the Rasch model.

Table 5.4: Total instances of unexpected responses by criteria

Criteria	Instances	% total instances
Support	3	0.001%
Organization	8	0.002%
Diction	5	0.001%
Sentence Variability	1	0.000%
Grammar	0	0.000%
Total unexpected responses	17	0.005%
Total responses provided by judges	3572	

Overall, then, the evidence suggests the scoring rubric criteria function well as a unidimensional assessment of a single construct, and that their application by raters would not appear to be introducing substantial construct-irrelevant variance in the writing sample scores.

5.1.1.3.3. Scoring rubric rating scale

With regard to the functioning of the rating scale, the average candidate ability measure and outfit mean square results produced by the Rasch analysis were considered as evidence. From Table 5.5, we see that the average examinee ability increases with each step up the scoring rubric scale. This is evidence that the examinees with higher ratings on the assessed skills are demonstrating more of the construct being assessed than those with lower ratings (Linacre, 1999; Park, 2004). Further, outfit mean square estimates for each rating category are found to be within the acceptable range, suggesting the rating scales are functioning as intended.

Table 5.5: Average candidate ability measures and outfit mean square results by category

Data				Quality Control		
Category Score	Times Category Used	%	Cumulative %	Average Measure	Change	Outfit Mean Square
1	433	12%	12%	-2.46	--	1.1
2	763	21%	33%	-1.61	.85	1.0
3	1094	31%	64%	-.68	.93	.9
4	759	21%	85%	.07	.75	1.0
5	441	12%	98%	.96	.89	.9
6	82	2%	100%	1.42	.46	1.5

Another source of evidence regarding the functioning of the rating scale comes from the category score probability curve, reported in Figure 5.4.

Figure 5.4: Category Score Probability Curve

```
   -4.0         -2.0          0.0          2.0          4.0          6.0
     ++------------+------------+------------+------------+------------++
   1 |                                                                  |
     |                                                              6666|
     |                                                           666    |
     |                                                        666       |
     |1                                                     66          |
   P | 11                                                  6            |
   r |   1                                               66             |
   o |    1                                             6               |
   b |     11                                          6                |
   a |       1                             5555555   66                 |
   b |        1                          55       556                   |
   i |         1222222  3333333         5          65                   |
   l |        221     2*       3344444**          6  55                 |
   i |      22   1   33 2      443   5  4        6     55               |
   t |    22      1 3    22  44   3 5    44    66        5              |
   y |  22         *       24      *3      44 6           55            |
     |22         33 11    442     5  3      6*              55          |
     |         33     1  4   22 55    33   6  44              555       |
     |       33       4**     5*2       **6     444              555    |
     |  33333      444   1***5   222*666  3333     44444            5555|
   0 |********************666*******1***********************************|
     ++------------+------------+------------+------------+------------++
    -4.0         -2.0          0.0          2.0          4.0          6.0
```

The horizontal axis of the figure reports examinee proficiency while the vertical represents probability. Each curve represents the probability of a candidate of a specific ability being assigned a particular score. Of primary interest is whether there is a distinct peak for each scale category curve, and whether the curves are evenly spaced, as “a series of hills” (Park, 2004, p. 15). Outside of categories 1 and 6, the probability curves for the rating scale are problematic. For levels 2, 3, 4, and to a slightly lesser extent, 5, there is more overlap, and less of a distinct peak, than we would hope. This indicates a lack of certainty in the segregation of candidates of different abilities into the most appropriate category for their demonstrated ability. At no ability range, for example, is a candidate much more than

50% likely to be judged as possessing writing ability at the level 2, 3 or 4 level. In other words, for categories 2, 3, 4, or 5, there would appear to be a lack of a clear portion of the ability range for which the category is the most probable given.

These results corroborate the separation (3.18) and strata (4.58) indices reported by the MFRM for the examinee ability facet, which indicate raters could reliably distinguish between only 3, or possibly 4, ability levels amongst candidates (Linacre, 1999; Park, 2004; Aryadoust, in press).

Together, rating scale results suggest that, while average examinee ability does increase with each step up the rating scale, there is reason for concern that one or both of the following may be occurring: i) the scoring rubric scale interval does not match the numbers of distinct ability levels observable in the examinee population (i.e., there are more scale levels than there are reliable strata of abilities addressed by the marking criteria in the target population); and, ii) the rating scales may not be uniformly understood or consistently applied by the raters (McNamara, 1996).

5.1.1.4 Warrant 1.4: Consistency across candidates and sessions

The final evaluation warrant addresses whether or not the characteristics, conditions, and scoring procedures are consistent for all candidates and testing 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 Student Services 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.

No further evidence regarding potential differences in testing conditions across testing sites (such as differences in furnishings, lighting, temperature regulation, etc., which might be at issue in some remote locations, for example) was gathered for this iteration of the study.

Finally, all examinees' results are processed via the same scoring methods. All writing samples are collected, copied with all identifying information redacted, and then distributed to raters for anonymous marking. Results from both raters are entered into computer and automatically compared with cut-scores and resulting placement recommendations are generated. (Further details of the process for scoring both assessments are provided in section 1.3.4.4.) As such, current policy and procedure at CMI would seem to promote consistency in the scoring process across individual test-takers.

In sum, then, while conditions across testing sites is an issue that needs to be investigated in future, the evidence considered for this validation supports Warrant 1.4, as testing characteristics and scoring procedures would appear to be largely consistent for all candidates and testing sessions.

5.1.2 Evaluation warrants for Accuplacer Companion

Turning to AC, three sources of evidence were considered in order to inform the evaluation warrants for the instrument: relevant research published by the test developers, current institutional policies and procedures, and stakeholder insights regarding the instrument and its administration. Ideally, Rasch modeling would also have been conducted for AC results, much like it was for the writing sample. Unfortunately, however, no record of which responses (or correct/incorrect results) for each test item by individual test-takers was available for the study and, as a result, Rasch analysis was not possible.

5.1.2.1 Warrant 1.1: Test characteristics

Whether or not the characteristics of Accuplacer Companion introduce construct-irrelevant variance was investigated using two sources of evidence: research provided by the test publisher, and local stakeholder opinion.

The publishers of Accuplacer assure users of both the adaptive, computer-based OnLine version and its paper-based derivative, Companion, that: i) 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 ii) 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 only with candidates for whom English was their first language and, therefore, likely grew up immersed in an English-speaking context. As this is not the case for the vast majority of CMI students, it is still entirely possible for there to be questions, answer options, or other texts in the instrument which contain language, cultural references, or other presumed background knowledge which may compromise Marshallese English Language Learners' 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.

One of the intentions of the focus group interview and items on follow-up questionnaire conducted with CMI English instructors was to gather opinions regarding the comprehensibility of the texts of both placement instruments for examinees. During the focus group, instructors presented what appeared to be a uniform position that "Accuplacer [Companion] is too difficult for CMI students" and that the "level of language and vocabulary… are far too advanced to be accessible to the vast majority of CMI applicants."

Furthermore, there was widespread agreement that “students accurately placed in Level 1 [Developmental English courses] would not understand very many of the questions of the English subtests”. Most instructors seemed to feel the majority of applicants were likely “guessing for most of the questions”.

One item on the follow-up questionnaire was intended to gather anonymous instructor opinions of the comprehensibility of Accuplacer Companion texts for test-takers. While the results, presented in Table 5.5, were not unanimous, the significant majority (86%, p=.00) felt comprehension of the texts would be problematic for most candidates.

Table 5.5: Stakeholder opinions regarding comprehensibility of Accuplacer Companion texts

Stakeholders	Questionnaire statement	Group	N	Observed Proportion	Expected Proportion	Sig. (2-tailed)
Instructors	Most applicants to the college will be able to understand the texts (such as the instructions, prompts, questions, etc.) of the instrument.	Agree*	2	.14	.50	.008❖
		Disagree+	12	.86		
		Total	14	1.00		
Test-takers	I understood the Accuplacer Companion English test instructions and questions	Agree*	75	.63	.50	.005
		Disagree+	46	.37		
		Total	123	1.00		

* Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses
❖Assumption of minimum 5 participants in each cell not met

In additional comments offered by instructors on the questionnaire, many expressed concerns the language of the test would be confusing for most examinees, as it was “obviously [intended] for native speakers”, while the vast majority of examinees are “far from native-like proficiency”. One instructor wrote, “I don't think students who will place into first level English… would be able to understand very many of the questions of the English sub-tests.” A number of respondents also identified content presenting potential cultural bias, including references to “King Kong”, “the American dream”, “Sesame Street” and “a sports complex” they felt would be “alien to our students”.

Instructor opinion, then, would seem to rebut Warrant 1.1, suggesting comprehension of the instrument texts could be introducing construct-irrelevant variance in the results of Accuplacer. Test-takers themselves, however, in responses to a post-exam questionnaire, did not express the same collective opinion as instructors. As seen in Table 5.5, of the 123 participants expressing a non-neutral opinion (i.e., did not indicate they "neither agree nor disagree"), the significant majority (63%, p=.00) report not having difficulties understanding the instructions and questions on the Accuplacer Companion English subtests.

Collectively, the evidence is problematic in that the various sources do not converge towards the same conclusion. While the significant majority of examinees report the texts were not confusing, nearly 40% of examinees providing non-neutral responses (25% of all respondents, total) stated they did find instructions and questions confusing represents a sizable portion of candidates whose scores may have been influenced by a factor not related to the intended construct. As such, the warrant, which states the instrument characteristics introduce only a minimal amount of construct-irrelevant variance, cannot be said to be supported by the evidence considered.

5.1.2.2 Warrant 1.2: Test conditions

CMI policy for the administration of placement testing sessions is that there is to be no time limit for the completion of either instrument. This policy would seem to support the evaluation warrant that testing conditions are not to introduce construct-irrelevant variance, as time constraints are likely to influence test-taker abilities to demonstrate the relevant skill(s).

However, as mentioned earlier in the chapter, many stakeholders felt that sufficiency of time to complete the instrument was an area of concern to be addressed during the study. To that end, an item included in the questionnaire for test-takers solicited their opinion as to whether or not they had time to "carefully read and answer all of the questions". Of the 133 respondents who provided a non-neutral response, the significant majority (63%, p = .00) agreed with the statement. As reported earlier for the same warrant relating to the writing sample, however, a number of examinees offered comments raising concerns about perceived time limits for the placement tests.

Overall, while CMI policy mandates no time limits, and the majority of examinees do not express concerns about time, the warrant states that test conditions introduce minimal construct-irrelevant variance. The finding that nearly 40% of examinees did report time constraint as a problem, leave open the possibility that time constraints, or at least the perception of them amongst examinees, might be introducing construct-irrelevant variance in a substantial proportion of examinees' results. The warrant, therefore, is not supported by the evidence considered.

5.1.2.3 Accuplacer Companion Warrant 1.3: Scoring Procedures

One of the primary advantages of standardized, objectively scored instruments like Accuplacer is the uniform, often automated, scoring procedures. As CMI employs computer scanning, marking, data entry, and data processing (including computing placement recommendations), there would seem little opportunity for variance in scoring procedures to introduce construct-irrelevant fluctuation, barring perhaps, errors in the marking keys or some other aspect of the process.

Publications from the test developers (College Board, 2003) assure Accuplacer (OnLine and Companion) 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 (e.g., CCCAA, 2007a, 2007b), 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, there would appear to be no evidence suggesting the scoring procedures for Accuplacer Companion introduce construct-irrelevant variance.

5.1.2.4 Warrant 1.4: Consistency across test-takers and sessions

As discussed when reviewing Warrant 1.4 for the writing sample, the existence of an established institutional protocol for all proctors to follow for all testing sessions, that the same two proctors take primary responsibility for all testing sessions held throughout the country, and the opinion of faculty members who have participated in multiple testing sessions that policy is followed and consistency maintained, all point towards the support of this warrant for both instruments.

Additionally, Accuplacer Companion answer sheets are automatically scanned, scored, and processed via computer. No variance across individuals would seem likely.

In sum, then, while conditions across testing sites is an issue that needs to be investigated in future, the evidence considered for this validation supports Warrant 1.4.

5.2 Generalizability claim

A single warrant for the generalizability claim – that instrument results demonstrate consistency – was articulated and evaluated for both placement assessments. As indices of reliability offer insight into the apparent consistency of scores across samples of observations, they provide evidence relevant to the generalizability claim (Kane et al., 1999). Evidence relating to the consistency of the writing sample will be presented first, followed by internal consistency estimates for Accuplacer Companion.

5.2.1 Writing sample reliability

Two sources of evidence from Rasch analysis results were used to inform the generalization warrant for the writing sample: the reliability index of the candidate facet analysis, and intra- and inter-rater consistency. As these findings were already presented and discussed relating to evaluation warrants in section 5.1, they will only be summarized briefly here.

As Rasch reliability estimates for the examinee facet are indications of the reproducibility of results in another, similar, sample (Aryadoust, in press), the high reliability index (.91) for the estimates of candidate ability supports the generalizability warrant for the writing sample. Further, fit statistics (infit/outfit mean square results) for the rater leniency/severity facet suggest all raters involved demonstrated sufficient internal consistency. Inter-rater consistency, however, was found to be a substantial issue, influencing variance in observed scores (fixed, all same, X^2 = 3262.9, df =357, p=.00), and resulting in considerable discrepancies between the placement recommendations derived from scores from different raters of the same essays. As CMI does not currently use any means of accounting for inter-rater variance or attenuating the impact on placement recommendations/decisions, this is a threat to the reliability of the instrument and rebuts its generalizability warrant.

5.2.2 Accuplacer Companion reliability

As results for specific test items, and therefore item score variance, were not available for the current study (see Section 4.4.1), typically preferred methods of establishing reliability estimates for objective instruments, such as Cronbach's alpha, were not possible for the current study. Instead, internal consistency estimates were derived using the Kuder-Richardson 21 formula, which does not require knowledge of variance for each test item. KR-21 results for Accuplacer Companion English section subtests Reading Comprehension, Sentence Skills, and the combined total score were .68, .57, and .76, respectively. While .68 and .57 estimates are well below the traditional .80 acceptability cut-off for reliability, the combined test scores are the basis for placement recommendations. As such, and given the known overly-conservative nature of the KR-21 formula (Brown, 2005), the .76 estimate of reliability was considered sufficient to consider the instrument reliable, supporting the generalizability claim.

5.3 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 target language use domain (i.e., in this case, the instructional domain), or otherwise believed to influence student success in the courses into which they are being placed. Two warrants were articulated for this claim, each of which will be dealt with in turn, with relevant evidence presented. As the same evidence and methodologies were used for both instruments, in the interests of brevity, results for both tests will be reported in each section.

5.3.1 Warrant 3.1: Relevance to the instructional domain

Warrant 3.1 states that instrument results are relevant to the requirements of the courses at CMI into which applicants might be placed. The ability of a placement instrument to substantially predict the final results of students in courses into which the instrument is used to place them would be powerful evidence that the competencies assessed by the test are relevant to those required for student success. For both instruments, it is important to note that final course grades used to establish correlational estimates were percentages (scores out of 100), thus avoiding the problem of restricted range. Tables 5.7 and 5.8 report correlational evidence for the WS and AC, respectively.

Given the writing sample is a direct measure of student writing ability, we might expect it to more strongly predict Reading and Writing (RW) course scores than Listening and Speaking (LS). This pattern does seem to hold for Level 1 students, where the instrument predicted approximately 16% to 17% of final course result variance of RW course results and 6% to 13% of LS course outcomes, depending on whether unadjusted or adjusted (i.e., the removal of attendance, participation, and/or missed assignments were removed) were considered. However, the writing sample results showed no predictive capacity for Level 2 or Level 3 course results, not only for LS, but also RW courses. There would seem to be no reason to suggest the writing sample task and scoring criteria (diction, organization, support, sentence variability, and grammar) are more suited to assessing Level 1 course requirements than Levels 2 and 3.

Table 5.7: Correlations between writing sample scores and final course results

Course	Final Course Result		Level 1	Level 2	Level 3	Credit
Listening & Speaking	Final Result	r^2	.098**	0.023	0.000	.a
		n	92	41	23	0
	Final Result Adjusted 1	r^2	.062*	0.029	0.006	.a
		n	88	41	22	0
	Final Result Adjusted 2	r^2	.129**	0.014	0.000	.a
		n	84	41	22	0
Reading & Writing	Final Result	r^2	.155**	0.020◆	0.094	.a
		n	99	20	24	0
	Final Result Adjusted 1	r^2	.172*	.a	0.096	.a
		n	29	0	24	0
	Final Result Adjusted 2	r^2	.a	.a	0.067	.a
		n	0	0	24	0

a insufficient participants
* significant at .05 level; **significant at .01 level
◆ original correlation (r) negative
Final Results Adjusted 1 - final course results with any influence of attendance and participation removed
Final Results Adjusted 2 - same as Final Results Adjusted 1, with any influence of missed/late assessments removed

Perhaps, then, the results reflect the issues found with the scoring rubric rating scales. Probability curves for the scales indicated indistinct peaks for all but the top (6) and bottom (1) categories used. Additionally, the separation index for the candidate facet suggested the instrument/raters could reliably distinguish between only three (or potentially four, if the strata index were used instead) categories of candidate ability. It is possible the use of 6 scales when only three to four categories of skills can be distinguished between, as well as the inter-rater consistency issues also reported earlier, are resulting in sufficient construct-irrelevant variance in writing sample scores that leave the predictive validity of the instrument, overall, insubstantial and insignificant, particularly in the ‘middle’ categories of test-taker ability.

Given the instrument is designed to assess the reading and sentence-related skills of native speakers, we might not be surprised to find Accuplacer Companion scores best predicted final course results in the most advanced (Level 3) RW course of the Developmental English program. Results indicate test scores accounted for 33% to 39% of variance in final course results. These findings are higher than those typically reported in other predictive validity studies at colleges in the US, usually ranging between 5-22% estimates of determination (Mattern & Packman, 2009). Perhaps unsurprisingly, given the instrument does not measure listening or speaking skills, Level 3 Listening and Speaking course results were not predicted to any significant extent by Accuplacer Companion scores.

Table 5.8: Correlations between Accuplacer Companion scores and final course results

Course	Final Course Result		Level 1	Level 2	Level 3	Credit
Listening & Speaking	Final Result unadjusted	r^2	.194**	0.035	0.088	.a
		n	93	41	23	0
	Final Result Adjusted 1	r^2	.158**	0.037	0.07	.a
		n	89	41	22	0
	Final Result Adjusted 2	r^2	0.246**	0.044	0.018	.a
		n	84	41	22	0
Reading & Writing	Final Result unadjusted	r^2	0.040*	0.181	.386**	.a
		n	100	20	24	0
	Final Result Adjusted 1	r^2	0.159*	.a	.329**	.a
		n	29	0	24	0
	Final Result Adjusted 2	r^2	.a	.a	.353**	.a
		n	0	0	24	0

.a insufficient participants
*significant at .05 level; **significant at .01 level
Final Results Adjusted 1 - final course results with any influence of attendance and participation removed
Final Results Adjusted 2 - same as Final Results Adjusted 1, but with any influence of missed assessments also removed

Results from the instrument did not significantly predict any Level 2 or LS or RW course results. Again, because the instrument is designed for use with native speakers, we might not expect it to predict final results for courses addressing 'intermediate' or 'pre-intermediate' English language learner reading and writing skills, and less so listening and speaking results. Perhaps oddly, however, results did show significant, and somewhat substantial, predictive capacity for not only Level 1 RW course outcomes (4-16%), but also LS course results (16-24%). It is unclear why the instrument would seem to demonstrate substantial predictive capacity for both Level 1 courses, none for Level 2 courses, and then for the RW course of Level 3.

Overall, however, the findings would seem to rebut the extrapolation claim for

Accuplacer Companion. The instrument demonstrated a significant predictive capacity for the outcomes of only three of the six Developmental English courses into which new students are placed. While two of these courses were the RW courses we might expect a reading and writing oriented instrument to be able to predict, the lack of prognostic capacity for the middle level of the program is problematic. Further, English course instructors are of the opinion the instrument addresses tasks and skills not widely relevant to the requirements of the instructional domain. With regard to the writing sample, while instructors argue for its relevance, it demonstrates insignificant predictive validity in all but the two Level 1 English courses. As such, it too, at least as it is currently administered and scored, cannot be said to demonstrate relevance to the instructional domain.

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

During the focus group interview, instructors (all of whom had reviewed copies of both placement instruments) appeared unanimous in the opinion that the writing sample required students to perform tasks similar to what would be required of them in their CMI English courses. One faculty member suggested “this is exactly what they do in their writing classes” and others pointed out it is also very similar to what will be expected of students in non-English courses at the college as well, as the vast majority will require written assignments and open-ended exam questions.

Responses to the item on the anonymous follow-up questionnaire addressing the relevance of the writing sample task to the instructional domain, reported in Table 5.9, corroborate this perception. The significant majority (86%, p=.01) of instructors reported agreeing that the writing sample asks students to do the same sorts of things they are required to do in their English courses at CMI.

Regarding AC, the clear majority opinion expressed during the instructor focus group seemed to be that the instrument tasks are generally dissimilar to the objectives and requirements of Developmental English classes. 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 Developmental English classes, and which CMI “credit level students would struggle with”. As we see from Table 5.9, while follow-up questionnaire item results confirm this was, indeed, the majority perception, the results were not statistically significant. 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 Developmental or even entry level credit level English courses at CMI. None of the comments offered addressed positive aspects of the instrument in relation to the extrapolation claim.

Table 5.9: Stakeholder opinions regarding relevance of Accuplacer tasks to instructional domain

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
The Writing Sample asks students to do the same sorts of things they will be expected to do in their classes at CMI.	Agree*	12	.86	.50	.008❖
	Disagree+	2	.14		
	Total	14	1.00		
The Accuplacer Companion English subtests ask students to do the same sorts of things they will be expected to do in their classes at CMI.	Agree*	3	.25	.50	.083❖
	Disagree+	9	.75		
	Total	12	1.00		

* Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses
❖Assumption of minimum 5 participants in each cell not met

Instructor opinion, then, would seem to largely support the first extrapolation warrant for the writing sample – that it is relevant to the requirements of the instructional domain – but refute it for Accuplacer Companion.

5.4 Summary of findings regarding the test score interpretation claims and warrants

As the extrapolation claim is the last of the three score interpretation inferences addressed in the validity framework for this study, Table 5.10 presents a brief summary of the findings to this point, before subsequently moving on to the placement assessment utilization claims.

Table 5.10: Summary of findings for the score interpretation claims

#	Claim/Warrant	Writing Sample	Accuplacer
1	**Evaluation Claim**: The characteristics, conditions, and scoring procedures of the assessment do not introduce construct-irrelevant variance in observed scores	✗	✗
1.1	The characteristics of the assessment do not introduce construct-irrelevant variance in observed scores	✗	✗
1.2	The conditions of the assessment do not introduce construct-irrelevant variance in observed scores	✗	✗
1.3	The scoring procedure for the assessment does not introduce construct-irrelevant variance in observed scores	✗	✓
1.4	The characteristics, conditions, and scoring procedures are consistent for all candidates and assessment sessions	✓	✓
2	**Generalizability Claim**: Results from the assessment are reliable and represent what a candidate would be expected to obtain over multiple similar tasks completed in multiple assessment settings.	✗	✓
2.1	Assessment results demonstrate consistency	✗	✓
3	**Extrapolation Claim**: The assessment provides information on skill, knowledge, or other characteristics relevant to the requirements of the instructional domain or otherwise believed to influence success in the courses into which they are being placed.	✗	✗
3.1	Instrument results are relevant to the requirements of the courses at CMI into which applicants might be placed	✗	✗
3.2	Characteristics of the instrument tasks are similar to those tasks required of students in the instructional domain	✓	✗

✓= supported
✗ = refuted

As Table 5.10 shows, only the single claim of generalizability for Accuplacer was supported by the evidence considered in this study. Issues of text comprehensibility and time allotment were reported by substantial enough numbers of examinees to raise concerns the amount of construct-irrelevant variance introduced into observed scores of both instruments is more than the ‘minimal’ allotted by Warrants 1.1 and 1.2. While Accuplacer scoring procedures are entirely automated, and therefore unlikely to chance across candidates, inter-rater discrepancies and rating scale issues rebutted Warrant 1.3 for the writing sample. Warrant 1.4, consistency across test-takers and testing sessions was the only warrant to be supported for both instruments. While internal consistency estimates supported the generalizability claim for Accuplacer, inter-rater reliability issues rebuffed this claim for the writing sample. While instructors argue for the relevance of the writing sample task to what students are required to do in their English courses, backing Warrant

3.2, results of both instruments demonstrated insufficient predictive capacities to support Warrant 3.1.

Having completed the review of test score interpretation claims, we shall turn our attention to the evaluation of the two placement assessment utilization claims: decisions and consequences.

5.5 Decisions Claim

The decisions claim asserts that placement decisions are equitable, values sensitive, and based on evidence that is sufficient and useful. Four warrants were articulated for this claim, each of which, along with relevant evidence, and are presented in turn in the following sections. As the same or similar evidence and methodologies were used for both instruments, results for both will be reported in each section.

5.5.1 Warrant 4.1: Equitability

The first decisions warrant asserts that the same instruments, processes and policies are utilized to inform placement decisions for all applicants. Current CMI policy is that all applicants must complete the PAS, and therefore must complete the same placement instruments, in order to be eligible to register at the institution. Upon reviewing the placement results and decisions of 2120 candidates from the Fall 2008 through Fall 2011 semester, all but three were found to have completed the same instruments – Accuplacer Companion and the writing sample – and been placed via the same placement rules and cut-scores. According to student services staff members involved in the placement testing process, until recently there was an exception at the college which allowed graduates from secondary schools in English-speaking nations to bypass the placement tests. Since the Fall 2008 semester, when the three students in question were admitted, it would appear no candidates have been placed by any other means than the current PAS.

5.5.2 Warrant 4.2: Full disclosure

Warrant 4.2 states that applicants are fully informed of the placement decision process. As CMI is required, by decree of the Ministry of Education, to test all secondary students soon to graduate around the country, it is unclear whether the high schools or the MOE might inform examinees of their own purposes for having students complete the PAS. For its own purposes – i.e., to exclude, place, or exempt students from enrolment in the Developmental English program at CMI – there would appear to be no standing CMI policy with regards to informing examinees of the purposes and procedures of the PAS, and potential outcomes for candidates.

According to faculty who participated in the focus group interview and follow-up questionnaire, 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 the constituent placement instruments, use of cut scores, or other aspects of the placement decision process. To the understanding of both faculty and Student Services staff, the only information most test-takers receive is a final placement decision and a date to come to the school to register.

During the focus group interview with English instructors, some expressed concerns that examinees are not aware of the importance of the writing sample to the placement decisions; that many examinees felt Accuplacer results were far more important and, resultantly, did not spend much time or effort on the writing sample. As one instructor wrote on the comments section of the faculty questionnaire, “The one issue I keep hearing from students is that if they had known how important the writing sample was to their placement, they would have done a better job!” Should this concern be well-placed, it not only is problematic for the current warrant of full disclosure, but is potentially another source of construct-irrelevant variance threatening the evaluation claim for the writing sample as well.

5.5.3 Warrant 4.3: Stakeholder input

Warrant 4.3 avows that numerous stakeholders, representative of all affected by the PAS, were consulted during the establishment and/or review of the placement process. According to faculty members participating in the focus group interview, and other stakeholders present at the time the current PAS was adopted (such as Student Services staff, the former head of Institutional Research, academic administrators, and the former Chair of the Developmental English department, all consulted informally), the selection of the constituent elements was decided upon largely by executive administrators, and the establishment of cut scores and other implementational procedures were carried out primarily by the Institutional Research department. As described in Chapter One, 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 the academic administrators, faculty members, or students, would seem to have occurred.

5.5.4 Warrant 4.4: Constituent instruments' combined sufficiency

The final decisions warrant establishes that the constituent methods of the PAS combine to account for substantial variance in first-semester students' performance in relevant courses. To investigate this warrant, results from both the writing sample and Accuplacer were included in predictive models for final course outcomes (as a score out of 100, rather than a letter grade, in order to avoid problems with restricted range). Table 5.11 presents the results of the regression models for the RW and LS courses. Predictive capacity estimates (r^2) for each individual instrument are provided for comparative purposes.

Results indicate the combined predictive capacity for each course, at each level, improved with the combination of both instrument results rather than either instrument alone. However, predictive validity estimates for the instruments combined were not always significant, and rarely substantial. Looking first to the RW courses, the two instruments combined accounted for a significant amount of variance in final course outcomes only for Level 1 and 3. While the approximately 43% of variance in Level 3 results is certainly substantial, it is doubtful one could argue the 17% of Level 1 variance accounted for is sufficient to warrant placement decisions be based on these results alone. With regard to LS courses, we might not expect assessments of writing ability and/or reading comprehension and sentence skills to offer much prognostic capacity for final course outcomes. While both instruments, alone and combined, significantly predict results in Level 1 LS courses, and combined account for an arguably substantial 25% of variance, they would seem to offer little insight into student success in Levels 2 and 3.

Table 5.11: Predictive Capacity of Placement Instruments for Course Results

		Both Instruments Combined								Writing Sample	Accuplacer
Course	Level	R	R^2	Adj. R^2	SE of Estimate	F	df1	df2	Sig.	r^2	r^2
Reading & Writing	1	.410a	.168	.151	13.451	9.697	2	96	.000	.155**	.040*
	2	.492a	.242	.153	17.798	2.715	2	17	.095	.020	.181
	3	.651a	.424	.369	19.535	7.731	2	21	.003	.094	.386**
Listening & Speaking	1	.500a	.250	.233	11.032	14.852	2	89	.000	.098**	.194**
	2	.260a	.068	.019	15.456	1.378	2	38	.264	.023	.035
	3	.306a	.093	.003	15.567	1.030	2	20	375	.000	.088

a. Predictors: (Constant), writing sample, Accuplacer
Adj. = adjusted
SE = standard error
*=sig. at p=.05 level; **=sig. at p=.01 level

As the instruments combined only account for a significant proportion of variance in the final outcomes of three of the six courses for which student results were available, and only 2 of these instances could be argued to represent substantial variance (25% of Level 1 LS and 43% of Level 3 RW), it would seem the sufficiency warrant for the decisions claim is not supported by the evidence.

5.5.5 Rebuttal 4.1: Overlapping assessments

Rebuttal 4.1 addressed the concern that two or more constituent assessments address the same or problematically similar constructs. This rebuttal was included in large part to attend to the assertion of Sullivan and Nielsen (2009) that writing samples are unnecessary for accurate placement as they address the same construct as standardised placement instruments: critical thinking.

To address this issue, correlational estimates were established (Table 5.12) between writing sample results and scores on the two Accuplacer Companion English subtests – Reading Comprehension and Sentence Skills – and the combined Accuplacer Companion English test results. Scores utilised for the analysis were from 1504 candidates who completed the PAS from Fall 2008 to Fall 2011, and for whom results from all instruments and subtests were available.

Table 5.12 Correlations between Accuplacer Companion and Writing Sample Results

		Reading Comprehension	Sentence Skills	Accuplacer Total
Writing Sample	Pearson Correlation	.455**	.486**	.530**
	Sig. (2-tailed)	.000	.000	.000
	N	1504	1504	1504

** Correlation is significant at the 0.01 level (2-tailed)

From the table, we see writing sample results show significant correlation with both subtests and total scores for the English section of Accuplacer. Correlations of .455 to .530, however, representing approximately 20% to 28% variance likely due to a common underlying factor, do not support the assertion that the two instruments are assessing a common, or overly similar, construct.

5.5.6 Rebuttal 4.2: Constituent instrument utility issues

The second decisions rebuttal addresses the possibility that one or more of the constituent instruments demonstrates utility issues which could raise concerns regarding its usefulness. Two sources of information were used to evaluate this rebuttal: score frequency distributions, and a review of the cut scores established at the institution in order to differentiate students into various ability categories.

Looking first to the writing sample, as seen in Figure 5.5, the scores from the 1504 candidates to have completed the placement instrument from Fall 2008 to Fall 2011 range across the entire spectrum of possible results. The estimate of skewness (.068, standard error of skewness .063) suggests the distribution of scores is normal.

Table 5.13 shows the cut scores for the writing sample, established in order to be able to differentiate applicants into the various placement categories. Cut scores are spread relatively evenly across the range of results possible, indicating no evidence of decisions being made based upon restricted ranges of scores, which can increase the likelihood of placement decision errors.

Figure 5.5: Frequency Distribution for Candidate Writing Sample Scores

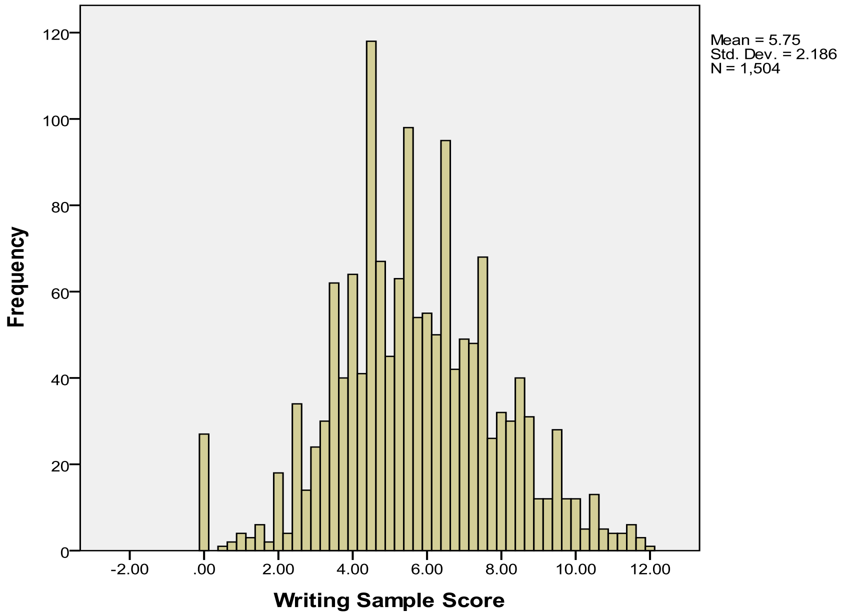


Table 5.13 Cut-scores established for the writing sample

Cut Score Range	Placement Recommendation
10 - 12	Credit English
8 - 9.9	Developmental English Level 1
6 – 7.9	Developmental English Level 2
4 – 5.9	Developmental English Level 1
0 – 3.9	Not ready for any CMI English course

Figure 5.6 shows the distribution of results for the 2117 candidates to have completed Accuplacer Companion from the Fall 2008 to Fall 2011 semester fall primarily towards the lower end of the range of possible scores.

Mean (20), median (19) and mode (17) results are very low for a test with 70 total items. The estimate of skewness (1.281), relative to the standard error of skewness (.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 5.14, established by the institution in order to separate candidates into placement categories.

Figure 5.6: Frequency Distribution for Accuplacer Companion

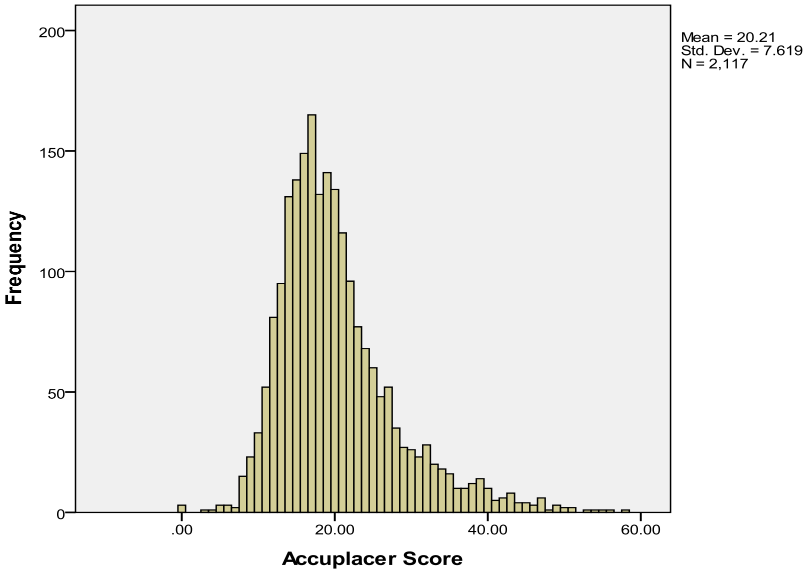


Table 5.14: Cut Scores Established for Accuplacer Companion

Cut Score Range	Placement Recommendation
43-70	Credit English
37-42	Developmental English Level 1
30-37	Developmental English Level 2
15-29	Developmental English Level 1
0-14	Not ready for any CMI English course

While the cut scores reported above are those used by the college since the adoption of Accuplacer Companion, 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 intended by Institutional Research, to those reported above. While the report does not indicate the original values, it does state the need for reducing the cut scores 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: i) admit sufficient numbers of the applicants to the school for that semester so as to avoid a substantial drop in enrolment numbers; and ii) place at least some of these new applicants in Levels 2 and 3 of the Developmental English program.

From the Table 5.14, we see that some placement categories are associated with very small score arrays. Ranges for Developmental English Levels 2 and 3 are only seven and five points wide, for example. Given that the standard error measurement for the instrument is 3.80 (i.e., a student with a score of 35 would likely get a score between 31 and 39, approximately 68% of the time, or a score between 27 and 43, 95% of the time), these are likely to be dangerously restricted ranges upon which to base high-stakes decisions about candidates, and ones which may result in a substantial number of placement errors. For AC, then, the evidence would seem to lend credence to Rebuttal 4.2.

5.6 Consequences

claim

The entire purpose of using assessments is for the benefit of various, hopefully all, stakeholders. Perhaps the most important claim that needs to be established and supported, then, is that the PAS and its constituent instruments, and the decisions informed by them, result in beneficial consequences for all affected. Three warrants and two rebuttals were developed to inform the claim, each of which are presented in turn.

5.6.1 Warrant 5.1: Beneficial consequences for individual stakeholders

The first warrant claims the PAS results in beneficial consequences for applicants, students and instructors. Evidence to inform this warrant came from two sources: student success rates in the English courses into which they were placed, and stakeholder opinion.

5.6.1.1 Consequences for applicants

Items included in the questionnaires for CMI English instructors and for test-takers who had completed Accuplacer and the writing sample were intended to gather stakeholder opinion regarding the instrument’s impact on examinee. Table 5.16 summarises participants’ responses and presents chi-square tests for significance.

Table 5.16: Stakeholder opinions regarding the impact of the placement instruments on examinees

	Questionnaire Item	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
Instructors	The Accuplacer Companion English subtests will have a positive impact on students' perceptions of themselves and their English language skills.	Agree❖	0	0	.50	**
		Disagree+	10	1.00		
		Total	10	1.00		
	The Accuplacer Companion English subtests will have a negative impact on students' desire to pursue a postsecondary education at CMI or another institution.	Agree❖	10	.91	.50	.007*
		Disagree+	1	.09		
		Total	11	1.00		
	The Writing Sample will have a positive impact on students' perceptions of themselves and their English language skills.	Agree❖	5	.63	.50	.480*
		Disagree+	3	.38		
		Total	8	1.00		
	The Writing Sample will have a negative impact on students' desire to pursue a postsecondary education at CMI or another institution.	Agree❖	7	.78	.50	.096*
		Disagree+	2	.22		
		Total	9	1.00		
Test-takers	Taking the Accuplacer Companion English test made me think I can be a successful student at CMI.	Agree❖	94	.77	.50	.000
		Disagree+	28	.23		
		Total	122	1.00		
	Taking the Accuplacer Companion English test made me feel good about my English abilities.	Agree❖	93	.76	.50	.000
		Disagree+	30	.24		
		Total	123	1.00		
	Taking the Accuplacer Companion English test made me want to study at CMI.	Agree❖	121	.86	.50	.000
		Disagree+	20	.14		
		Total	141	1.00		
	The Writing Sample made me think I can be a successful student at CMI.	Agree❖	98	.80	.50	.000
		Disagree+	25	.20		
		Total	123	1.00		
	The Writing Sample made me feel good about my English abilities.	Agree❖	89	.83	.50	.000
		Disagree+	18	.17		
		Total	107	1.00		
	The Writing Sample made me want to study at CMI.	Agree❖	120	.86	.50	.000
		Disagree+	19	.14		
		Total	139	1.00		

❖Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
*Assumption of minimum 5 participants in each cell violated
** Unable to compute due to 0 participants in one cell

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 higher education. Responses do not indicate the same significant majority concern regarding negative washback of the WS on test-takers.

Looking to the responses of the examinees themselves, they do not report experiencing negative effects due to completing either instrument, with regard to their perceptions of their language abilities, likelihood of being successful at CMI, or desire to pursue a higher education. As the warrant pertains to examinee perceptions, firsthand feedback was felt to take precedent over the perceptions of instructors, and thus the evidence would seem to support Warrant 5.1 for both instruments.

5.6.1.2 Consequences for new students

As mentioned in the introductory chapter, one of the impetuses for the current study was the concern that many new students were being misplaced by the PAS and that this might be at least part of the reason for low pass, completion, and retention rates for new students at the college. In the Fall 2009 semester, for example, final grade distributions in Developmental Reading and Writing courses and credit level Composition courses indicated only 46% of new students achieved a grade of C+ or higher, and some 37% did not pass, making a failing grade the most common result for new students. These results did not include students who ‘dropped' the course or were withdrawn due to lack of attendance, which are also identified as problems at the college (CMI, 2008, 2009).

From Fall 2009 to Fall 2011, near the end of each semester, first-semester students were asked to complete a questionnaire intended to gain insights on their perceptions of the consequences of their placement as a result of the PAS. Table 5.17 provides a summary of the results, including chi-square tests for statistical significance.

The significant, often approaching unanimous, majority of students report no problems understanding their instructors or what is required of them in their English courses, and are confident in their ability not only to pass, but also achieve an A or a B in both the RW and LS courses they were placed into. Some 78% and 83% of new students reported feeling they were placed in the correct LS and RW course, respectively, for their abilities.

Curiously, however, similarly large ratios of respondents stated they felt their LS (68%) and RW (72%) course was too easy for them. It is unclear why many respondents would report being in the correct level and also report being in a level below where they feel they should have been placed. While not unsubstantial, comparatively small number of students (15%) reported being in a LS or RW level too difficult for them.

Table 5.17: First-semester student questionnaire responses

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
I am doing well in my English classes.	Agree❖	63	.95	.50	.000
	Disagree+	3	.05		
	Total	66	1.00		
I think I am in the right Listening and Speaking class for my ability.	Agree❖	49	.78	.50	.000
	Disagree+	15	.23		
	Total	64	1.00		
I think my Listening and Speaking class is too difficult for me.	Agree❖	10	.15	.50	.000
	Disagree+	57	.85		
	Total	67	1.00		
I think my Listening and Speaking class is too easy for me.	Agree❖	39	.68	.50	.024
	Disagree+	18	.32		
	Total	57	1.00		
If I do my best I can pass my Listening and Speaking class.	Agree❖	76	.97	.50	.000
	Disagree+	2	.03		
	Total	78	1.00		
If I do my best I can get an A or B in my Listening and Speaking class.	Agree❖	70	.95	.50	.000
	Disagree+	4	.05		
	Total	74	1.00		
I am in the right Reading and Writing class for my ability.	Agree❖	54	.83	.50	.000
	Disagree+	11	.17		
	Total	65	1.00		
I think my Reading and Writing class is too difficult for me.	Agree❖	10	.15	.50	.000
	Disagree+	57	.85		
	Total	67	1.00		
I think my Reading and Writing class is too easy for me.	Agree❖	43	.72	.50	.005
	Disagree+	17	.28		
	Total	60	1.00		
If I do my best I can pass my Reading and Writing class.	Agree❖	80	.99	.50	.000
	Disagree+	1	.01		
	Total	81	1.00		
If I do my best I can get an A or B in my Reading and Writing class.	Agree❖	67	.92	.50	.000
	Disagree+	6	.08		
	Total	73	1.00		
I can understand what my teachers want me to do in my English classes.	Agree❖	88	1.00	.50	.000
	Disagree+	0	.00		
	Total	88	1.00		
I can do what my teachers ask me to do in my English classes.	Agree❖	88	.99	.50	.000
	Disagree+	1	.01		
	Total	89	1.00		
I can pass my English classes this semester.	Agree❖	79	.98	.50	.000
	Disagree+	2	.02		
	Total	81	1.00		

❖Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses

Comments offered by respondents in the open-ended section of the questionnaire, however, were largely negative, and generally reported frustration with being placed in courses that were too easy. Only one comment offered, “Though I feel I should have started… with Credit levels of English… [Level 3 classes] refreshed my memory and now credit levels will be much more understanding”, suggested satisfaction with the placement decision made for them, though this student also reported feeling they should have been placed in English courses one level above where they were. The remaining comments indicate frustration with being placed in courses that are too easy:

- "…all these things I learn in Level 1, I already learned them from high school"
- "It's boring!"
- "I think my English class is to easy for me"
- "I think I should have started in English 96/98 [Level 3]"
- "I don't want to learn things I've already knew. I want to learn new things and I want to have a challenge in my English classes. I know every student wants a challenge."
- "I have a BIG suggestion to make! Correct/revise your (CMI's) placement test! It has caused me GREAT STRESS by making my time & financial aid such a waste for this fall semester 09!"

Overall, the results of the evidence considered would appear somewhat difficult to reconcile. Should low success, completion, and retention rates be a result of over- or underwhelmed students misplaced by the PAS, we might expect the majority of students to report being in the wrong level for their abilities. While this is the case, perhaps, with the majority reporting they are in a RW and/or LS course that is too easy for them, even larger majorities report being in the right level for their abilities. These results are somewhat confusing, in that many of the same respondents must be reporting both being in the correct level and also being in a level too easy for them. Perhaps many feel they are in a level that is acceptable, but believe they could have coped with the level above the one they were placed into. In either event, the vast majority of students reporting not only being able to pass but also achieve an A or B in the course they were placed into does not support the position that most students are frustrated and/or floundering due to misplacement.

5.6.1.3 Consequences for Instructors

Two sources of evidence were considered for this aspect of the warrant: estimates of student misplacement, according to instructor feedback; and faculty opinion regarding the usefulness of each instrument for placing incoming students.

From Fall 2009 to Fall 2011, towards the end of each semester, English instructors with first-semester students in their classes were asked where they felt, based solely on the English language skills relevant to the course – i.e., listening and speaking, or reading and writing abilities – each new student should have ideally been placed. Table 5.18 presents a summary of the findings.

Table 5.18: Estimates of correct placement according to instructors' opinions

Students placed…	LS	%	RW	%
In a level too difficult	62	15.78%	59	13.59%
In the correct level	223	56.74%	294	67.74%
In a level too easy	107	27.23%	80	18.43%
In a level they cannot pass	36	9.2%	52	11.9%
All misplaced	169	43.00%	139	32.03%
n	393		434	

LS = Listening & Speaking course
RW=Reading & Writing course

According to instructor opinion, then, substantial numbers of new students are being misplaced by the current PAS, with estimates of 43% in LS courses and 32% in RW courses being misplaced. While the ratio of correctly placed students according to instructor opinion is far lower than those derived from new student opinion, the amount of students placed in levels too high for their abilities approximates 15% for both courses and for both groups of stakeholders.

Various sections of this paper have reported the negative perspectives of many faculty members towards one constituent instrument of the PAS, Accuplacer Companion. English instructors widely felt, for example, that the instrument texts were (due to complexity or content) likely confusing to most test-takers, that it did not address skills relevant to the instructional domain, and that it likely had negative washback effects on test-takers' perceptions of their language abilities and desires to pursue a higher education.

Opinions regarding the writing sample were largely the opposite (with the exception that it was not felt to negatively impact students' self-perceptions, as opposed to improving them). It is perhaps unsurprising, then, that instructors responses to questionnaire items (summarized in Table 5.19) indicate widely held perceptions that Accuplacer is not useful for informing admissions or placement decisions, but the writing sample is.

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 overwhelmed or underwhelmed, and the mixed abilities 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. Further, it was quite clear 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 5.19: Instructor opinion regarding the usefulness of constituent placement instruments

	Item	Group	N	Observed prop.	Test prop.	Exact Sig. (2-tailed)
Accuplacer Companion	The Accuplacer Companion English subtests are useful for choosing which applicants are able to enroll in English courses at CMI.	Agree❖	1	.09	.50	.007*
		Disagree+	10	.91		
		Total	11	1.00		
	The Accuplacer Companion English subtests are useful for placing incoming students in the Developmental or Credit level English classes best suited for their current language abilities.	Agree❖	2	.18	.50	.035*
		Disagree+	9	.82		
		Total	11	1.00		
Writing Sample	The Writing Sample is useful for choosing which applicants are able to enroll in English courses at CMI.	Agree❖	9	.90	.50	.011*
		Disagree+	1	.10		
		Total	10	1.00		
	The Writing Sample is useful for placing incoming students in the Developmental or Credit level English classes best suited for their current language abilities.	Agree❖	11	.92	.50	.004*
		Disagree+	1	.08		
		Total	12	1.00		

❖Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
*Assumption of minimum 5 participants in each cell violated

Overall, results indicate that instructors view the PAS, particularly Accuplacer Companion, as negatively impacting themselves by misplacing several new students each semester.

5.6.2 Warrant 5.2: Confidentiality of results

CMI policy establishes that the results of the PAS and its constituent instruments 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 Chair of Developmental English, they may be allowed to review the writing sample of the student in order to help inform their opinion as to the accuracy of the students' placement. The department chair and the instructor of the level into which the student might be moved may also review the writing sample if both the current instructor and the student as well, decide changing levels would be in the best interest of the student. It should be noted that current policy only allows for students to be moved if they agree. As such, students would have to give their permission for anyone other than their current instructor to be able to review their writing sample, as the ad hoc committee would not be formed if students' did not wish to pursue the change in level.

Overall, then, this policy and the re-placement process would seem to assure the confidentiality of placement decisions and the scores of individual students on both placement instruments.

5.6.3 Warrant 5.3: Promotion of effective teaching and learning

The final consequences warrant asserts that the PAS and/or its constituent instruments promote effective teaching and learning. While first-semester students report fewer instances of initial misplacement, indicating 80% of new pupils feel they have been placed into courses appropriate for their abilities compared with instructors' 60-70%, both groups of stakeholders indicate approximately 15% of incoming students are placed in courses too difficult for them. Instructors, during the focus group interview, complained of mixed abilities classes, "scrambles" at the beginning of every semester to identify and move misplaced students while class enrolment can still be changed, and a number of first-semester students being either under-challenged or, worse, having little chance of success.

From questionnaire responses and focus group comments, clearly faculty members view the problem to be largely Accuplacer, and not the writing sample.

While these are matters of perception, low predictive validity estimates, highly skewed results, and resultant necessity to use restricted ranges of scores for placement decisions (and the greater risk of misplacement as a result), are not. Results, then, would not seem to support Warrant 5.3 for Accuplacer.

With regard to the WS, instructors clearly hold that the instrument is relevant to the requirements of the instructional domain, and is useful for informing placement students. Examinee opinions also are largely positive towards the instrument. Despite these positive perceptions, however, the writing sample's lack of predictive capacity with regard to final course results, and issues with inter-rater reliability and rating scale function, suggest it is also problematic. As a result, it, too, cannot be said to be having a positive effect on teaching and learning at the institution.

5.7 Summary of evidence regarding utility claims for the PAS and constituent instruments

As was presented at the end of the claims addressing test score interpretation, Table 5.20 provides a summary of the findings relating to each of the utilization claims and warrants for the overall PAS and, where appropriate, the individual placement instruments as well.

As Table 5.20 shows, the only utilization warrants supported by the evidence, for individual placement tests or the overall PAS, were the confidentiality of the results and consistency of the instruments, processes and policies used to place all candidates. As a result, neither the decision nor the consequences claim was supported for the PAS or either placement instrument used.

Table 5.20: Summary of findings for the test utilization claims

#	Claim/Warrant/Rebuttal	Evidence supports the warrant/claim for:		
		Writing Sample	Accuplacer	PAS
4	**Decisions Claim**: Placement decisions are equitable, values sensitive, and based on evidence that is sufficient and useful.	n/a	n/a	✗
4.1	The same instrument, processes, and policies are utilized to inform placement decisions for all applicants	n/a	n/a	✓
4.2	Applicants are fully informed of the placement decision process.	n/a	n/a	✗
4.3	Numerous stakeholders, representative of all affected by the PAS, were consulted during the establishment and/or review of the placement process.	n/a	n/a	✗
4.4	The constituent methods of the PAS combine to account for substantial variance in first-semester students' performance in relevant courses.	n/a	n/a	✗
R4.1	Rebuttal: Two or more constituent assessments assess the same or problematically similar constructs	✓	✓	✓
R4.2	Rebuttal: One or more constituent assessments demonstrate utility issues substantial enough to raise concerns regarding usefulness	✗	✗	✗
5	**Consequences Claim**: Use of the PAS and its constituent methods, and decisions informed by them, result in beneficial consequences for all stakeholders.	✗	✗	✗
5.1	The PAS results in beneficial consequences for applicants, new students, and instructors.	✗	✗	✗
5.2	Results of the PAS and constituent assessments are confidential	✓	✓	✓
5.3	The PAS and its component instruments promote effective teaching and learning, making it beneficial to students, instructors, and program(s) affected	✗	✗	✗

✓= results back the warrant (or refute the rebuttal)
✗ = results refute the warrant (or back the rebuttal)
n/a = not applicable for individual instrument

Having completed the review of results for all warrants and claims, the next chapter shall discuss the findings in light of the specific research questions established, and offer some conclusions regarding the findings, process, and implications for future research.

CHAPTER SIX

Summary, Discussion and Conclusion

6.0 Introduction

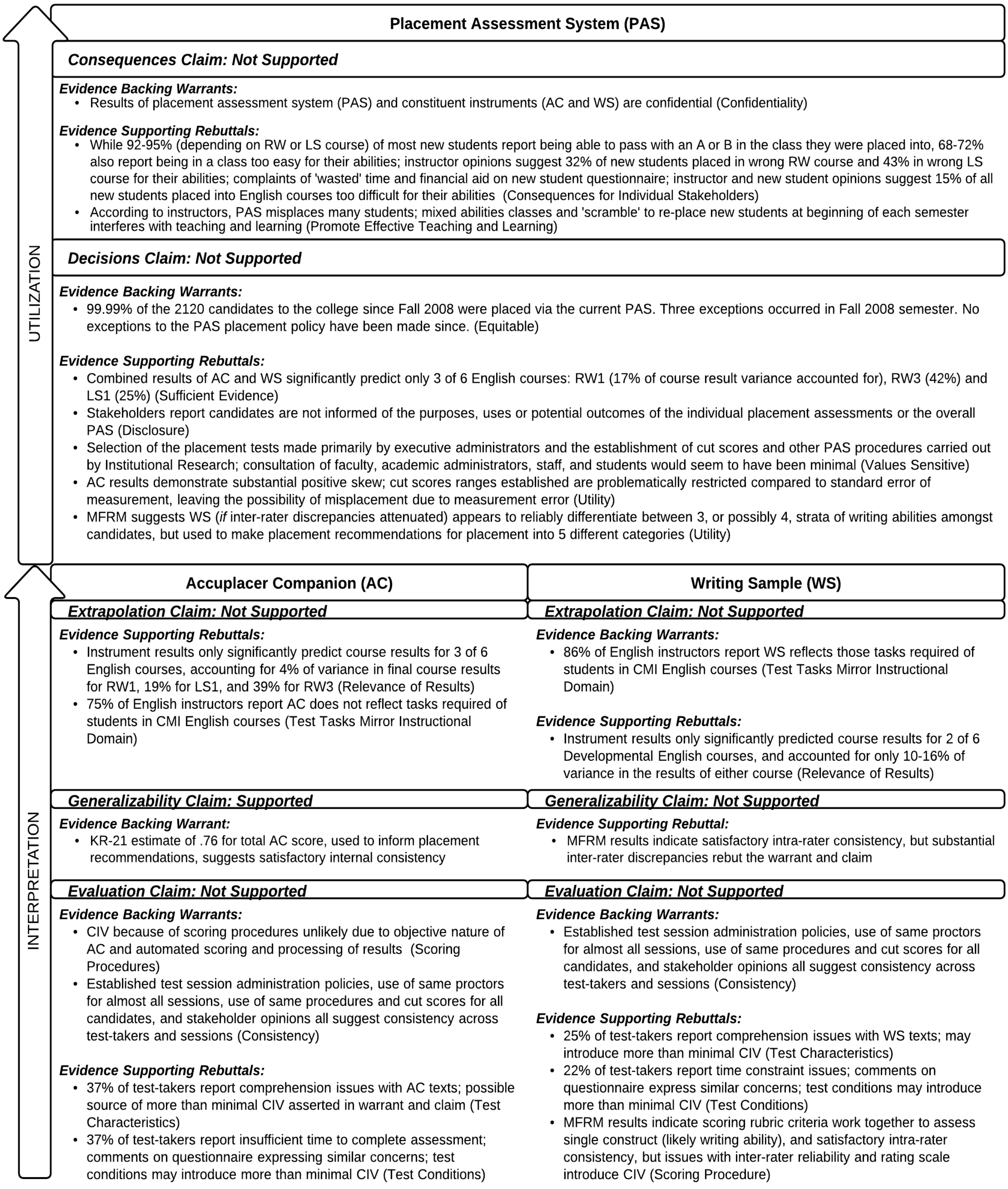
This chapter will provide a brief summary of the results for the validation framework for both placement instruments and, where applicable, the overall placement assessment system (PAS). Subsequently, each of the research questions articulated in Chapter One will be addressed, in light of the evidence considered in the current study. The chapter ends with a discussion of future implications for future research, followed by concluding remarks.

6.1 Summary of the Results

The primary purpose of this study was to construct a validity argument and apply it to the current instruments and overall PAS used by the College of the Marshall Islands to exempt, exclude or place students within their Developmental English program. Figure 6.1 provides a summary of the validation framework for the PAS and constituent instruments, and the main findings informing the warrants and rebuttals relating to each claim.

Specific results of the investigation will be discussed in detail in the relevant research questions in the following sections of the chapter. Overall, however, the outcomes indicate significant problems in a variety of areas for both placement instruments and the overarching PAS, with the generalizability claim for Accuplacer Companion being the only claim in the framework supported by the evidence.

Figure 6.1: Validation Framework Restated in Light of Evidence



Placement Assessment System (PAS)

Consequences Claim: Not Supported

Evidence Backing Warrants:

- Results of placement assessment system (PAS) and constituent instruments (AC and WS) are confidential (Confidentiality)

Evidence Supporting Rebuttals:

- While 92-95% (depending on RW or LS course) of most new students report being able to pass with an A or B in the class they were placed into, 68-72% also report being in a class too easy for their abilities; instructor opinions suggest 32% of new students placed in wrong RW course and 43% in wrong LS course for their abilities; complaints of 'wasted' time and financial aid on new student questionnaire; instructor and new student opinions suggest 15% of all new students placed into English courses too difficult for their abilities (Consequences for Individual Stakeholders)
- According to instructors, PAS misplaces many students; mixed abilities classes and 'scramble' to re-place new students at beginning of each semester interferes with teaching and learning (Promote Effective Teaching and Learning)

Decisions Claim: Not Supported

Evidence Backing Warrants:

- 99.99% of the 2120 candidates to the college since Fall 2008 were placed via the current PAS. Three exceptions occurred in Fall 2008 semester. No exceptions to the PAS placement policy have been made since. (Equitable)

Evidence Supporting Rebuttals:

- Combined results of AC and WS significantly predict only 3 of 6 English courses: RW1 (17% of course result variance accounted for), RW3 (42%) and LS1 (25%) (Sufficient Evidence)
- Stakeholders report candidates are not informed of the purposes, uses or potential outcomes of the individual placement assessments or the overall PAS (Disclosure)
- Selection of the placement tests made primarily by executive administrators and the establishment of cut scores and other PAS procedures carried out by Institutional Research; consultation of faculty, academic administrators, staff, and students would seem to have been minimal (Values Sensitive)
- AC results demonstrate substantial positive skew; cut scores ranges established are problematically restricted compared to standard error of measurement, leaving the possibility of misplacement due to measurement error (Utility)
- MFRM suggests WS (*if* inter-rater discrepancies attenuated) appears to reliably differentiate between 3, or possibly 4, strata of writing abilities amongst candidates, but used to make placement recommendations for placement into 5 different categories (Utility)

Accuplacer Companion (AC)

Extrapolation Claim: Not Supported

Evidence Supporting Rebuttals:

- Instrument results only significantly predict course results for 3 of 6 English courses, accounting for 4% of variance in final course results for RW1, 19% for LS1, and 39% for RW3 (Relevance of Results)
- 75% of English instructors report AC does not reflect tasks required of students in CMI English courses (Test Tasks Mirror Instructional Domain)

Generalizability Claim: Supported

Evidence Backing Warrant:

- KR-21 estimate of .76 for total AC score, used to inform placement recommendations, suggests satisfactory internal consistency

Evaluation Claim: Not Supported

Evidence Backing Warrants:

- CIV because of scoring procedures unlikely due to objective nature of AC and automated scoring and processing of results (Scoring Procedures)
- Established test session administration policies, use of same proctors for almost all sessions, use of same procedures and cut scores for all candidates, and stakeholder opinions all suggest consistency across test-takers and sessions (Consistency)

Evidence Supporting Rebuttals:

- 37% of test-takers report comprehension issues with AC texts; possible source of more than minimal CIV asserted in warrant and claim (Test Characteristics)
- 37% of test-takers report insufficient time to complete assessment; comments on questionnaire expressing similar concerns; test conditions may introduce more than minimal CIV (Test Conditions)

Writing Sample (WS)

Extrapolation Claim: Not Supported

Evidence Backing Warrants:

- 86% of English instructors report WS reflects those tasks required of students in CMI English courses (Test Tasks Mirror Instructional Domain)

Evidence Supporting Rebuttals:

- Instrument results only significantly predicted course results for 2 of 6 Developmental English courses, and accounted for only 10-16% of variance in the results of either course (Relevance of Results)

Generalizability Claim: Not Supported

Evidence Supporting Rebuttal:

- MFRM results indicate satisfactory intra-rater consistency, but substantial inter-rater discrepancies rebut the warrant and claim

Evaluation Claim: Not Supported

Evidence Backing Warrants:

- Established test session administration policies, use of same proctors for almost all sessions, use of same procedures and cut scores for all candidates, and stakeholder opinions all suggest consistency across test-takers and sessions (Consistency)

Evidence Supporting Rebuttals:

- 25% of test-takers report comprehension issues with WS texts; may introduce more than minimal CIV (Test Characteristics)
- 22% of test-takers report time constraint issues; comments on questionnaire express similar concerns; test conditions may introduce more than minimal CIV (Test Conditions)
- MFRM results indicate scoring rubric criteria work together to assess single construct (likely writing ability), and satisfactory intra-rater consistency, but issues with inter-rater reliability and rating scale introduce CIV (Scoring Procedure)

LS1 = Level 1 Listening and Speaking course; RW1 = Level 1 Reading and Writing course; RW3 = Level 3 Reading and Writing course
CIV = construct-irrelevant variance; PAS = Placement Assessment System; AC = Accuplacer Companion; WS = writing sample

6.2 Research Questions

Five research questions were articulated to address areas of insight believed to be of direct importance to CMI, and likely to other institutions using standardised instruments and/or writing samples to inform placement decisions.

6.2.1 Research Question (RQ) 1: Relative advantages regarding score interpretation

With regard to test score interpretation, what do the results of the study indicate relating to: i) the relative advantages of AC and WS in the CMI context; and ii) current debates in the literature regarding standardised instruments and WS's for placement purposes? The evidence relating to the three elements of test score interpretation addressed in this study – evaluation, generalizability, and extrapolation – will be presented and discussed in turn.

6.2.1.1 Relative advantages regarding evaluation

Two areas of investigation, test characteristics – specifically, instrument texts – and scoring procedures of the two instruments, were examined as potential sources of construct irrelevant variance (CIV) and resultant threats to the evaluation claim. While an additional potential source of CIV via test conditions, time allotment, was also investigated, this was considered to be a factor related to the administration of the instruments, and not an area in which one instrument would potentially offer an advantage over the other. As evidence did uncover potential problems, however, this aspect of the study will be discussed later, when considering potential means of improving the PAS.

6.2.1.1.1 Relative advantages regarding text comprehension

While a larger portion of examinees reported problems understanding the texts of AC (25% of all respondents, 37% of respondents who did not select "neither agree nor disagree") than the locally developed WS texts (18% overall, 22% of non-neutral respondents), both results were considered too large to likely represent the 'minimal' source of CIV allowed for in the test characteristics warrant and overall evaluation claim. As such, neither can really be said to offer an advantage in assuring issues related to accessibility of the texts do not interfere with test score interpretation.

Given that AC is intended for use with native speakers of the English language, which candidates to the CMI almost uniformly are not, the sizable percentage of examinees reporting difficulties understanding the instruments' texts is unsurprising. The finding corroborates concerns in the literature regarding CIV and validity issues due, at least in part, to comprehension problems when instruments designed for native speakers are used with English Language Learners (ELLs) (Crawford, 2004; Solorzano, 2004). With regard to the WS, however, the relatively large portion of examinees reporting comprehension issues was somewhat unexpected, particularly since the instrument texts are created by writing instructors with experience teaching and assessing Marshallese ELLs, and reviewed by an external consultant with expertise in assessing writing. None of the literature reviewed for this study would seem to address the issue of potential CIV introduced via locally developed WS texts, other than perhaps variance in familiarity with the topic of the prompt. It would seem the widely held presumption is that instructions designed with specific examinees in mind, by those with experience working with said examinee population, will be readily accessible to test-takers. At CMI, and perhaps elsewhere, this would appear to be a presumption that needs to be rethought.

6.2.1.1.2 Relative advantages regarding test scoring

One area of clear advantage, at least as the two instruments are currently employed at CMI, was the likely amount of CIV introduced by the scoring procedures of the two instruments. Given the objective nature of AC, lack of apparent issues with problematic questions, options, or answers, and automated scoring procedures, little threat of substantial CIV being introduced in via the scoring system was perceived. With the WS, however, multi-facet Rasch measurement (MFRM) outcomes indicated that, while raters were found to be internally consistent, significant and substantial CIV was introduced by inter-rater discrepancies (fixed, all-same, X^2=596.3, df=14, p=.00). Further, as CMI does not attenuate such variance across raters, either through mathematical modeling, such as Rasch analysis, or other means, this influence was found to demonstrably influence both observed scores and resulting placement recommendations based on those scores.

These results confirm the widely held position that the objective, closed-ended nature of standardised instruments results in little CIV due to the scoring procedures, while open-ended, judged performances are prone to potentially validity-threatening CIV, particularly from inconsistencies in rater behaviour.

However, MFRM results also affirm the position that rater variance can be attenuated and, if it is, performance assessments can reliably differentiate amongst examinees based on the construct intended. Results of the Rasch analysis, which attended to inter-rater discrepancies in the model, indicated that the WS reliably (reliability index = .91) differentiated between different strata of abilities amongst candidates. Further, none of the individual criteria on the scoring rubric used to assess the placement essays was found to be contributing CIV, indicating they worked well collectively to assess a single construct: writing ability. Results, then, substantiate both the dangers of performance assessment scoring without attending to differences in rater behaviour, and that such assessments can reliably differentiate amongst candidate ability if steps are taken to mitigate this source of CIV (McNamara, 1996; Park, 2004).

MFRM outcomes also uncovered another source of CIV in WS results, problematic functioning of the scoring rubric rating scale – a scale of 1-6 for each of the criteria – possibly due to differential understanding and/or application of the various categories across raters. None of the articles reviewed, other than those utilizing Rasch analyses, addressed rating scale issues as a potential source of CIV in performance assessments. For those that did, none found problematic functioning. It would seem that this is an area in which more widespread investigation is needed.

6.2.1.1.3 Relative advantages regarding generalizability

While raters were found to be internally consistent, and the potential for reliable outcomes for the WS was evident from the MFRM results, as CMI does not currently attend to inter-rater discrepancies in their use of the assessment, the test scores used by the institution are demonstrably unreliable. The internal consistency estimate for AC results was found to be .76. While this is short of the .80 traditional crucible of acceptable reliability, it was considered sufficient given the conservative nature of the KR-21 formula utilized. (Sections 1.5 and 4.4.1 provide details on why this method was necessary.)

Results align with the widely held view of closed-ended, objective, standardised instruments with substantial numbers of test items holding the advantage of consistency over open-ended, judged performance assessments which are prone to inconsistency in rater behaviour. The evidence reported in the literature suggests very high estimates of consistency for standardised tests such as Accuplacer (College Board, 2003; Mattern & Packman, 2009) or IELTS (Cambridge ESOL, 2007); levels which writing sample results may be unlikely to approximate simply because of their open-ended nature and the subjective evaluation of individual raters (East, 2009; Haswell, 2005; Jonsson & Svingby, 2007; Lumley, 2002; Rezaei & Lovorn, 2010; Stemler, 2004).

However, while a number of studies have reported problematic inconsistency between WS raters (Brown, et al., 2004; Jonsson & Svingby, 2007), others report reassuringly high inter-rater reliability when steps are taken to address rater consistency (East, 2009; James & Templeman, 2009; Jonsson & Svingby, 2007; Matzen & Hoyt, 2004; Park, 2004). Using MFRM, results for the candidate ability facet included indices of reliability (.91), separation (3.18) and strata (4.58) suggesting that the assessment reliably separated examinees into three or possibly four categories of writing ability. As Rasch estimates of reliability represent the reproducibility of the outcomes, given a similar sample of examinees, this is evidence for the claim of generalizability of the assessment, *if* CIV from facets such as inter-rater variance are attenuated.

6.2.1.3 Relative advantages regarding extrapolation

Instructor opinions clearly indicate the perception that the WS task is relevant to the requirements of CMI English courses (86% agree, p=.01) while the majority do not feel the same about AC items and tasks (75% disagree, though not statistically significant, p=.08). This perceived advantage, however, was not borne out by the evidence considered, as neither instrument's results demonstrated satisfactory relevance to the requirements of the instructional domain at CMI. AC scores were found to significantly predict final course outcomes in only three of the six Developmental English courses at CMI (r^2= .04, .19, and .39), and results of the WS accounted for significant variance in only two courses (r^2=.10 and .16).

While the one predictive validity estimate of .39 (for the Level 3 Reading and Writing course) is higher than most other studies report, overall results for AC roughly match the rather limited, or entirely lacking, predictive capacity found in other investigations involving Accuplacer (CCCAA, 2007a, 2007b, 2008; College Board, 2003; Mattern & Packman, 2004), and various other standardised instruments (Armstrong, 2000; Behrman, 2000; Goodman, Freed & McManus, 1990; Hill, et al., 1999; Holderer, 1992; Hughes & Nelson, 1991; Isonio, 1991, 1992; James & Templeman, 2009; Rasor & Barr, 1995; Sullivan & Nielsen, 2009).

Reasons often offered for the generally poor predictive validity of standardised tests include their closed-ended nature, necessarily general scope (so as to be usable at a wide range of programs), indirect assessment of language skills, and resulting limited reflection of the complex requirements of the instructional domain (Armstrong, 2000; Behrman, 2000; Murphy & Yancey, 2008; Williams, 1990). Conversely, these are all areas of suggested advantage for performance assessments, which, because of their open-ended nature and direct assessment of abilities (Hughes, 2003), are argued to better reflect the complex cognitive processes and specific competencies required of students in the courses into which they are being placed (East, 2009; Jonsson & Svingby, 2007; Kim, 2008; Matzen & Hoyt, 2004; Messick, 1989; Stoynoff, 2009).

Results from the current study, however, do not support this position, at least not as the WS is currently employed at CMI. Given the reported issues with inter-rater reliability and the rating scale, though, we might expect that prediction of course outcomes would be limited at best. However, perhaps surprisingly given the frequent offering of instructional domain relevance as an argument for the use of writing samples for placement, convergence of WS results with course outcomes would seem to have been infrequently investigated and/or published, particularly of late. In the only two relevant studies found which were conducted within the past decade, Matzen and Hoyt (2004) report locally developed and marked essays significantly predicted final course results, and did so to a greater extent than the standardised placement tests included in the study, while May (2007) found no significant predictive capacity for WS's. It would appear, then, that the widely offered advantage of relevance of the WS task to the instructional domain requires more empirical evidence to corroborate its rational appeal.

6.2.2 RQ 2: Relative Advantages Regarding Utilization

With regard to issues of utilization, what do the results of the study indicate relating to: i) the relative advantages of AC and WS in the CMI context; and ii) current debates in the literature regarding standardised instruments and WS's for placement purposes? The evidence relating to the three elements of assessment utilization addressed in this study – decisions, specifically sufficiency and utility issues, and consequences for the individual stakeholders and overall teaching and learning at the program – will be presented and discussed in turn.

6.2.2.1 Relative advantages regarding decisions

Investigations into the decisions claim for the PAS addressed two areas useful for informing the relative advantages of AC and the WS for placement decisions at CMI: sufficiency and utility.

6.2.2.1.1 Relative advantages regarding sufficiency

While the use of results from both instruments was found to be a better predictor than either instrument alone, for all six Developmental English courses, results still only significantly predicted outcomes in the same three courses as AC outcomes alone: Reading and Writing Level 1 ($r^2 = .17$, p=.00) and Level 3 ($r^2 = .42$, p=.00), and Listening and Speaking Level 1 ($r^2 = .25$, p=.00). As such, the combination of the two instruments, which, along with possession of a high school diploma or GED certificate, represents the entirety of the information upon which placement decisions are based, cannot be argued to be sufficient evidence upon which to base placement decisions for students entering into the language program.

The finding that the combination of WS and standardised test results offers better prediction of students' course results than either assessment alone corroborates results from a variety of other studies reporting the same (Breland, et al., 1987; Galbraith, 1986; Garrow, 1989; Isonio, 1994; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998). Further, the finding supports arguments from educational testing organizations, test publishers, and researchers alike that a consideration of a variety of types and sources of evidence should inform placement decisions (AERA, APA & NCME, 1999; Bachman & Palmer, 1996; Board of Governors of the California Community Colleges, 2008; Camara & Lane, 2006; College Board, 2003; Lynch, 2001; Solorzano, 2008).

6.2.2.1.2 Relative advantages regarding utility

Both instruments were found to demonstrate problematic utility issues. For AC, results were significantly positively skewed, to the extent that decisions were being made about candidates based on a restricted range of scores. In order to be able to discriminate amongst students within this restricted range, cut score ranges for the instrument results were also extremely small, approximating the standard error of measurement for the instrument and increasing the likelihood of misclassifications due to measurement error alone.

While WS results did not appear to demonstrate this problem of skewed and/or restricted range of scores, the evidence did suggest a somewhat similar utility issue. MFRM outcomes for the candidate ability facet suggest that the assessment reliably discriminates amongst three, possibly four, strata of candidate ability (if rater variance was attenuated in the model). As assessment recommendations consist of five possible classifications, however, the assessment is perhaps being used to make distinctions amongst candidates that are finer than it reliably is able to deliver.

The dangers of making decisions about examinees based on restricted ranges of results, and of using instruments designed for native speakers with ELLs (Crawford, 2004; Solorzano, 2008) are widely acknowledged. However, there would appear to be little discussion in the literature of rating scale utility issues, such as those uncovered here and reported earlier in the chapter with regard to the apparent inconsistent application of the rating scale categories by raters. As rating scale functioning certainly impacts observed scores and decisions based upon them, this would seem an important area for further investigation.

6.2.2.2 Relative Advantages Regarding Consequences

The evidence considered in the current study was found to indicate that neither placement assessment alone, nor the overall PAS, appeared to result in beneficial consequences for individual stakeholders or the teaching and learning occurring in the Developmental English program. While new students widely report confidence in their ability to pass the English courses into which they have been placed by the PAS, most also report being in a course that is too easy for them (72% for RW courses, p=.00, 68% for LS, p=.00) and several respondents complained of “wasted” time and financial aid. Instructor and new student opinion converged in suggesting 15% of students are placed in levels too difficult for their abilities. Instructors reported that, of all of their first-semester students from Fall 2008 to Spring 2010, 32% and 43% were placed in the wrong RW and LS course, respectively. During the focus group interview and on questionnaire comments they also complained of ‘scrambles’ at the beginning of the semester to try to identify and reassign misplaced students, and of mixed abilities classes, which added to their frustrations and impeded the teaching and learning in their courses.

While instructors clearly viewed AC as the primary reason for the perceived PAS performance issues, with 91% suggesting it was not useful for informing placement decisions (p=.04) and 92% indicating the WS is (p=.00), the evidence suggests both instruments, at least as currently executed at CMI, are contributing to any troubles with the PAS.

Given the problems of potential sources of CIV, scoring, relevance, sufficiency and/or utility for the individual instruments, and resultant problems for the overall placement system, it is not surprising that the consequences claim for the PAS was not supported. These issues also make it difficult to compare the findings of the current study with those from the literature. In large, most would seem to argue the inclusion of writing samples leads to greater placement accuracy (Breland, Camp, Jones, Morris, & Rock, 1987; Galbraith, 1986; Garrow, 1989; Isonio, 1991, 1992; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998). However, recent evidence would seem to be rather scarce, with conflicting results as to the relative contributions of including WS outcomes in placement decisions (Matzen & Hoyt, 2004; May, 2007). As for the results from CMI, they would seem to offer a reminder that the use of any instrument, be it a standardised test or locally developed and marked performance assessment, is no guarantee of beneficial outcomes for any stakeholder without careful consideration of all aspects of its design, and monitoring of its performance in the local context, for local purposes.

6.2.3 RQ 3: What opportunities for improvement to the PAS has the validation study revealed?

This validation inquiry has highlighted numerous problems associated with both placement instruments, and with the resultant placement assessment system their outcomes inform. This section will present a number of the potential opportunities for CMI to improve the PAS, which could result in greater performance and more beneficial outcomes for all stakeholders and the institution as well.

6.2.3.1 Changes regarding the standardised placement instrument

While recent evidence may be lacking, most studies (Breland et al., 1987; Galbraith, 1986; Garrow, 1989; Isonio, 1994; Matzen & Hoyt, 2004; Wolcott, 1996; Wolcott & Legg, 1998), including the current investigation, suggest the combination of a standardised instrument and locally developed and marked WS offer better placement accuracy than either assessment method alone. As such, CMI would likely do well to maintain this combination of procedures. Given the problems with text comprehension, restricted range of results, and perceived and demonstrated lack of relevance of the instrument tasks to the instructional domain, however, the institution would almost certainly do well to consider transitioning from AC to an instrument designed to place ELLs within a multi-leveled English language program.

Due to the results of the current investigation, and continued concerns expressed by stakeholders regarding the use of AC, the institution did, indeed, adopt a new paper-and-pencil standardised test, Accuplacer ESL. Piloting of the instrument suggested its results accounted for significant, and in some instances quite substantial, variance in students' final outcomes in five of the six Developmental English courses (Table 6.1). For every course, the instrument was a better predictor than either AC or the WS.

Table 6.1: Predictive capacity of Accuplacer ESL for CMI Developmental English courses

Course	Level	n	r^2	Sig. (2-tailed)
Reading and Writing	1	82	.202	.000
	2	15	.227	.072
	3	11	.771	.000
Listening and Speaking	1	44	.314	.000
	2	28	.265	.005
	3	12	.405	.026

Additionally, normal distribution of scores, and opinions solicited from CMI

English faculty indicating the instrument tasks are relevant to the requirements of the instructional domain (79% of all respondents agreed or strongly agreed, n=19, p=.01), suggest the instrument may avoid the issues of restricted range of results and lack of relevance found for the Companion version.

6.2.3.2 Changes regarding the writing sample

The validation study identified a number of areas in which improvements could be made for the functioning and usefulness of the WS. The higher than expected proportion of instructors and examinees who expressed concerns or reported problems, respectively, with regard to the accessibility of the instrument texts suggests that efforts to avoid such problems may need to be revisited. While the establishment of a group of writing instructors experienced in teaching and assessing Marshallese ELLs, and the employment of an outside advisor to review the prompts and instructions they develop, are considered to be positive aspects of the WS text creation process, perhaps all prompts should go through the additional process of being piloted with current CMI students and feedback gained from post-assessment questionnaires or focus group interviews in order to identify problematic instructions, topics, or content in the text.

Possibly the most significant problem that needs to be dealt with regarding the WS is the impact of inter-rater discrepancies and the resulting CIV demonstrated to impact placement recommendations, and also demonstrated, via MFRM analysis, to be mitigatable. Results from a number of investigations addressing rater consistency suggest steps such as rater training, norming sessions, and a common rubric can be effective in assuring inter-rater differences are not validity threatening (East, 2009; James & Templeman, 2009; Jonsson & Svingby, 2007; Matzen & Hoyt, 2004). While CMI does use a common rubric, the institution does not regularly engage in any form of rater training or norming, and has not done so since the inception of the current rubric. As most current faculty/raters have joined the college since this time, few raters will have participated in any form of training.

Instating such procedures, then, may go a long way in improving inter-rater reliability. Another effective means of addressing the problem would be to utilize Rasch analysis, as MFRM has been demonstrated in a number of contexts to be an effective means of not only assessing the extent of, but dampening the impact of, inter-rater discrepancies (McNamara, 1996; Park, 2004).

The institution may also wish to review the current scoring rubric fort the WS. While results of the MFRM suggest the criteria, and the way the criteria are understood and applied by the raters, would appear to be operating well in assessing a singular construct (presumably, writing ability), functioning of the rating scale was found to be problematic. As this could be due to inconsistency in understanding and/or application of the scoring categories across raters, the institution should make sure to include discussions and possible revisions of the rating scale in any implemented rater training sessions.

One possible revision for the rating scale might be to align it with the five potential placement categories – exempted, excluded, or placed within Level 1, 2, or 3 of the Developmental English program -- into which its outcomes are used to place students. This would allow for the use of the relevant, specific student learning outcomes (SLOs) from RW courses from each level to be used as descriptors for evaluating placement essays. As all raters are English instructors, familiar with the SLOs of the program, this could help make comprehension and application of the rating scale more uniform and results of the instrument, resultantly, more meaningful and useful.

6.2.3.3 Changes regarding the overall placement assessment system

In addition to the insights for potential improvements relating to the specific placement instruments, a number relevant to the overall PAS were also uncovered. First amongst these, for both ethical and functional reasons, may be to ensure that all test-takers, prior to completing any instruments, understand the purposes, potential outcomes, and processes involved in the PAS. Indications that examinees might not know that the WS score is equally weighted for placement decisions as that of the standardised instrument, for example, may be resulting in some applicants not dedicating the time and attention to the instrument that they would otherwise.

Indications from stakeholders, particularly examinees themselves, that time constraints for both placement assessments are an issue also warrants investigation. CMI policy states that there are to be no time limitations on placement assessments. Should test conditions not mirror this policy, or should test-takers perceive a time constraint that does not in fact exist, this is something that must be addressed, as it is important that all examinees be able to demonstrate their skills to the best of their ability, without factors such as time constraints, or the risk of discrepancies in the speed with which candidates are able to perform introducing CIV.

Consideration of additional sources of evidence to bolster the sufficiency and likely predictive capacity and positive consequences of the PAS should also be a priority. One obvious area of concern should be the lack of assessment of oral and aural language skills, considering the PAS places students not only in RW, but LS courses as well. Certainly, placing students based on skills associated with the written language alone is not uncommon; indeed, it is the norm in 2-year higher education language programs in the US (Hughes & Scott-Clayton, 2011; Sullivan, 2008). It is also particularly problematic in a context like the Marshall Islands, where access to electricity or limited test-taker experience with computers are issues that constrain assessment method options. However, the institution needs to address this rather glaring gap between the skills addressed by their PAS and those required in the instructional domain.

Additionally, the apparent loss of the writing sample results for some 600 applicants to the college, and the lack of attempts to keep a record of which writing sample prompt candidates responded to highlight significant problems in the attention to meaningful data collection and simple record-keeping at the institution. While changes in leadership and focus at the institution lead to improvements in both areas, perhaps as reflected in the positive outcomes such as full re-accreditation by the ACCJC and WASC, and support for endeavours such as the current placement assessment validation study, efforts to improve in these areas must be ongoing.

Finally, and arguably most importantly, the institution needs to continue the reiterative investigation of the validity and usefulness of the PAS and its constituent instruments. It is unlikely that much of the information uncovered by this initial iteration of the validation study, and the opportunities for improvement as a result, would have been revealed otherwise. Future investigations can continue to inform further improvements and eventually result in a placement assessment system that results in beneficial consequences for all stakeholders.

6.3 Implications for future research

Outcomes of the validation study indicate a number of potentially fruitful areas of further study. With regard to the standardised placement tests (or standardised tests used for placement), results serve to confirm those of several others in the literature of the problematic nature of using instruments designed for native speakers with language learners, including problems with comprehension of the instrument texts, the possibility of making decisions based on dangerously restricted ranges of results, and the likely issue of lack of relevance of such instruments with the requirements of an instructional domain designed for language learners and language learning.

With regard to performance assessments, such as WSs, the large portion of examinees reporting difficulties understanding WS texts was considered somewhat surprising, particularly given that the instructions, prompts, topics, etc. were developed by writing instructors with experience teaching and assessing Marshallese, ELL, higher education students, and that the prompts were reviewed by an external consultant considered an assessment authority. This would not seem to be an area of attention in many studies in the literature, and would seem well worth further investigation given the substantial CIV which could be introduced based on differential comprehension of the task and content across examinees.

In addition to the current study, only two other investigations (Matzen & Hoyt, 2004; May, 2007) into the predictive capacity of WSs for English language courses in a higher education context were found to have been conducted in the past decade. Only one of these three studies has reported significant predictive capacity for the assessment method. Recent evidence backing up the rational argument of greater relevance to the instructional domain would appear to be lacking, and something that would be highly valuable in the ongoing discussions of the value of WS's and other performance assessments for placement purposes.

While scoring rubrics are not an uncommon area of consideration in studies investigating WSs or other performance assessments, the rating scales are not as frequently scrutinized, especially if the study does not employ Rasch analysis. As the functioning of this facet potentially impacts observed scores, as demonstrated in the current study, and impacts decisions made based upon those scores, this is another area deserving of further attention in the literature.

Results of the current study offer a good example of some of the considerable insights offered via Rasch analysis. Use of the procedure made it possible to determine what facets of the WS procedure were functioning reasonably well – internal rater consistency and scoring rubric criteria – and which were not – inter-rater variance and the rating scale. As a result, CMI not only knows of the issues causing the WS to be less than effective for placement, but knows where to focus efforts to address these problems. The advantages of Rasch analysis may not be something new to the literature, indeed it has been reported and widely recognized for some time (Linacre, 1999; McNamara, 1996; Park, 2004), but this has not translated into widespread use, or even awareness, of the approach, even where performance assessments are used to inform high-stakes decisions about test-takers and programs. This is troubling not only because of the missed benefits and continued problematic functioning and decisions likely occurring because of it, but also because the financial and human resources required to acquire and implement the necessary software is a minor investment relative to the potential benefits for stakeholders.

Finally, the articulation and implementation of an argument-based validation framework, one which addresses both test score interpretation and utilization (Bachman, 2005; Bachman & Palmer, 2010), and which incorporates a logical mechanism such as Toulmin's model, has resulted in considerable insights regarding the functioning and usefulness of the PAS and its constituent placement instruments at CMI. Equally important, it has not only revealed problems, but provided a path towards addressing those problems and improving the performance of the placement system and the consequences for all stakeholders at the college. As with the dearth of MFRM utilization amongst test-using institutions, so too are there far too few programs benefiting from the implementation of similar validation studies, and then benefiting other institutions by sharing these findings through publication in the literature or reporting them elsewhere. For the field of validation and the practice of ethical and effective assessment to make substantial strides forward, this must be made a priority amongst all test-using institutions.

6.4 Conclusion

Use of a single, standardised instrument to make high-stakes decisions about test-takers is pervasive in higher education. Contrary to longstanding best practices encouraged by researchers, professional organizations, test publishers, and many accrediting bodies, few, if any such institutions have endeavoured to meaningfully validate the instrument(s) they use for their specific context and purposes. The current study attempted to address this void by developing and applying an argument-based validation framework for two widely adopted placement assessment methods – a standardised placement test, Accuplacer

Companion, and a locally developed and marked writing sample – used by CMI to exempt, exclude, or place students within their Developmental English language program.

Results indicated not only a number of expected results, such as problems associated with using instruments designed for native speakers with language learners, but also a number of outcomes not expected by local stakeholders, such as the lack of relevance of WS results to English course outcomes. Results of the validation study are argued to have not only revealed substantial problems in the functioning and consequences of the individual placement instruments and the overall PAS at the institution, but to have provided insights to opportunities for improvement that would not likely have been revealed otherwise.

As such, the study is argued to be evidence of the substantial value of *in sitiu*, argument-based validation studies, particularly those attending to both test score interpretation and utilization, and which employ Toulmin's informal argument model, for any institution employing assessments to inform high-stakes decisions about individuals and/or programs.

REFERENCES

ACCJC (2003). *Accreditation Notes*. Accrediting Commission for Community and Junior Colleges. Retrieved from http://www.accjc.org/wp-content/uploads/2010/09/april03.pdf

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/ACCJC_WASC_Accreditation_Standards.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.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(02), 79-113.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*: Cambridge University Press.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*(3), 280-297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied linguistics, 14*(2), 115-115.

APA, AERA, & NCME (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, D.C.: The Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.) *Test Validity* (pp.19-32). Hillsdale, NJ: Lawrence Erlbaum.

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.

Aryadoust, V. (2011). Validity arguments of the speaking and listening modules of international English language testing system: A synthesis of existing research. *The Asian ESP Journal, 7*(2), 28-54.

Aryadoust, V. (in press). Evaluating the psychometric quality of an ESL placement test of writing: A many-facets Rasch study. *Linguistics Journal.*

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. F. (1996). Review of Seoul National University Criterion-Referenced English Proficiency Test (SNUCREPT). *Language Research, 32*(2), 373-383.

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. (2006). *Linking validity and use in educational assessment*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Francisco, CA.

Bachman, L. F. (2007). *Language assessment: Opportunities and challenges*. Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Costa Mesa, CA. Retrieved from http://202.197.121.116/Downloads/LangTst/tst_004.doc

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

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

Bachman, L. F., & Palmer, D. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

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

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

Bailey, T., Jeong, D. W., & Cho, S. W. (2010b). *Student Progression through Developmental Sequences in Community Colleges* (No. 45). New York, NY: Teachers College, Columbia University. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED512395

Behrman, E. H. (2000). Developmental Placement Decisions: Content-Specific Reading Assessment. *Journal of Developmental Education, 23*(3), 12-18.

Bejar, I. I., & Jamieson, N. (2000). *TOEFL 2000 listening framework: A working paper.* Educational Testing Service.

Belcher, M. J. (1993). *Preparedness of High School Graduates for College: A Statewide Look at Basic Skills Tests Results for 1990-91.* Miami-Dade Community College Office of Institutional Research. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED366394

Board of Governors of the California Community Colleges. (2008). *Report on the System's Current Programs in English as a Second Language (ESL) and Basic Skills*. Sacramento, CA: Board of Governors of the California Community Colleges.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* College Board Report No. 0-87447-280-6. New York: College Entrance Examination Board.

Bright, P., & Chutaro, E. (2007). *Marshall Islands Population Atlas*. Noumea, New Caledonia: Secretariat of the Pacific Community.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*(2), 105-121.

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. McGraw-Hill.

Brown, R. S., & Niemi, D. N. (2007). *Investigating the Alignment of High School and Community College Assessments In California* (No. 07-3). The National Center for Public Policy and Higher Education. Retrieved from http://www.highereducation.org/reports/brown_niemi/index.shtml

Burt, M., & Keenan, F. (1995). *Adult Learner Assessment: Purposes and Tools*. National Clearinghouse for ESL Literacy Education.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper*. Educational Testing Service.

Cabrillo College Office of Institutional Research (1999). *Content, consequential and cut-score validity study of ACCUPLACER Reading Test.* Cabrillo College. Retrieved from http://pro.cabrillo.edu/pro/pro_reports/readassess.pdf

Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice, 25*(3), 35-41.

Cambridge ESOL. (2007). IELTS Handbook. Cambridge: Cambridge ESOL. Retrieved from http://www.cambridgeesol.org/assets/pdf/resources/IELTS_Handbook.pdf

Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8*, 67-84.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Rep. No. 19). Princeton, NJ: Educational Testing Service.

CCCAA (2007a). *California Community Colleges Assessment Association Test-Development Feasibility Project.* Sacramento, CA: California Community Colleges Assessment Association.

CCCAA (2007b). *Special Meeting to Discuss the Feasibility of a CCC-Owned Assessment Instrument*. Sacramento, CA: California Community College Assessment Association.

CCCAA (2008). *Standards, policies and procedures for the evaluation of assessment instruments used in California community colleges.* Sacramento, CA: California Community Colleges Assessment Association.

CCCCO (2009). *Focus on Results: Accountability Reporting for the California Community Colleges*. Sacramento, CA: California Community Colleges Chancellor's Office.

Cellan, J. (2007). International Pharmacy Graduate Language Assessment. *The Canadian Modern Language Review, 64*(1), 226-233.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

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.

Clapham, C. (2000). Assessment and Testing. *Annual Review of Applied Linguistics, 20*, 147-161.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones & B. Spolsky (Eds.), *Language Testing Proficiency* (pp. 10-24). Arlington, VA: Center for Applied Linguistics.

CMI (2008). *College of the Marshall Islands Faculty Orientation Handbook*. Majuro, Republic of the Marshall Islands: College of the Marshall Islands.

CMI (2009). *A Report of the Self Study: Institutional Self Study Report in Support of Reaffirmation of Accreditation*. Majuro, Republic of the Marshall Islands: College of the Marshall Islands.

CMI (2010). *Developmental English Course Outlines*. Majuro, Republic of the Marshall Islands: College of the Marshall Islands.

Cohen, A. M., & Brawer, F. B. (1987). *The Collegiate Function of Community Colleges.* San Francisco, CA: Jossey Bass.

College Board. (2003). *Accuplacer OnLine Technical Manual.* College Board. Retrieved from http://isp.southtexascollege.edu/ras/research/pdf/ACCUPLACER_OnLine_Technical_Manual.pdf

College of the Canyons (1993). *College of the Canyons Predictive Validity Studies.* Valencia, CA: College of the Canyons.

College of the Canyons (1996). *English Writing Placement Recommendations at College of the Canyons: An Analysis of Disproportionate Impact*. Santa Clarita, CA: College of the Canyons.

Crawford, J. (2004). *No Child Left Behind: Misguided approach to school accountability for English language learners*. Center on Educational Policy's Forum on Ideas to Improve the NCLB Accountability Provisions for Students with Disabilities and English Language Learners. Retreived from http://users.rcn.com/crawj/langpol/Crawford_NCLB_Misguided_Approach_for_ELLs.pdf

Crisp, V., & Shaw, S. (2010). *How hard can it be? Issues and challenges in the development of a validation method for traditional written examinations*. Paper presented at the International Association for Educational Assessment Annual Conference, Bangkok, Thailand. Retreived from http://www.iaea2010.com/fullpaper/501.pdf

Cronbach, L. J. (1969). Validation of educational measures. In *Proceedings of the 1969 Invitational Conference on Testing Problems* (pp. 35-52). Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco, CA: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3-18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, Theory, and Public Policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. Retrieved from http://psychclassics.yorku.ca/Cronbach/construct.htm.

Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly*, *1*(1), 5–18.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 Writing Framework: A Working Paper.* Princeton, NJ: Educational Testing Service. Retrieved from http://www1.ets.org/Media/Research/pdf/RM-00-05.pdf

Denham, P. A., & Oner, J. A. (1992). *IELTS research project: Validation study of listening sub-test*. Canberra, Australia: University of Canberra.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88-115.

Engemann, J. F., & Gallagher, T. (2006). The Conundrum of Classroom Writing Assessment. *Brock Education Journal, 15*(2), 35-44.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing, 14*(2), 113-139.

Gabe, L. A. C. (1989). *Relating College-Level Course Performance to ASSET Placement Scores*. Retrieved from ERIC database. (ED451856)

Galbraith, F. L. (1986). *The use of multiple choice items and holistically scored writing samples to assess student writing ability* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 8626947)

Garrow, J. R. (1989). *Assessing and improving the adequacy of college composition placement* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 8921432)

Goh, C., & Aryadoust, S. V. (2010). Investigating the Construct Validity of the MELAB Listening Test through the Rasch Analysis and Correlated Uniqueness Modeling. *SPAAN Fellow Working Papers in Second or Foreign Language Assessment, 8*, 31-68.

Goodman, J. F., Freed, B., & McManus, W. J. (1990). Determining exemptions from foreign language requirements: Use of the modern language aptitude test. *Contemporary Educational Psychology, 15*(2), 131-141.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*(3), 505-521.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*(1), 7-22.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haiyang, S. (2010). An application of classical test theory and many-facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87-102.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second Language Writing* (pp. 69-87). Cambridge: CUP.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*(1), 1-9.

Haswell, R. (2005). Post-secondary Entrance Writing Placement. Corpus Christi, TX: Texas A&M University. Retrieved from http://comppile.org/profresources/placement.htm

Hebel, S. (2001). Universities push to influence state tests for high-school students. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Universities-Push-to-Influence/1724

Heilenman, L. K. (1983). The use of a cloze procedure in foreign language placement. *Modern Language Journal, 67*(2), 121-126.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MA: Newberry House Publishers.

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports*, *2*, 52–63.

Hillocks Jr, G. (2002). *The testing trap: how state writing assessments control learning*. New York, NY: Teachers College Press.

Hillocks Jr, G. (2003). Fighting back: Assessing the assessments. *English Journal, 92*(4), 63-70.

Holderer, R. W. (1992). *The use of the Modified Primary Trait Scoring Guide for placing students in freshman composition* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 9321577)

Hughes, A. A. (1989). *Testing for language teachers*: Cambridge University Press.

Hughes, A. A. (2003). *Testing for language teachers*: Cambridge University Press.

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. Retrieved from http://www.tc.columbia.edu/centers/ncpr/conference/PDF/NCPR_Panel%202_HughesClaytonPaper.pdf

Hughes, R. E., & Nelson, C. H. (1991). Placement Scores and Placement Practices: An Empirical Analysis. *Community College Review, 19*(1), 42-46.

Isonio, S. (1991). *Implementation and Initial Validation of the APS English Test and The APS English-Writing Test at Golden West College: Evidence for Predictive Validity*. Retrieved from Eric database. (ED345781)

Isonio, S. (1992). *English Placement Recommendations at Golden West College: An Analysis of Disproportionate Impact*. Retrieved from Eric database. (ED346908)

Isonio, S. (1994). *Relationship between APS Writing Test Scores and Instructor Preparedness Ratings: Further Evidence for Validity*. Retrieved from Eric database. (ED370617)

Jackson, T. R., Draugalis, J., Slack, M. K., Zachry, W. M., & D'Agostino, J. (2002). Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education*, *66*(3), 233–242.

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167-178.

James, C. L., & Templeman, E. (2009). A case for faculty involvement in EAP placement testing. *TESL Canada Journal, 26*(2), 82-99.

Joint Committee on Testing Practices (2005). Code of fair testing practices in education (Revised). *Educational Measurement: Issues and Practice*, *24*(1), 23-26.

Jones, J., & Jackson, R. (1991). *The Impact of Writing Placement Testing and Remedial Writing Programs on Student Ethnic Populations at Oxnard College*. Retrieved from Eric database. (ED335081)

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 135.

Kane, M. T. (2006). Validation. *Educational Measurement, 4*, 17-64.

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

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA: Online Journal of Language Studies*, *11*(3), 179–197.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. In R. Tulloch (Ed.), *International English Language Testing System Research Reports* (Vol. 3, pp. 85-108). Canberra, Australia: IELTS Australia.

Khairani, A. Z., & Nordin, M. S. (2011). The development and construct validation of the mathematics proficiency test for 14-year-old students. *Asia Pacific Journal of Educators and Education*, *26*(1), 33–50.

Kim, J. Y. (2008). Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3337823)

Klee, C. A., & Rogers, E. S. (1989). Status of articulation: placement, advanced placement credit, and course options. *Hispania, 72*(3), 763-773.

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing, 2*(1), 76-104.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing, 1*(2), 134-134.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing, 19*, 1-29.

Kunnan, A. J. (1998). Approaches to validation in language assessment. In Kunnan, A. J. (Ed.) *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium* (pp. 1-18). Lawrence Erlbaum.

Kunnan, A. J. (2000). Fairness and justice for all. In Kunnan, A. J. (Ed.) *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.

Kunnan, A. J. (2003). Test fairness. In M. Milanovic & C. Weir (Eds.), *Select Papers from the European Year of Languages Conference, Barcelona*. Cambridge: Cambridge University Press.

Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*(4), 366-389.

Linacre, J. M. (1989). *Many-facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category unity. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, J. M. (2010). *A user's guide to FACETS Rasch-model computer programs*. Retrieved from http://www.winsteps.com.

Linacre, J. M. (2011). *FACETS* Rasch measurement [*computer program*], Version 3.68.1. Chicago: Winsteps.com.

Linacre, J. M., Engelhard Jr., G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educaitonal Research, 21*(6), 569-577.

Llosa, L. (2008). Building and supporting a validity argument for a standards based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27*(3), 32-42.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351-351.

Lytle, S. L., & Wolfe, M. (1989). *Adult Literacy Education: Program Evaluation and Learner Assessment.* Retrieved from Eric database. (ED315665)

Mann, W., & Marshall, C. R. (2010). Building an Assessment Use Argument for sign language: the BSL - Nonsense Sign Repetition Test. *International Journal of Bilingual Education and Bilingualism, 13*(2), 243-243.

Marion, S. F., & Pellegrino, J. (2009). *Validity Framework for Evaluating the Technical Quality of Alternate Assessments Based on Alternate Achievement Standards*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Francisco, CA.

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.

Marwick, J. D. (2004). Charting a path to success: The association between institutional placement policies and the academic success of latino students. *Community College Journal of Research & Practice*, *28*(3), 263-280.

Mathay, G. A. (1992). *Learning Outcomes Assessment Activities, 1989-1992*. Retrieved from Eric database. (ED354027)

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://professionals.collegeboard.com/profdownload/pdf/09b_765_PredValidity_WEB_091124.pdf

Matzen, R. N., & Hoyt, J. E. (2004). Basic writing placement with holistically scored essays: Research evidence. *Journal of Developmental Education*, *26*(1), 2-34.

May, J. S. (2007). Analyzing the Placement of Community College Students in English as a Second Language for Academic Purposes (EAP) Courses (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3281566)

McLeod, S., Horn, H., & Haswell, R. H. (2005). Accelerated classes and the writers at the bottom: A local assessment story. *College Composition and Communication*, *56*(4), 556–580.

McNamara, T. F. (1996). *Measuring second language performance*: Longman.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick. *Language Assessment Quarterly, 3*(1), 21-21.

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

Messick, S. J. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256.

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analyses. In Cumming, A. & Berwick, R. (Eds.), *Validation in Language Testing* (pp. 15-38). Clevedon: Multilingual Matters.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470-476.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: validity in writing assessment. In C. Bazerman (Ed.), *Handbook of Research on Writing: History, Society, School, Individual, Text* (pp. 365-385) New York, NY: Lawrence Erlbaum Associates.

Nakamura, Y. (2004). *A Comparison of Holistic and Analytic Scoring Methods in the Assessment of Writing.* Paper presented at the JALT Pan-SIG Conference, Tokyo, Japan. Retrieved from http://jalt.org/pansig/2004/HTML/Nakamura.htm

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.

Palmer, A., Groot, P. J. M., & Trosper, G. A. (1981). *The Construct Validation of Tests of Communicative Competence*. Paper presented at the Annual TESOL Conference, Washington, DC.

Pardo-Ballester, M. C. (2007). The development of a Web-based Spanish listening placement exam (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3278322)

Pardo-Ballester, M. C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly*, *7*(2), 137.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Columbia University Working Papers in TESOL & Applied Linguistics, 4*(1), 1-21.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72-92.

Raatz, U. (1985). Tests of reduced redundancy: The C-test, a practical example. In V. Kohonen & A. J. Pitkanen (Eds.), *Language Testing in School, AFinLA Yearbook 1985*. Tampere, Finland: AFinLA.

Rasor, R. A., & Barr, J. (1995). *Refinement in Assessment Validation: Technicalities of Dealing with Low Correlations and Instructor Grading Variation*. Retrieved from Eric database. (ED393883)

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Saunders, P. I. (2000). *Meeting the Needs of Entering Students through Appropriate Placement in Entry-Level Writing Courses.* Retrieved from Eric database. (ED447505)

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based Test (iBT): Exploration in a field trial sample* (TOEFL iBT Research Report No. TOEFLiBT-04). Retrieved from http://praxis.ets.org/Media/Research/pdf/RR-08-09.pdf

Sawyer, R. (1989). *Validating the use of ACT Assessment scores and high school grades for remedial course placement in college* (ACT Research Report No. 89-4). Retrieved from http://www.act.org/research/reports/pdf/ACT_RR89-4.pdf

Sawyer, R. (2007). Indicators of Usefulness of Test Scores. *Applied Measurement in Education, 20*(3), 255-271.

Schumaker, R. E., & Smith Jr., E. V. (2007). A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.

Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education*, *19*, 405-450.

Shohamy, E. (1993). *The Power of Tests: The Impact of Language Tests on Teaching and Learning. NFLC Occasional Papers.* Retrieved from Eric database. (ED362040)

Shohamy, E. G. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340-349.

Shohamy, E. G. (1998). Critical language testing and beyond. *Studies In Educational Evaluation, 24*(4), 331-345.

Shohamy, E. G. (2001). *The power of tests: a critical perspective on the uses of language tests*. Longman.

Shohamy, E. G., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298-317.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English Language Learners. *Review of Educational Research, 78*(2), 260-329.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing, 14*(3), 242-247.

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing* (Vol. 1, pp. 5–21). Frankfurt: P. D. Lang.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4).

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching, 42*(01), 1-40.

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.

Sullivan, P., & Nielsen, D. (2009). Is a Writing Sample necessary for “Accurate Placement”? *Journal of Developmental Education*, *33*(2), 2-11.

Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing*, *10*(2), 193 -207.

Toulmin, S. E. (1958). *The Uses of Argument* (1st ed.). Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The Uses of Argument* (3rd ed.). Cambridge: Cambridge University Press.

United Nations. (2004). *World Population Prospects: The 2004 Revision, Analytical Report*. United Nations. Retrieved from http://www.un.org/esa/population/publications/WPP2004/WPP2004_Volume3.htm

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321-344.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.

White, E. M. (1989). *Developing Successful College Writing Programs.* San Francisco, CA: Jossey-Bass

White, E. M. (1990). Language and reality in writing assessment. *College Composition and Communication*, *41*(2), 187-200.

Williams, K. L. (1990). Three new tests for overseas students entering postgraduate and vocational training courses. *ELT Journal, 44*(1), 55-65.

Wolcott, W. (1996). Evaluating a basic writing program. *Journal of Basic Writing, 15*(1), 57-69.

Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. Retrieved from Eric database. (ED423541)

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions. 16*(3), 888.

Wright, B. D., & Stone, M. H. (1988). *Reliability in Rasch Measurement* (Research Memorandum No. 53). Chicago: MESA.

Wrigley, H. S. (1992). *Learner Assessment in Adult ESL Literacy. ERIC Q & A*. Retrieved from Eric database (ED353863).

Xi, X. (2004). Investigating language performance on the graph description task in a semi-direct oral test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 2,* 83-134. Retrieved from http://www.ling.lsa.umich.edu.simsrad.net.ocs.mq.edu.au/UMICH/eli/Home/_Projects/Scholarships/Spaan/PDFs/Spaan_Papers_V2_2004.pdf#page=90

Xi, X. (2008). Methods of Test Validation. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (pp. 2316-2335). Boston, MA: Springer US. Retrieved from http://www.springerlink.com.simsrad.net.ocs.mq.edu.au/content/x26628716822v82r

Zinn, A. (1998). Ideas in practice: Assessing writing in the developmental classroom. *Journal of Developmental Education, 22*(2), 28-34.

APPENDIX A: Developmental English Course Student Learning Outcomes

General Outcomes for Developmental English Listening and Speaking Courses

	Level 1	Level 2	Level 3
1	Demonstrate beginning oral and aural skills in English communication tasks in various academic settings	Demonstrate intermediate oral and aural skills in English communication tasks in various academic settings	Demonstrate pre-college oral and aural skills in English communication tasks in various academic settings
2	Build upon and use an adequate beginning level of skill in listening	Attain an intermediate level of skill in listening	Attain a pre-college level of skill in listening
3	Build upon and use an adequate beginning level of skill in speaking	Attain an intermediate level of skill in speaking	Attain a pre-college level of skill in speaking
4	Demonstrate effective beginning academic note-taking skills	Demonstrate effective intermediate academic note-taking skills	Demonstrate effective pre-college academic note-taking skills

Student Learning Outcomes for Developmental Listening and Speaking Courses

	Level 1	Level 2	Level 3
1	Respond to beginning communicative exchanges, such as simple statements, questions, and commands in dialogues and role play situations	Respond to intermediate communicative exchanges, such as simple statements, questions, and commands in dialogues and role play situations	Attain appropriate organizational strategies to initiate, sustain, and close pre-college communicative exchanges in dialogues and role play situations
1a			Perform impromptu speaking activities with other students that draw on personal experience and knowledge
2	Listen effectively:		
2a	Discriminate among English phoneme stops (/p/, /b/, /t/, /d/, /k/, and /g/)	Discriminate among the English fricatives (/f/, /v/, /sh/, /zh/, /s/, /z/, /Θ/ and /δ/) and distinguish between fricatives and stops	Discriminate among the English affricates (/ch/, /dzh/) and discriminate between stops, fricatives and affricates
2b	Apply new vocabulary to basic conversations	Apply new vocabulary to intermediate conversations	Apply new vocabulary to advanced conversations
2c	Respond correctly to impromptu questions	Respond correctly to impromptu questions	Respond correctly to impromptu questions
3	Speak effectively:		
3a	Acquire correct pronunciation of English phoneme stops (/p/, /b/, /t/, /d/, /k/, and /g/)	Acquire correct pronunciation of the English fricatives (/f/, /v/, /sh/, /zh/, /s/, /z/, /Θ/ and /δ/)	Acquire correct pronunciation of English affricates (/ch/, /dzh/)
3b	Manipulate new vocabulary in short interchanges	Manipulate new vocabulary in short interchanges	Manipulate new vocabulary in longer interchanges
3c	Model good speaking skills in short presentations	Employ narrative and descriptive strategies in presentations and dialogues	Formulate and present opinions with supporting details and hypotheses
4	Acquire effective note-taking skills:		
4a	Use note-taking skills to write the main ideas from a variety of lecture topics	Detect main ideas and supporting details from lectures	Detect implied main ideas from a variety of lecture topics
4b		Write the main ideas and supporting details "in their own words" from lectures	Paraphrase the main ideas and supporting details from lectures
4c			Analyze various organizational strategies in lectures

General Outcomes for Developmental English Reading and Writing Courses

	Level 1	Level 2	Level 3
1	Demonstrate a pre-intermediate level of skill in reading	Demonstrate an intermediate level of skill in reading	Demonstrate a pre-college level of skill in reading
2	Read for pleasure to foster life-long learning	Read for pleasure to foster life-long learning	Read for pleasure to foster life-long learning
3	Demonstrate a pre-intermediate level of skill in writing	Demonstrate an intermediate level of skill in writing	Demonstrate a pre-college level of skill in writing
4	Demonstrate a pre-intermediate level of skill in grammar and mechanics	Employ an intermediate level of skill in grammar and mechanics	Demonstrate a pre-college level of skill in grammar and mechanics
5	Use appropriate word processing skills for written assignments	Use appropriate information technology for research and writing	Use appropriate information technology for research and writing

Student Learning Outcomes for Developmental English Reading and Writing Courses

	Level 1	Level 2	Level 3
1	Demonstrate active reading and referencing skills	Demonstrate active reading and referencing skills using intermediate level text materials	Read and record notes from academic textbook chapters and materials
1a	Locate facts and information by skimming and scanning	Locate facts and information by skimming and scanning using intermediate level text material	Locate facts and information by skimming and scanning
1b	Make logical inferences and predictions	Make logical inferences and predictions for events and information	Make logical inferences and predictions for events and information
1c	Make connections to texts	Make text-to-self and test-to-world connections from independent readings	Make text-to-self and test-to-world connections from independent readings
1d	Respond to readings by expressing personal opinions and ideas about issues in texts	Respond to readings with personal opinions	Respond to readings with personal opinions
1e	Self-select independent reading material appropriate to their level	Identify signal words to determine rhetorical forms and structure	Identify signal words to determine rhetorical forms and structure to improve comprehension of the text
1f	Identify signal words to determine rhetorical forms and structure	Use a dictionary to determine meaning, word forms, and pronunciation of words	Use a dictionary to determine meaning, word forms, and pronunciation of words
1g	Use a dictionary to locate meaning, word forms, and pronunciation of words	Determine meaning of new words by using contextual clues	Use context clues and linguistic clues to determine the meaning of new words
1h	Utilize note-taking skills to record essential information	Utilize note-taking skills to record essential information	Utilize note-taking skills to record essential information from textbook chapters
2	Select, read, and discuss a reading solely for pleasure	Select, read, and discuss a reading solely for pleasure	Select, read, and discuss a reading solely for pleasure
3	Use the writing process to produce a variety of styles of papers	Use the writing process to produce a variety of styles of papers	Write 3 to 5 paragraph essays with emphasis on closed form essay structure and direct thesis statement
3a	Write a narrative and a comparison/contrast paragraph with appropriate transitional devices	Write focused narrative, descriptive, definition, comparison/contrast, and opinion paragraphs	Use an advanced academic vocabulary in written work
3b	Write single-paragraph summaries	Write simple summaries from reading materials	
3c	Use academic vocabulary in written work	Use an academic vocabulary in written work	Use an academic vocabulary in written work

Student Learning Outcomes for Developmental English Reading and Writing Courses (cont.)

	Level 1	Level 2	Level 3
4	Demonstrate appropriate grammar and mechanics	Demonstrate appropriate grammar and mechanics	Demonstrate accurate grammar and mechanics
4a	Write sentences with subject-verb agreement	Write sentences with subject-verb agreement	Write sentences free of subject-verb agreement errors
4b	Show control of simple present, present continuous, simple past, present perfect, and future “will + going to” tenses	Use articles and plurals	Use count and non-count nouns correctly
4c	Use articles and plurals	Show control of verb tenses with a focus on the perfect and perfect continuous tenses	Show mastery of past, present, future, continuous constructions and present perfect tenses
4d	Use simple and compound sentences	Use conditionals in writing	Use modals in writing
4e	Identify and use appropriate labels for parts of speech and sentence formations	Use simple, compound and complex sentences in writing	Use simple, compound, complex, and compound-complex sentences in writing
4f	Use appropriate capitalization and end punctuation	Use appropriate capitalization, end pronunciation, and commas	Use commas, semicolons, and quotation marks
4g		Identify and use appropriate labels for English parts of speech and sentence formations	Employ adverbial, adjectival, and prepositional phrases
4h		Construct adverbial, adjectival, and prepositional phrases	
5	Use appropriate formats for submitting types assignments	Use word processing and Internet skills	Use word processing and Internet skills
5a		Use appropriate formats for submitting typed academic writing	Use appropriate formats for submitting typed academic writing
5b		Electronically submit writing assignments	Utilize academic sites for research
5c		Use academic sites provided by the instructor for research	

APPENDIX B: Analytic Scoring Rubric for Writing Sample

Criteria	Descriptor	Score	
Support	Addresses the topic, but position is very unclear; provides minimal or no support	1	
	Position is very underdeveloped	2	
	Lacks development or is repetitive in parts; position provides uneven support	3	
	Position provides some development	4	
	Takes a clear position and supports with reason and/or examples	5	
	Clear position, well-chosen reasons and/or examples, uses persuasive strategy	6	
Organization	Disorganized or unfocused in much of the response	1	
	Organized in parts or responses	2	
	Parts are disjointed and/or lack transitions	3	
	Generally organized, but has few or no transitions	4	
	Well-organized, lacks some transitions	5	
	Focused and well-organized with effective transitions	6	
Sentence Variation	No control over sentence boundaries and sentence structure	1	
	Minimal control over sentence boundaries and sentence structure	2	
	Uneven control over sentence boundaries and sentence structure	3	
	Sentence structure simple and unvaried	4	
	Some variety in sentence structure	5	
	Consistent variety in sentence structure	6	
Diction	Word choice inaccurate in much or all of the response	1	
	Word choice often inaccurate	2	
	Some inaccurate word choices	3	
	Word choice is mostly accurate	4	
	Words occasionally used inaccurately	5	
	Consistently precise in word choice	6	
Errors in Grammar	Severely impede understanding across the response	1	
	Interfere with understanding on much of the response	2	
	Frequently interfere with understanding	3	
	Sometimes interfere with understanding	4	
	Are present, but do not interfere with understanding	5	
	Are few and do not interfere with understanding	6	

APPENDIX C: English Instructor Focus Group Interview Consent Form

COLLEGE OF THE MARSHALL ISLANDS
JITDAM KAPEEL
1993

Research Project: Assessing the Assessments: Focus Group Interview

Research Aims:

This research is being conducted for two reasons. First, CMI wants to know how well our current placement tests are helping us in selecting new students and placing them in their first semester classes at the college. Your opinions and experiences with the Accuplacer Companion test, the Writing Sample, and placement system overall will help us with this. Second, this research is being conducted by Robert Johnson, an instructor in the Developmental English Department at CMI (tel. 625-3394, Ext 249/251, rjohnson@cmi.edu) as partial fulfilment of the requirements of the Doctor of Applied Linguistics degree, under the supervision of Dr. Mehdi Riazi (tel +612 9850-7951, Mehdi.Riazi@ling.mq.edu.au) and Dr. Stephen Moore (tel. +612 9850-8742, Stephen.Moore@ling.mq.edu.au) of the Applied Linguistics Department at Macquarie University.

Participants' Role and Rights:

As a participant in this part of the study, you will be asked a series of questions. It is entirely up to you whether you wish to answer and how you wish to answer each question. The focus group will last approximately 20-30 minutes.

All information and results will be kept private and confidential (they will not be made public) and all information collected will be used for research purposes only. Access to this information will be strictly limited to the principal researcher (named below) and the Institutional Research and Planning Department at CMI. Information gathered may be used for publications in the future, but no participants will ever be identified by name. Your privacy will be protected.

You may also request a summary of the results of this research project be sent to you via email.

I would like a summary of the results of this study sent to my email address. Yes □ No □

email: ________________________

Alternatively, please contact Robert Johnson (625-3394, Ext 249/251, rjohnson@cmi.edu) if you would like a summary of the results of this research project. (The summary will not contain any private information, but the overall findings of the study.)

Participating in this research is completely your decision and is not a requirement. There is no penalty for choosing not to participate and you may stop at any time, without consequence.

Principal Researcher's signature:

____________________________________ Date: ________________________________

Robert Johnson, Instructor
Developmental English Department, CMI
Email: rjohnson@cmi.edu
Tel: 625-3394 Ext 249 or 251

Participants' Signature:

I (the participant) have read (or, where appropriate, have had read to me) and understand the information above, and any questions I have asked have been answered to my satisfaction. I agree to participate in this research, knowing that I can withdraw at any time, without consequence. I have been given a copy of this form to keep.

Participant's Signature: ______________________________________Date: ________________

Participant's Name: ____
(please print)

Note: This research has been reviewed and approved by the campus community of the College of the Marshall Islands. If you have any complaints or concerns about this research, please contact the principal researcher (above), Ellia Zebedy, Vice President for Research, Planning and Grants at CMI (email: ezebedy@cmi.edu or telephone: 625-3394) or Don Hess, Interim Vice President of Academic Affairs and Student Life and Development (email: dhess@cmi.edu or telephone: 625-3394).

Additionally, the ethical aspects of this study have been approved by the Macquarie University Ethics Review Committee (Human Research). If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Ethics Review Committee through the Director, Research Ethics (telephone +60 9850 7854; email ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

APPENDIX D: First-year Student Placement Assessment and Course Outcome Consent Form

Research Project: Assessing the Assessments: Students' Results

Research Aims:

This research is being conducted for two reasons. First, CMI wants to know how well our current placement tests are helping us in selecting new students and placing them in their first semester classes at the college. Your course and test results will help us with this. Second, this research is being conducted by Robert Johnson, an instructor in the Developmental English Department at CMI (tel. 625-3394, Ext 249/251, rjohnson@cmi.edu) as partial fulfilment of the requirements of the Doctor of Applied Linguistics degree, under the supervision of Dr. Mehdi Riazi (tel +612 9850-7951, Mehdi.Riazi@ling.mq.edu.au) and Dr. Stephen Moore (tel. +612 9850-8742, Stephen.Moore@ling.mq.edu.au) of the Applied Linguistics Department at Macquarie University.

Participants' Role and Rights:

As a participant in this part of the study, you will not need to do anything. If you agree, your results for your class and scores on the tests below will be used by the researcher, Robert Johnson, and the Institutional Research and Planning Department at CMI, and will help us decide how well our placement procedures are working.

All information and results will be kept private and confidential (they will be kept secret) and all information will be used for research purposes only. Access to this information will be strictly limited to the principal researcher, Robert Johnson, and the Institutional Research and Planning Department at CMI. Information gathered may be used for publications in the future, but no participants will ever be identified by name. Your privacy will be protected.

Yes □ No □ I agree to the use of my Accuplacer: Companion results in this study.
(Accuplacer: Companion is the test you completed when you applied to CMI.)

Yes □ No □ I agree to the use of my Writing Sample results in this study.
(The Writing Sample is the essay you wrote when you applied to CMI.)

Yes □ No □ I agree to the use of my C-test results in this study.
(The C-test is the test you completed the first week of your English class.)

Yes □ No □ I agree to the use of my class and/or assignment results in this study.

You may also request a summary of the results of this research project be sent to you via email.

Yes □ No □ I would like a summary of the results of this study sent to my email address.

email: ________________________

Alternatively, please contact Robert Johnson (625-3394, Ext 249/251, rjohnson@cmi.edu) if you would like a summary of the results of this research project. (The summary will not contain any private information, but the overall findings of the study.)

Participating in this research is completely your decision and is not a requirement. There is no penalty for choosing not to participate and you may stop at any time, without consequence.

Principal Researcher's signature:

__________________________________ Date: ______________________________

Robert Johnson, Instructor
Developmental English Department, CMI
Email: rjohnson@cmi.edu
Tel: 625-3394 Ext 249 or 251

Participants' Signature:

I (the participant) have read (or, where appropriate, have had read to me) and understand the information above, and any questions I have asked have been answered to my satisfaction. I agree to participate in this research, knowing that I can withdraw at any time, without consequence. I have been given a copy of this form to keep.

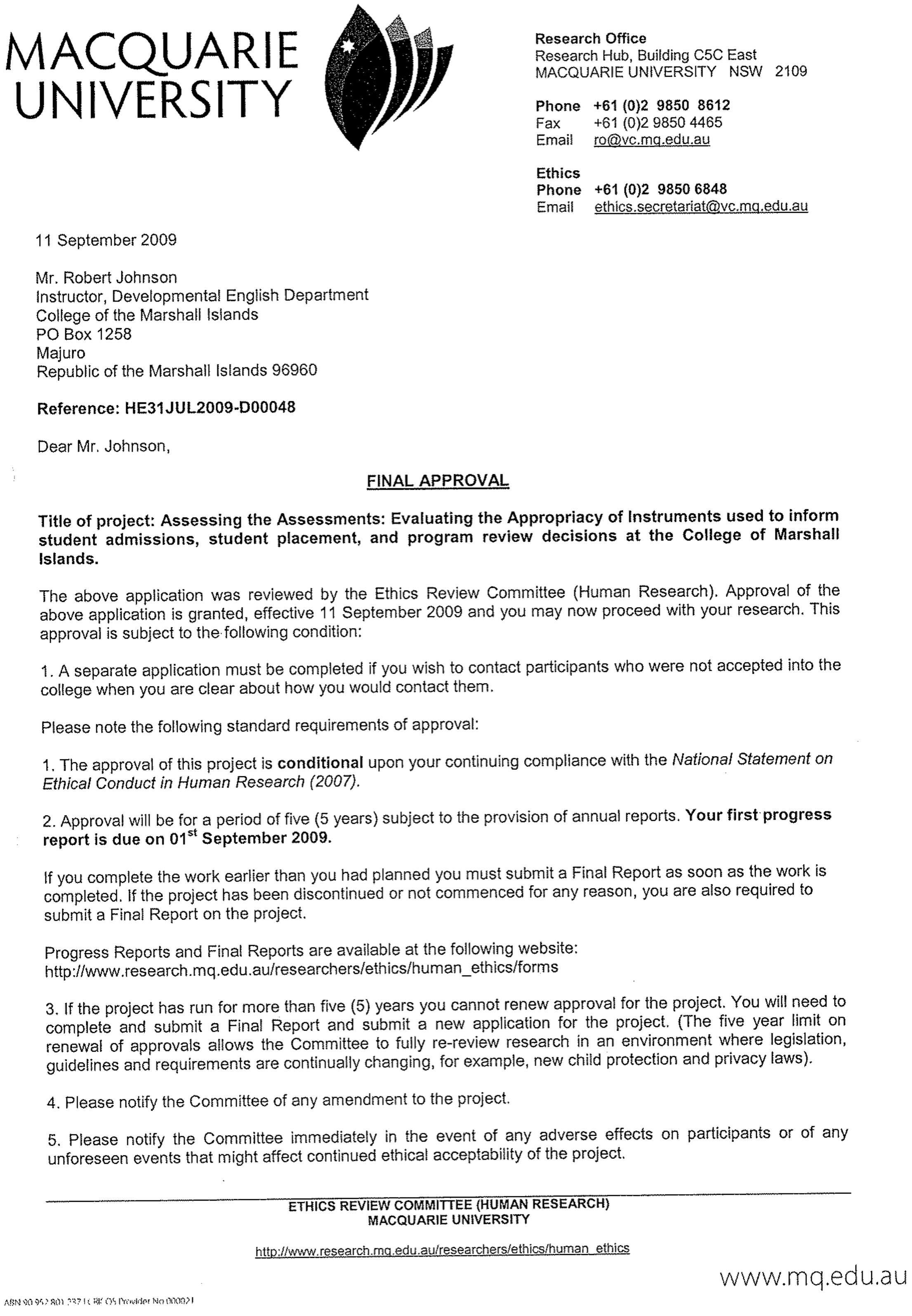
Participant's Signature: ____________________________________Date: ________________

Participant's Name: ____
(please print)

Note: This research has been reviewed and approved by the campus community of the College of the Marshall Islands. If you have any complaints or concerns about this research, please contact the principal researcher (above), Ellia Zebedy, Vice President for Research, Planning and Grants at CMI (email: ezebedy@cmi.edu or telephone: 625-3394) or Don Hess, Interim Vice President of Academic Affairs and Student Life and Development (email: dhess@cmi.edu or telephone: 625-3394).

Additionally, the ethical aspects of this study have been approved by the Macquarie University Ethics Review Committee (Human Research). If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Ethics Review Committee through the Director, Research Ethics (telephone +60 9850 7854; email ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

APPENDIX E: Ethics Review Committee (Human Research) Approval



MACQUARIE
UNIVERSITY

Research Office
Research Hub, Building C5C East
MACQUARIE UNIVERSITY NSW 2109

Phone **+61 (0)2 9850 8612**
Fax +61 (0)2 9850 4465
Email ro@vc.mq.edu.au

Ethics
Phone **+61 (0)2 9850 6848**
Email ethics.secretariat@vc.mq.edu.au

11 September 2009

Mr. Robert Johnson
Instructor, Developmental English Department
College of the Marshall Islands
PO Box 1258
Majuro
Republic of the Marshall Islands 96960

Reference: HE31JUL2009-D00048

Dear Mr. Johnson,

FINAL APPROVAL

Title of project: Assessing the Assessments: Evaluating the Appropriacy of Instruments used to inform student admissions, student placement, and program review decisions at the College of Marshall Islands.

The above application was reviewed by the Ethics Review Committee (Human Research). Approval of the above application is granted, effective 11 September 2009 and you may now proceed with your research. This approval is subject to the following condition:

1. A separate application must be completed if you wish to contact participants who were not accepted into the college when you are clear about how you would contact them.

Please note the following standard requirements of approval:

1. The approval of this project is **conditional** upon your continuing compliance with the *National Statement on Ethical Conduct in Human Research (2007).*

2. Approval will be for a period of five (5 years) subject to the provision of annual reports. **Your first progress report is due on 01st September 2009.**

If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report on the project.

Progress Reports and Final Reports are available at the following website:
http://www.research.mq.edu.au/researchers/ethics/human_ethics/forms

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. Please notify the Committee of any amendment to the project.

5. Please notify the Committee immediately in the event of any adverse effects on participants or of any unforeseen events that might affect continued ethical acceptability of the project.

ETHICS REVIEW COMMITTEE (HUMAN RESEARCH)
MACQUARIE UNIVERSITY

http://www.research.mq.edu.au/researchers/ethics/human_ethics

www.mq.edu.au

ABN 90 952 801 237 CRICOS Provider No 00002J

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at: http://www.research.mq.edu.au/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide Macquarie University's Research Grants Officer with a copy of this letter as soon as possible. The Research Grants Officer will not inform external funding agencies that you have final approval for your project and funds will not be released until the Research Grants Officer has received a copy of this final approval letter.

Yours sincerely

Dr Karolyn White
Director of Research Ethics
Chair, Ethics Review Committee (Human Research)

Cc: Dr Stephen H Moore, Department of Linguistics, Macquarie University

APPENDIX F: Instructor Questionnaire regarding First-Semester Student Placement

Instructor Questionnaire **Student Placement**

Instructor: Course: Semester:

Instructions:

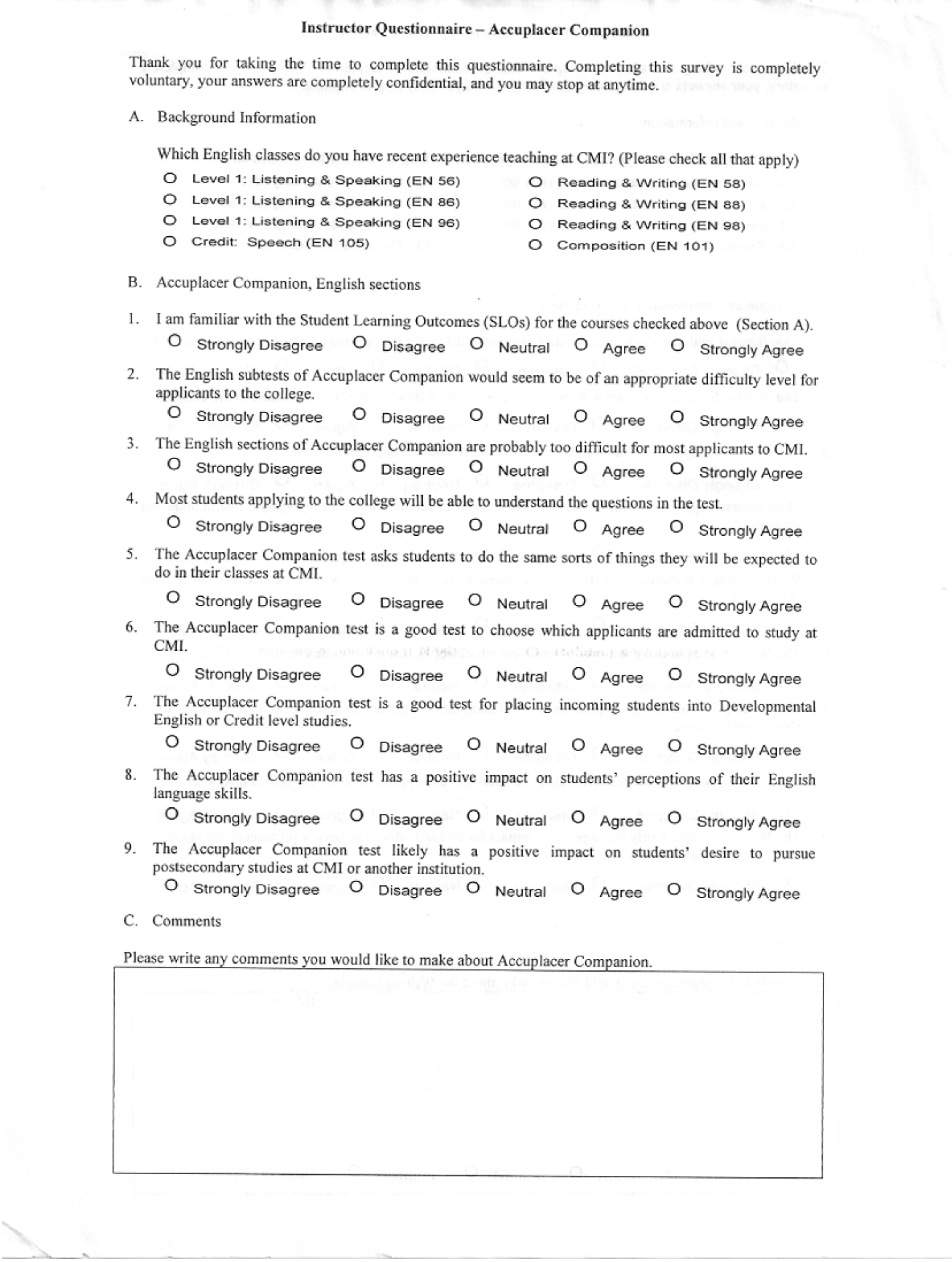
Listed below are each of the new (first-semester) students enrolled in your English class for this semester. For each student, indicate which level of English class you feel they should have, ideally, been placed in at the beginning of the semester, based entirely on the language skills relevant to the course.

Placement Level Key:

0 – not ready for any Developmental English course at CMI
1 – Developmental English Level 1
2 – Developmental English Level 2
3 – Developmental English Level 3
4 – Credit English

	Student Name	Student Number	Ideal Placement				
			4	3	2	1	0
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							

APPENDIX G: Instructor Questionnaire regarding Placement Instruments



Instructor Questionnaire – Accuplacer Companion

Thank you for taking the time to complete this questionnaire. Completing this survey is completely voluntary, your answers are completely confidential, and you may stop at anytime.

A. Background Information

Which English classes do you have recent experience teaching at CMI? (Please check all that apply)

○ Level 1: Listening & Speaking (EN 56)
○ Level 1: Listening & Speaking (EN 86)
○ Level 1: Listening & Speaking (EN 96)
○ Credit: Speech (EN 105)
○ Reading & Writing (EN 58)
○ Reading & Writing (EN 88)
○ Reading & Writing (EN 98)
○ Composition (EN 101)

B. Accuplacer Companion, English sections

1. I am familiar with the Student Learning Outcomes (SLOs) for the courses checked above (Section A).
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

2. The English subtests of Accuplacer Companion would seem to be of an appropriate difficulty level for applicants to the college.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

3. The English sections of Accuplacer Companion are probably too difficult for most applicants to CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

4. Most students applying to the college will be able to understand the questions in the test.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

5. The Accuplacer Companion test asks students to do the same sorts of things they will be expected to do in their classes at CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

6. The Accuplacer Companion test is a good test to choose which applicants are admitted to study at CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

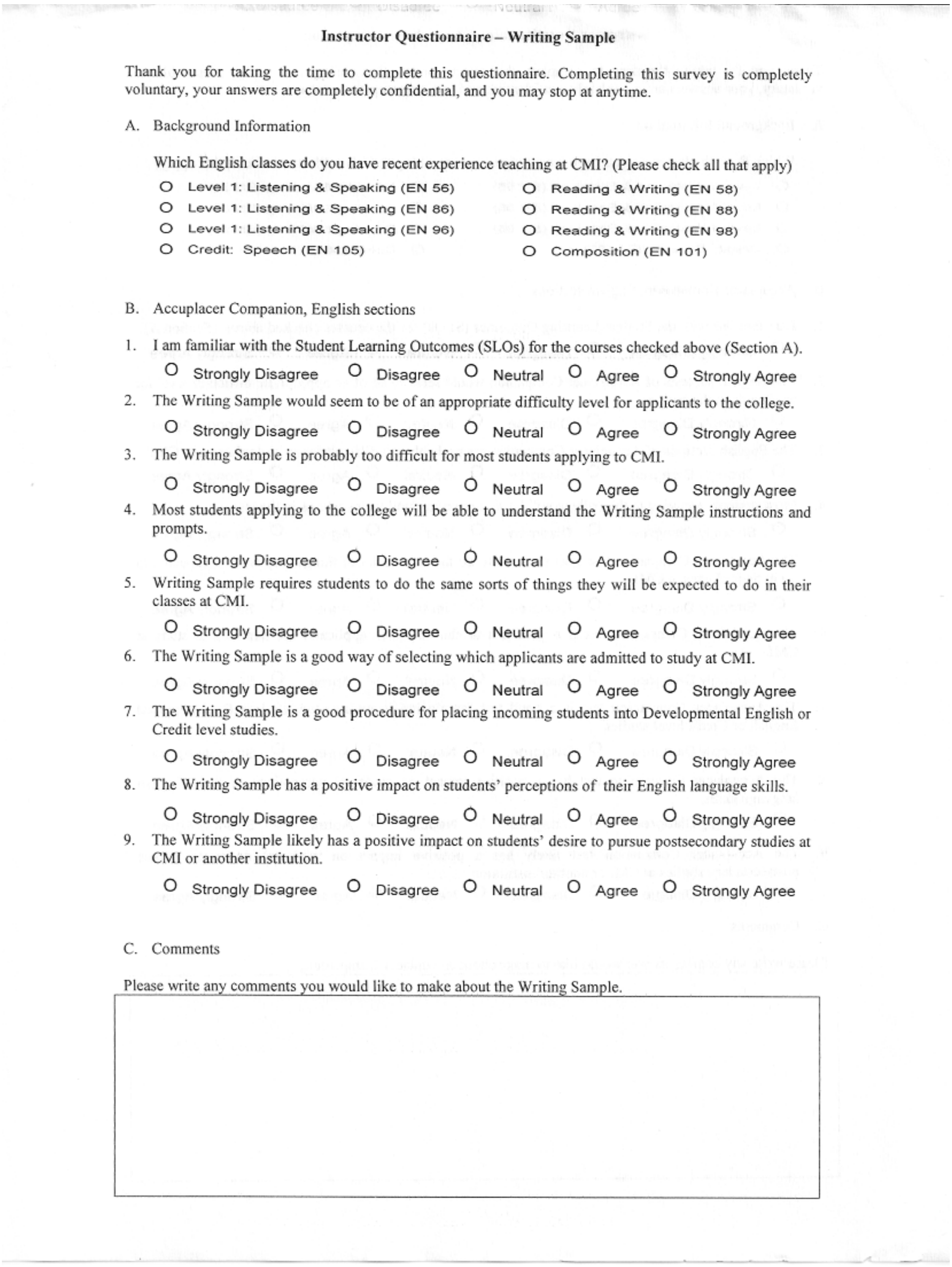
7. The Accuplacer Companion test is a good test for placing incoming students into Developmental English or Credit level studies.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

8. The Accuplacer Companion test has a positive impact on students' perceptions of their English language skills.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

9. The Accuplacer Companion test likely has a positive impact on students' desire to pursue postsecondary studies at CMI or another institution.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

C. Comments

Please write any comments you would like to make about Accuplacer Companion.



Instructor Questionnaire – Writing Sample

Thank you for taking the time to complete this questionnaire. Completing this survey is completely voluntary, your answers are completely confidential, and you may stop at anytime.

A. Background Information

Which English classes do you have recent experience teaching at CMI? (Please check all that apply)

○ Level 1: Listening & Speaking (EN 56)
○ Level 1: Listening & Speaking (EN 86)
○ Level 1: Listening & Speaking (EN 96)
○ Credit: Speech (EN 105)
○ Reading & Writing (EN 58)
○ Reading & Writing (EN 88)
○ Reading & Writing (EN 98)
○ Composition (EN 101)

B. Accuplacer Companion, English sections

1. I am familiar with the Student Learning Outcomes (SLOs) for the courses checked above (Section A).
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

2. The Writing Sample would seem to be of an appropriate difficulty level for applicants to the college.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

3. The Writing Sample is probably too difficult for most students applying to CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

4. Most students applying to the college will be able to understand the Writing Sample instructions and prompts.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

5. Writing Sample requires students to do the same sorts of things they will be expected to do in their classes at CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

6. The Writing Sample is a good way of selecting which applicants are admitted to study at CMI.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

7. The Writing Sample is a good procedure for placing incoming students into Developmental English or Credit level studies.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

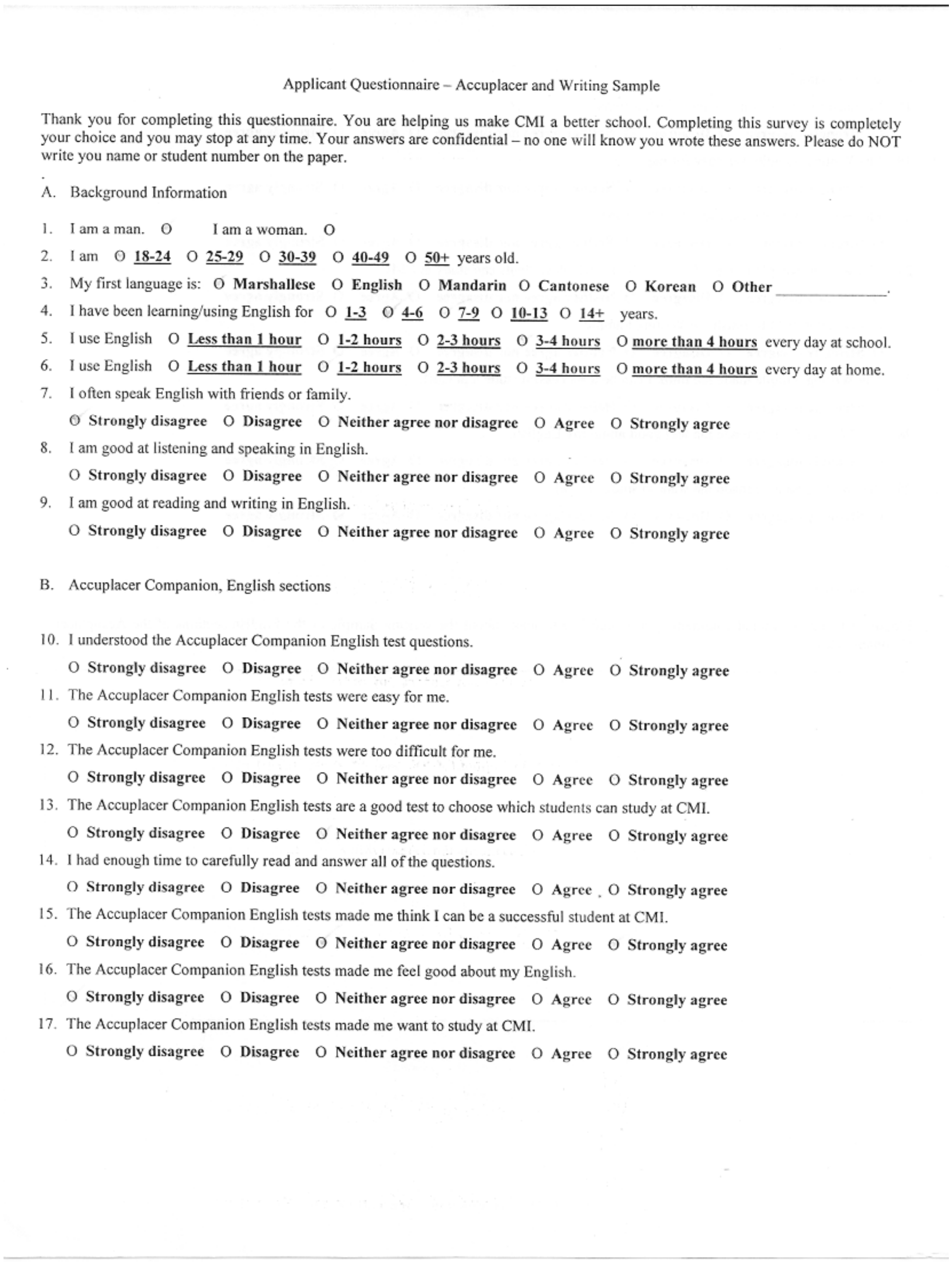
8. The Writing Sample has a positive impact on students' perceptions of their English language skills.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

9. The Writing Sample likely has a positive impact on students' desire to pursue postsecondary studies at CMI or another institution.
○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

C. Comments

Please write any comments you would like to make about the Writing Sample.

APPENDIX H: Candidate Questionnaire regarding Placement Instruments



Applicant Questionnaire – Accuplacer and Writing Sample

Thank you for completing this questionnaire. You are helping us make CMI a better school. Completing this survey is completely your choice and you may stop at any time. Your answers are confidential – no one will know you wrote these answers. Please do NOT write you name or student number on the paper.

.

A. Background Information

1. I am a man. ○ I am a woman. ○
2. I am ○ **18-24** ○ **25-29** ○ **30-39** ○ **40-49** ○ **50+** years old.
3. My first language is: ○ **Marshallese** ○ **English** ○ **Mandarin** ○ **Cantonese** ○ **Korean** ○ **Other** ______________.
4. I have been learning/using English for ○ **1-3** ○ **4-6** ○ **7-9** ○ **10-13** ○ **14+** years.
5. I use English ○ **Less than 1 hour** ○ **1-2 hours** ○ **2-3 hours** ○ **3-4 hours** ○ **more than 4 hours** every day at school.
6. I use English ○ **Less than 1 hour** ○ **1-2 hours** ○ **2-3 hours** ○ **3-4 hours** ○ **more than 4 hours** every day at home.
7. I often speak English with friends or family.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
8. I am good at listening and speaking in English.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
9. I am good at reading and writing in English.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

B. Accuplacer Companion, English sections

10. I understood the Accuplacer Companion English test questions.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
11. The Accuplacer Companion English tests were easy for me.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
12. The Accuplacer Companion English tests were too difficult for me.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
13. The Accuplacer Companion English tests are a good test to choose which students can study at CMI.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
14. I had enough time to carefully read and answer all of the questions.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
15. The Accuplacer Companion English tests made me think I can be a successful student at CMI.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
16. The Accuplacer Companion English tests made me feel good about my English.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
17. The Accuplacer Companion English tests made me want to study at CMI.
    ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

C. Writing Sample

18. I understood the Writing Sample instructions.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

19. The Writing Sample was easy for me.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

20. The Writing Sample was too difficult for me.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

21. The Writing Sample is a good way to choose which students can study at CMI.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

22. I had enough time to finish the Writing Sample.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

23. The Writing Sample made me think I can be a successful student at CMI.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

24. The Writing Sample made me feel good about my English.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

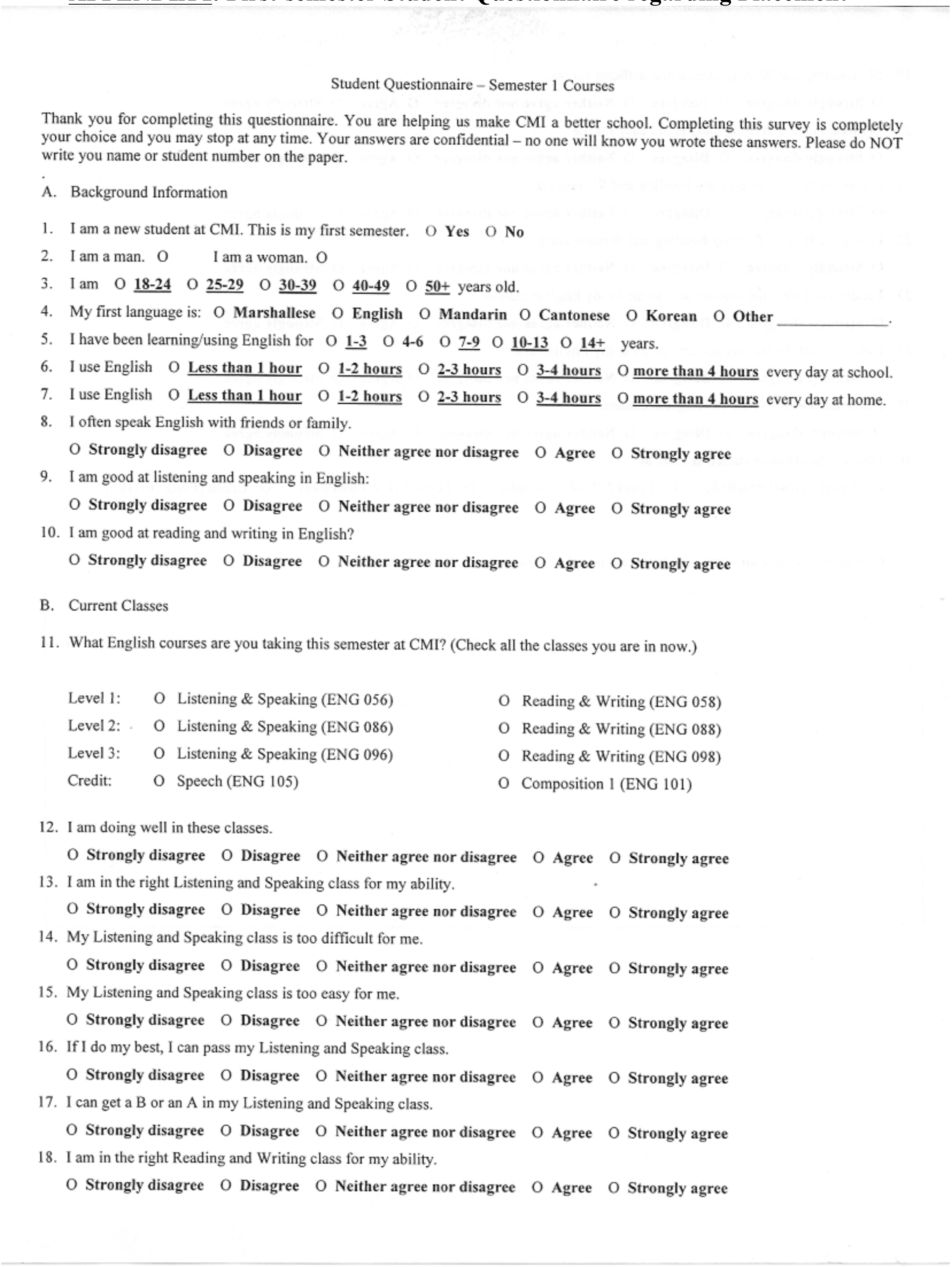
25. The Writing Sample made me want to study at CMI.

○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

D. Comments

Please write any additional_comments you would like to make about the Writing Sample or the English sections of the Accuplacer Companion test.

APPENDIX I: First-semester Student Questionnaire regarding Placement



Student Questionnaire – Semester 1 Courses

Thank you for completing this questionnaire. You are helping us make CMI a better school. Completing this survey is completely your choice and you may stop at any time. Your answers are confidential – no one will know you wrote these answers. Please do NOT write you name or student number on the paper.

.

A. Background Information

1. I am a new student at CMI. This is my first semester. ○ **Yes** ○ **No**
2. I am a man. ○ I am a woman. ○
3. I am ○ **18-24** ○ **25-29** ○ **30-39** ○ **40-49** ○ **50+** years old.
4. My first language is: ○ **Marshallese** ○ **English** ○ **Mandarin** ○ **Cantonese** ○ **Korean** ○ **Other** ____________.
5. I have been learning/using English for ○ **1-3** ○ **4-6** ○ **7-9** ○ **10-13** ○ **14+** years.
6. I use English ○ **Less than 1 hour** ○ **1-2 hours** ○ **2-3 hours** ○ **3-4 hours** ○ **more than 4 hours** every day at school.
7. I use English ○ **Less than 1 hour** ○ **1-2 hours** ○ **2-3 hours** ○ **3-4 hours** ○ **more than 4 hours** every day at home.
8. I often speak English with friends or family.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
9. I am good at listening and speaking in English:
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
10. I am good at reading and writing in English?
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

B. Current Classes

11. What English courses are you taking this semester at CMI? (Check all the classes you are in now.)

Level 1:	○ Listening & Speaking (ENG 056)	○ Reading & Writing (ENG 058)
Level 2:	○ Listening & Speaking (ENG 086)	○ Reading & Writing (ENG 088)
Level 3:	○ Listening & Speaking (ENG 096)	○ Reading & Writing (ENG 098)
Credit:	○ Speech (ENG 105)	○ Composition 1 (ENG 101)

12. I am doing well in these classes.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
13. I am in the right Listening and Speaking class for my ability.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
14. My Listening and Speaking class is too difficult for me.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
15. My Listening and Speaking class is too easy for me.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
16. If I do my best, I can pass my Listening and Speaking class.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
17. I can get a B or an A in my Listening and Speaking class.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**
18. I am in the right Reading and Writing class for my ability.
   ○ **Strongly disagree** ○ **Disagree** ○ **Neither agree nor disagree** ○ **Agree** ○ **Strongly agree**

19. My Reading and Writing class is too difficult for me.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

20. My Reading and Writing class is too easy for me.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

21. If I do my best, I can pass my Reading and Writing class.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

22. I can get a B or an A in my Reading and Writing class.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

23. I understand what the teacher says to me in my English classes.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

24. I can do what the teacher asks me to do in my English classes.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

25. I have the ability to pass my English classes.

O **Strongly disagree** O **Disagree** O **Neither agree nor disagree** O **Agree** O **Strongly agree**

26. I think I should have started at CMI in:

O Level 1 (ENG 056/058) O Level 2 (ENG 086/088) O Level 3 (ENG 096/098) O Credit (ENG 101/105)

C. Comments: Is there anything else you want to tell us about your English classes?

APPENDIX J: Rasch Analysis Results for Candidate Ability Facet

```
+---------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit       Outfit   |Estim.| Correlation |                     |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num Examinees       |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|    58      10      5.8   5.84|   5.40   .78 |  .91    .0  1.02    .2| 1.00 |   .20   .27 | 127 127             |
|    54      10      5.4   5.33|   3.49   .52 | 2.62   2.6  2.64   2.6| -.60 |  -.28   .32 | 592 592             |
|    50      10      5.0   5.14|   3.07   .44 | 1.24    .6  1.32    .7|  .61 |   .07   .36 | 147 147             |
|    49      10      4.9   5.05|   2.88   .43 |  .52  -1.1   .53  -1.1| 1.47 |   .75   .37 | 139 139             |
|    49      10      4.9   5.05|   2.88   .43 |  .97    .0  1.02    .2|  .96 |  -.57   .37 | 159 159             |
|    48      10      4.8   4.96|   2.70   .42 |  .47  -1.2   .50  -1.2| 1.46 |   .86   .38 | 155 155             |
|    48      10      4.8   4.94|   2.67   .42 |  .45  -1.3   .46  -1.3| 1.55 |   .53   .45 | 605 605             |
|    46      10      4.6   4.90|   2.58   .40 | 1.81   1.6  1.60   1.2|  .05 |  -.19   .45 | 525 525             |
|    47      10      4.7   4.87|   2.53   .41 |  .43  -1.5   .46  -1.3| 1.60 |   .13   .39 | 158 158             |
|    45      10      4.5   4.85|   2.49   .39 |  .60   -.9   .61   -.8| 1.40 |   .67   .43 | 211 211             |
|    47      10      4.7   4.81|   2.43   .41 |  .39  -1.6   .36  -1.7| 1.66 |   .37   .49 | 124 124             |
|    46      10      4.6   4.78|   2.36   .40 |  .23  -2.4   .21  -2.5| 1.82 |   .63   .40 | 140 140             |
|    46      10      4.6   4.78|   2.36   .40 |  .12  -3.3   .11  -3.3| 1.93 |   .83   .40 | 156 156             |
|    44      10      4.4   4.77|   2.35   .38 |  .66   -.7   .62   -.8| 1.41 |   .28   .44 | 524 524             |
|    47      10      4.7   4.76|   2.34   .41 | 1.64   1.3  1.61   1.3|  .40 |   .78   .49 | 809 809             |
|    51      10      5.1   4.71|   2.25   .46 | 1.12    .4   .98    .0|  .78 |  -.41   .44 | 222 222             |
|    46      10      4.6   4.67|   2.18   .40 |  .36  -1.8   .39  -1.6| 1.56 |   .70   .50 | 846 846             |
|    45      10      4.5   4.62|   2.11   .39 |  .90    .0   .88   -.1| 1.03 |   .35   .51 | 125 125             |
|    47      10      4.7   4.55|   1.99   .41 |  .57  -1.0   .56  -1.0| 1.37 |   .40   .39 | 578 578             |
|    42      10      4.2   4.50|   1.92   .38 |  .73   -.5   .69   -.6| 1.26 |   .69   .61 | 476 476             |
|    41      10      4.1   4.49|   1.91   .37 | 1.69   1.5  1.61   1.3|  .20 |   .58   .46 | 204 204             |
|    42      10      4.2   4.45|   1.84   .37 | 1.55   1.2  1.51   1.1|  .36 |   .63   .44 | 632 632             |
|    42      10      4.2   4.40|   1.77   .37 | 1.03    .2  1.03    .2|  .94 |   .48   .43 | 131 131             |
|    42      10      4.2   4.40|   1.77   .37 |  .31  -2.1   .30  -2.2| 1.82 |   .86   .43 | 145 145             |
|    43      10      4.3   4.39|   1.76   .38 | 2.99   3.2  2.99   3.2|-1.20 |  -.42   .50 | 640 640             |
|    41      10      4.1   4.30|   1.63   .37 |  .86   -.1   .88   -.1|  .96 |   .30   .43 | 149 149             |
|    41      10      4.1   4.30|   1.63   .37 | 1.08    .3  1.07    .3|  .99 |   .07   .43 | 150 150             |
|    41      10      4.1   4.30|   1.63   .37 |  .58  -1.0   .58  -1.0| 1.55 |   .64   .43 | 151 151             |
|    43      10      4.3   4.28|   1.60   .38 |  .90    .0   .85   -.2| 1.04 |   .46   .45 | 532 532             |
|    44      10      4.4   4.26|   1.58   .38 | 1.70   1.4  1.69   1.4|  .21 |  -.30   .41 | 308 308             |
|    44      10      4.4   4.26|   1.58   .38 |  .49  -1.3   .45  -1.4| 1.71 |   .76   .41 | 552 552             |
|    44      10      4.4   4.26|   1.58   .38 |  .94    .0  1.03    .2| 1.17 |   .18   .41 | 595 595             |
|    40      10      4.0   4.26|   1.57   .37 |  .50  -1.3   .50  -1.3| 1.59 |   .62   .45 | 462 462             |
|    40      10      4.0   4.26|   1.57   .37 |  .37  -1.8   .36  -1.8| 1.65 |   .56   .45 | 823 823             |
|    44      10      4.4   4.25|   1.56   .38 | 2.63   2.8  2.57   2.7| -.84 |  -.02   .41 | 584 584             |
|    41      10      4.1   4.22|   1.52   .37 | 1.60   1.3  1.62   1.3|  .27 |   .75   .53 | 123 123             |
|    42      10      4.2   4.21|   1.50   .38 |  .41  -1.6   .39  -1.7| 1.68 |   .69   .50 | 622 622             |
|    40      10      4.0   4.20|   1.50   .37 |  .58  -1.0   .59  -1.0| 1.48 |   .50   .44 | 148 148             |
|    42      10      4.2   4.14|   1.42   .38 |  .21  -2.7   .21  -2.7| 1.92 |   .93   .47 | 619 619             |
|    41      10      4.1   4.13|   1.40   .37 |  .88   -.1   .84   -.2|  .96 |   .34   .49 | 498 498             |
|    36      10      3.6   4.13|   1.40   .36 |  .58  -1.0   .57  -1.0| 1.46 |   .35   .51 | 442 442             |
|    38      10      3.8   4.11|   1.37   .36 | 2.23   2.3  2.19   2.2| -.22 |   .68   .44 | 472 472             |
|    39      10      3.9   4.10|   1.36   .36 |  .55  -1.1   .55  -1.1| 1.40 |   .06   .44 | 134 134             |
|    39      10      3.9   4.10|   1.36   .36 |  .56  -1.0   .56  -1.0| 1.38 |   .03   .44 | 144 144             |
|    39      10      3.9   4.10|   1.36   .36 |  .42  -1.6   .43  -1.6| 1.54 |   .28   .44 | 157 157             |
|    41      10      4.1   4.10|   1.36   .37 |  .15  -3.1   .15  -3.1| 1.84 |   .79   .50 | 621 621             |
|    46      10      4.6   4.09|   1.34   .40 |  .22  -2.5   .22  -2.5| 1.81 |   .65   .49 | 219 219             |
|    38      10      3.8   4.06|   1.30   .36 | 1.59   1.3  1.57   1.2|  .31 |   .67   .46 | 456 456             |
|    38      10      3.8   4.06|   1.30   .36 |  .21  -2.7   .21  -2.7| 1.73 |   .50   .46 | 811 811             |
|    36      10      3.6   4.02|   1.26   .36 | 1.03    .2  1.04    .2|  .97 |  -.31   .47 | 519 519             |
|    36      10      3.6   4.02|   1.26   .36 |  .52  -1.2   .52  -1.2| 1.52 |   .38   .47 | 528 528             |
|    40      10      4.0   4.02|   1.25   .37 |  .70   -.6   .68   -.7| 1.46 |   .61   .49 | 469 469             |
|    37      10      3.7   4.01|   1.24   .36 | 1.72   1.5  1.72   1.5|  .36 |   .05   .45 | 473 473             |
|    38      10      3.8   4.00|   1.23   .36 |  .82   -.3   .81   -.3| 1.13 |  -.24   .44 | 138 138             |
|    43      10      4.3   4.00|   1.22   .39 | 1.14    .4  1.19    .5|  .84 |   .18   .63 | 730 730             |
|    37      10      3.7   4.00|   1.22   .37 |  .65   -.7   .67   -.7| 1.31 |   .39   .63 | 482 482             |
|    37      10      3.7   4.00|   1.22   .37 | 3.12   3.3  3.04   3.2|-1.43 |   .82   .63 | 488 488             |
|    39      10      3.9   3.99|   1.21   .37 | 2.43   2.5  2.50   2.6| -.62 |  -.42   .55 | 613 613             |
|    39      10      3.9   3.98|   1.20   .37 | 1.03    .2  1.04    .2|  .93 |  -.07   .52 | 644 644             |
|    45      10      4.5   3.97|   1.18   .39 |  .16  -3.0   .16  -3.0| 1.86 |   .77   .50 | 216 216             |
|    39      10      3.9   3.95|   1.17   .37 |  .87   -.1   .87   -.1| 1.10 |   .12   .54 | 830 830             |
|    44      10      4.4   3.95|   1.16   .39 |  .32  -2.0   .30  -2.1| 1.76 |   .72   .55 | 731 731             |
|    41      10      4.1   3.94|   1.16   .37 | 1.16    .4  1.14    .4|  .74 |   .26   .43 | 596 596             |
|    37       9      4.1   3.90|   1.10   .39 | 4.27   4.3  4.16   4.2|-2.76 |   .18   .43 | 485 485             |
|    37      10      3.7   3.90|   1.10   .36 |  .38  -1.7   .38  -1.8| 1.57 |   .20   .45 | 142 142             |
|    41      10      4.1   3.90|   1.10   .37 | 1.84   1.7  1.75   1.5|  .12 |   .31   .44 | 575 575             |
|    39      10      3.9   3.86|   1.05   .37 | 1.33    .8  1.32    .8|  .54 |   .10   .47 | 628 628             |
|    36      10      3.6   3.86|   1.04   .36 | 1.61   1.3  1.60   1.3|  .41 |   .83   .46 | 822 822             |
|    41      10      4.1   3.85|   1.04   .37 | 3.21   3.5  3.00   3.2|-1.35 |   .39   .43 | 435 435             |
|    44      10      4.4   3.85|   1.03   .39 |  .39  -1.7   .37  -1.7| 1.70 |   .64   .50 | 221 221             |
|    40      10      4.0   3.84|   1.02   .37 | 4.48   4.7  4.40   4.6|-3.05 |  -.14   .44 | 557 557             |
|    40      10      4.0   3.84|   1.02   .37 | 4.48   4.7  4.40   4.6|-3.05 |  -.14   .44 | 558 558             |
|    40      10      4.0   3.84|   1.02   .37 |  .74   -.5   .73   -.5| 1.43 |   .59   .44 | 597 597             |
|    34      10      3.4   3.82|   1.00   .36 | 1.11    .3  1.14    .4|  .93 |  -.43   .47 | 512 512             |
|    38      10      3.8   3.82|   1.00   .36 | 1.61   1.3  1.57   1.2|  .49 |   .74   .50 | 541 541             |
+---------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit       Outfit    |Estim.| Correlation |                      |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num Examinees        |
|-------------------------------+--------------+---------------------+------+-------------+----------------------|
|    35      10      3.5   3.82|    .99   .36 | 1.98  1.9  1.98  1.9|  .01 |   .27   .45 | 637 637              |
|    36      10      3.6   3.80|    .97   .36 |  .72  -.5   .71  -.6| 1.35 |   .61   .45 | 137 137              |
|    35      10      3.5   3.79|    .95   .36 | 1.47  1.1  1.42  1.0|  .46 |   .67   .63 | 480 480              |
|    37      10      3.7   3.78|    .94   .36 | 1.25   .6  1.23   .6|  .85 |   .81   .56 | 610 610              |
|    41      10      4.1   3.77|    .93   .38 |  .22 -2.6   .20 -2.6| 1.81 |   .84   .64 | 171 171              |
|    38      10      3.8   3.76|    .92   .36 | 2.09  2.1  2.07  2.0| -.30 |   .14   .48 | 649 649              |
|    37      10      3.7   3.74|    .90   .36 | 1.19   .5  1.18   .5|  .95 |   .75   .54 | 805 805              |
|    43      10      4.3   3.73|    .88   .38 |  .16 -3.0   .15 -3.1| 1.91 |   .87   .51 | 217 217              |
|    35      10      3.5   3.72|    .87   .36 |  .59  -.9   .59  -.9| 1.48 |   .28   .45 | 634 634              |
|    37      10      3.7   3.72|    .86   .36 |  .53 -1.2   .53 -1.2| 1.53 |   .78   .46 | 555 555              |
|    37      10      3.7   3.72|    .86   .36 |  .57 -1.0   .57 -1.0| 1.48 |   .74   .46 | 604 604              |
|    35      10      3.5   3.70|    .84   .36 |  .45 -1.5   .45 -1.4| 1.56 |   .16   .45 | 152 152              |
|    34      10      3.4   3.68|    .82   .36 |  .45 -1.4   .45 -1.4| 1.57 |   .78   .63 | 484 484              |
|    34      10      3.4   3.66|    .79   .36 |  .36 -1.8   .36 -1.8| 1.68 |   .30   .44 | 448 448              |
|    33      10      3.3   3.66|    .79   .36 | 1.60  1.3  1.59  1.3|  .41 |   .36   .49 | 517 517              |
|    37      10      3.7   3.65|    .78   .36 |  .61  -.9   .61  -.9| 1.44 |   .69   .48 | 612 612              |
|    37      10      3.7   3.65|    .78   .36 |  .80  -.3   .79  -.3| 1.27 |   .27   .48 | 648 648              |
|    34      10      3.4   3.65|    .78   .36 |  .54 -1.1   .55 -1.1| 1.48 |  -.03   .46 | 810 810              |
|    38      10      3.8   3.63|    .75   .36 | 1.43  1.0  1.42  1.0|  .57 |   .03   .44 | 566 566              |
|    38      10      3.8   3.63|    .75   .36 | 1.08   .3  1.07   .2|  .96 |   .38   .44 | 600 600              |
|    37      10      3.7   3.62|    .74   .36 |  .76  -.4   .76  -.4| 1.35 |   .71   .49 | 623 623              |
|    34      10      3.4   3.62|    .74   .36 |  .83  -.2   .83  -.2| 1.11 |   .50   .45 | 630 630              |
|    35      10      3.5   3.60|    .72   .36 |  .35 -1.9   .35 -1.9| 1.74 |   .65   .54 | 126 126              |
|    34      10      3.4   3.60|    .71   .36 |  .29 -2.2   .28 -2.2| 1.76 |   .45   .45 | 141 141              |
|    34      10      3.4   3.60|    .71   .36 | 2.44  2.5  2.45  2.5| -.71 |   .84   .45 | 153 153              |
|    38      10      3.8   3.58|    .70   .36 |  .60  -.9   .60  -.9| 1.56 |   .67   .45 | 571 571              |
|    38      10      3.8   3.58|    .70   .36 | 1.87  1.7  1.84  1.7|  .12 |   .10   .45 | 580 580              |
|    33      10      3.3   3.58|    .69   .36 | 2.54  2.7  2.55  2.7| -.67 |  -.39   .63 | 494 494              |
|    35      10      3.5   3.57|    .68   .36 |  .34 -1.9   .35 -1.9| 1.75 |   .65   .48 | 443 443              |
|    36      10      3.6   3.55|    .66   .38 |  .66  -.7   .64  -.8| 1.48 |   .92   .77 | 479 479              |
|    36      10      3.6   3.55|    .65   .36 |  .74  -.5   .74  -.5| 1.32 |   .58   .48 | 533 533              |
|    33      10      3.3   3.55|    .65   .36 |  .30 -2.1   .30 -2.1| 1.72 |   .66   .46 | 818 818              |
|    33      10      3.3   3.55|    .65   .36 |  .99   .1  1.00   .1| 1.00 |   .69   .46 | 836 836              |
|    39      10      3.9   3.55|    .65   .37 |  .13 -3.3   .13 -3.2| 1.99 |   .95   .64 | 170 170              |
|    38      10      3.8   3.54|    .64   .36 | 2.76  2.9  2.74  2.9| -.97 |  -.18   .44 | 429 429              |
|    35      10      3.5   3.54|    .64   .36 |  .66  -.7   .66  -.7| 1.32 |   .01   .54 | 817 817              |
|    37      10      3.7   3.53|    .62   .36 |  .34 -1.9   .34 -1.9| 1.70 |   .60   .45 | 563 563              |
|    37      10      3.7   3.53|    .62   .36 |  .63  -.8   .63  -.8| 1.49 |   .86   .45 | 590 590              |
|    37      10      3.7   3.53|    .62   .36 | 1.72  1.5  1.70  1.5|  .22 |   .44   .45 | 598 598              |
|    37      10      3.7   3.53|    .62   .36 | 1.03   .2  1.02   .1|  .97 |   .25   .45 | 599 599              |
|    31      10      3.1   3.51|    .61   .36 |  .21 -2.6   .21 -2.6| 1.87 |   .67   .47 | 522 522              |
|    41      10      4.1   3.51|    .60   .37 |  .83  -.2   .82  -.3| 1.27 |   .41   .52 | 215 215              |
|    37      10      3.7   3.49|    .58   .36 | 1.15   .4  1.14   .4|  .91 |   .73   .45 | 108 108              |
|    33      10      3.3   3.49|    .58   .36 |  .25 -2.4   .25 -2.4| 1.82 |   .48   .45 | 132 132              |
|    33      10      3.3   3.49|    .58   .36 |  .20 -2.6   .20 -2.7| 1.87 |   .58   .45 | 135 135              |
|    33      10      3.3   3.49|    .58   .36 |  .27 -2.2   .27 -2.2| 1.75 |   .71   .45 | 136 136              |
|    33      10      3.3   3.49|    .58   .36 |  .22 -2.6   .22 -2.6| 1.93 |   .78   .45 | 143 143              |
|    33      10      3.3   3.49|    .58   .36 | 1.05   .2  1.05   .2|  .98 |   .46   .45 | 146 146              |
|    35      10      3.5   3.48|    .56   .36 | 1.51  1.1  1.50  1.1|  .53 |   .59   .52 | 625 625              |
|    34      10      3.4   3.47|    .55   .36 |  .88  -.1   .89  -.1| 1.17 |   .86   .53 | 645 645              |
|    33      10      3.3   3.46|    .53   .36 |  .23 -2.5   .23 -2.5| 1.88 |   .89   .52 | 606 606              |
|    31      10      3.1   3.45|    .53   .36 |  .59  -.9   .59  -.9| 1.37 |   .59   .49 | 526 526              |
|    32      10      3.2   3.45|    .52   .36 |  .72  -.5   .72  -.5| 1.37 |   .73   .46 | 635 635              |
|    32      10      3.2   3.45|    .52   .36 |  .72  -.5   .72  -.5| 1.29 |  -.07   .46 | 831 831              |
|    47      15      3.1   3.44|    .51   .30 |  .71  -.8   .71  -.8| 1.29 |   .57   .55 | 439 439              |
|    36      10      3.6   3.41|    .48   .36 |  .27 -2.3   .27 -2.3| 1.81 |   .75   .45 | 585 585              |
|    34      10      3.4   3.41|    .47   .36 |  .34 -1.9   .34 -1.9| 1.70 |   .42   .50 | 537 537              |
|    31      10      3.1   3.41|    .47   .36 |  .56 -1.0   .57 -1.0| 1.48 |   .63   .45 | 629 629              |
|    34      10      3.4   3.41|    .47   .36 |  .61  -.9   .61  -.9| 1.44 |   .52   .46 | 451 451              |
|    34      10      3.4   3.41|    .47   .36 | 1.80  1.6  1.81  1.6|  .20 |   .56   .46 | 607 607              |
|    30      10      3.0   3.40|    .46   .36 |  .55 -1.0   .54 -1.1| 1.57 |   .55   .47 | 200 200              |
|    30      10      3.0   3.40|    .46   .36 |  .56 -1.0   .55 -1.1| 1.52 |   .74   .47 | 202 202              |
|    32      10      3.2   3.39|    .45   .36 |  .16 -3.0   .16 -3.0| 1.90 |   .83   .45 | 133 133              |
|    33      10      3.3   3.37|    .42   .36 | 3.66  3.8  3.64  3.8|-2.12 |   .71   .48 | 441 441              |
|    30      10      3.0   3.35|    .39   .36 | 1.22   .6  1.25   .6|  .69 |  -.20   .49 | 523 523              |
|    31      10      3.1   3.35|    .39   .36 | 1.19   .5  1.18   .5|  .72 |  -.08   .46 | 458 458              |
|    31      10      3.1   3.35|    .39   .36 |  .87  -.1   .87  -.1| 1.15 |  -.49   .46 | 804 804              |
|    35      10      3.5   3.32|    .36   .36 |  .41 -1.6   .41 -1.6| 1.75 |   .77   .45 | 437 437              |
|    35      10      3.5   3.32|    .36   .36 |  .84  -.2   .83  -.2| 1.15 |   .24   .45 | 478 478              |
|    35      10      3.5   3.32|    .36   .36 | 1.12   .4  1.12   .4|  .92 |   .61   .45 | 556 556              |
|    35      10      3.5   3.32|    .36   .36 | 1.00   .1  1.00   .1| 1.05 |   .88   .45 | 568 568              |
|    35      10      3.5   3.32|    .36   .36 |  .86  -.2   .85  -.2| 1.22 |   .48   .45 | 591 591              |
|    34      10      3.4   3.31|    .35   .36 |  .24 -2.4   .24 -2.4| 1.81 |   .58   .49 | 611 611              |
|    34      10      3.4   3.31|    .35   .36 |  .33 -1.9   .33 -1.9| 1.66 |   .65   .49 | 627 627              |
|    28      10      2.8   3.31|    .35   .37 |  .43 -1.5   .43 -1.5| 1.64 |   .46   .51 | 438 438              |
|    35      10      3.5   3.31|    .35   .36 |  .48 -1.3   .47 -1.4| 1.60 |   .45   .45 | 586 586              |
+--------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                      |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num Examinees        |
|-------------------------------+--------------+---------------------+------+-------------+----------------------|
|   33      10       3.3   3.31|    .34   .36 |  .50 -1.2   .51 -1.2| 1.54 |   .08   .50 | 531 531              |
|   33      10       3.3   3.31|    .34   .36 | 1.20   .5  1.21   .5|  .68 |  -.15   .50 | 536 536              |
|   26       8       3.3   3.30|    .33   .40 | 1.27   .6  1.26   .6|  .72 |   .13   .43 | 616 616              |
|   30      10       3.0   3.30|    .33   .36 | 1.43  1.0  1.43  1.0|  .52 |   .09   .45 | 470 470              |
|   33      10       3.3   3.27|    .30   .36 |  .14 -3.1   .14 -3.1| 1.94 |   .75   .52 | 626 626              |
|   30      10       3.0   3.25|    .27   .36 |  .71  -.5   .72  -.5| 1.26 |   .35   .44 | 447 447              |
|   28      10       2.8   3.25|    .27   .37 |  .93   .0   .93   .0| 1.05 |  -.15   .54 | 454 454              |
|   30      10       3.0   3.24|    .26   .36 | 1.00   .1  1.01   .1| 1.02 |   .68   .46 | 631 631              |
|   30      10       3.0   3.24|    .26   .36 |  .88  -.1   .88  -.1| 1.17 |   .79   .46 | 833 833              |
|   35      10       3.5   3.23|    .24   .36 | 1.46  1.0  1.45  1.0|  .54 |   .50   .45 | 427 427              |
|   35      10       3.5   3.23|    .24   .36 | 1.72  1.5  1.71  1.5|  .29 |   .47   .45 | 431 431              |
|   34      10       3.4   3.22|    .23   .36 |  .67  -.7   .66  -.7| 1.37 |   .44   .45 | 550 550              |
|   34      10       3.4   3.22|    .23   .36 | 1.00   .1   .99   .1| 1.04 |   .53   .45 | 553 553              |
|   34      10       3.4   3.22|    .23   .36 | 1.36   .9  1.37   .9|  .59 |   .74   .45 | 562 562              |
|   34      10       3.4   3.21|    .22   .36 |  .88  -.1   .87  -.1| 1.15 |   .19   .45 | 583 583              |
|   32      10       3.2   3.21|    .21   .36 |  .44 -1.5   .44 -1.5| 1.61 |   .43   .50 | 549 549              |
|   32      10       3.2   3.21|    .21   .36 | 2.56  2.6  2.58  2.7| -.69 |  -.04   .46 | 608 608              |
|   29      10       2.9   3.20|    .21   .36 |  .64  -.8   .63  -.8| 1.33 |   .56   .44 | 633 633              |
|   29      10       2.9   3.19|    .20   .36 | 2.72  2.8  2.72  2.8| -.85 |   .17   .45 | 471 471              |
|   34      10       3.4   3.19|    .19   .36 |  .79  -.3   .78  -.3| 1.15 |   .73   .45 | 104 104              |
|   34      10       3.4   3.19|    .19   .36 |  .56 -1.0   .57 -1.0| 1.44 |   .78   .45 | 105 105              |
|   44      15       2.9   3.19|    .19   .30 | 2.26  2.7  2.32  2.8| -.47 |  -.02   .55 | 455 455              |
|   34      10       3.4   3.18|    .18   .36 | 1.09   .3  1.09   .3|  .91 |  -.07   .45 | 572 572              |
|   34      10       3.4   3.18|    .18   .36 |  .61  -.9   .61  -.9| 1.43 |   .51   .45 | 573 573              |
|   31      10       3.1   3.17|    .16   .36 |  .70  -.6   .69  -.6| 1.33 |   .66   .48 | 449 449              |
|   31      10       3.1   3.16|    .15   .36 |  .45 -1.4   .45 -1.4| 1.61 |   .85   .53 | 647 647              |
|   32      10       3.2   3.14|    .13   .36 |  .54 -1.1   .54 -1.1| 1.52 |   .71   .48 | 530 530              |
|   29      10       2.9   3.14|    .13   .36 | 1.66  1.4  1.64  1.3|  .23 |   .00   .46 | 636 636              |
|   34      10       3.4   3.13|    .12   .36 | 2.82  3.0  2.81  3.0| -.99 |   .18   .45 | 587 587              |
|   31      10       3.1   3.13|    .11   .36 | 1.58  1.2  1.58  1.2|  .40 |   .79   .54 | 821 821              |
|   31      10       3.1   3.10|    .08   .36 |  .99   .1  1.00   .1|  .97 |   .40   .50 | 546 546              |
|   27      10       2.7   3.10|    .07   .37 |  .72  -.5   .72  -.5| 1.31 |   .61   .47 | 518 518              |
|   27      10       2.7   3.10|    .07   .37 |  .35 -1.9   .36 -1.8| 1.84 |   .62   .47 | 529 529              |
|   31      10       3.1   3.09|    .07   .36 |  .67  -.7   .69  -.6| 1.32 |   .28   .50 | 464 464              |
|   33      10       3.3   3.09|    .06   .36 |  .82  -.3   .81  -.3| 1.22 |   .84   .45 | 570 570              |
|   37      10       3.7   3.09|    .06   .36 | 1.21   .6  1.20   .5|  .88 |   .85   .53 | 212 212              |
|   30      10       3.0   3.06|    .02   .36 |  .39 -1.7   .39 -1.7| 1.68 |   .58   .56 | 617 617              |
|   29      10       2.9   3.05|    .01   .37 |  .48 -1.3   .48 -1.3| 1.70 |   .86   .51 | 602 602              |
|   29      10       2.9   3.05|    .01   .37 |  .63  -.8   .62  -.8| 1.35 |   .57   .51 | 609 609              |
|   33      10       3.3   3.03|   -.01   .36 | 2.78  2.9  2.78  2.9|-1.04 |   .49   .45 | 434 434              |
|   32      10       3.2   3.02|   -.03   .36 |  .29 -2.1   .29 -2.1| 1.75 |   .61   .45 | 481 481              |
|   31      10       3.1   3.01|   -.04   .37 |  .47 -1.4   .44 -1.5| 1.67 |   .86   .77 | 174 174              |
|   31      10       3.1   3.01|   -.04   .37 | 1.40   .9  1.39   .9|  .48 |   .34   .77 | 182 182              |
|   30      10       3.0   3.00|   -.05   .36 |  .56 -1.0   .58 -1.0| 1.48 |   .72   .50 | 547 547              |
|   26      10       2.6   2.99|   -.06   .37 |  .73  -.5   .73  -.5| 1.29 |   .65   .46 | 513 513              |
|   26      10       2.6   2.99|   -.06   .37 |  .73  -.5   .73  -.5| 1.29 |   .65   .46 | 527 527              |
|   25      10       2.5   2.99|   -.07   .38 |  .29 -2.2   .31 -2.1| 1.90 |   .75   .50 | 446 446              |
|   29      10       2.9   2.95|   -.11   .37 | 1.24   .6  1.24   .6|  .73 |   .04   .56 | 614 614              |
|   30      10       3.0   2.94|   -.13   .36 |  .33 -2.0   .32 -2.0| 1.73 |   .62   .48 | 539 539              |
|   27      10       2.7   2.94|   -.14   .37 |  .63  -.8   .64  -.8| 1.45 |   .49   .45 | 459 459              |
|   27      10       2.7   2.94|   -.14   .37 |  .75  -.4   .76  -.4| 1.31 |   .34   .45 | 463 463              |
|   31      10       3.1   2.92|   -.16   .36 | 1.31   .7  1.32   .8|  .63 |   .06   .45 | 559 559              |
|   29      10       2.9   2.90|   -.18   .37 |  .62  -.8   .62  -.8| 1.44 |   .05   .50 | 542 542              |
|   29      10       2.9   2.89|   -.20   .37 | 1.20   .5  1.19   .5|  .75 |   .73   .50 | 468 468              |
|   31      10       3.1   2.89|   -.20   .36 | 2.63  2.7  2.63  2.7| -.73 |   .65   .45 | 113 113              |
|   35      10       3.5   2.89|   -.20   .36 |  .78  -.4   .77  -.4| 1.22 |   .38   .54 | 214 214              |
|   28      10       2.8   2.88|   -.21   .37 |  .37 -1.8   .39 -1.7| 1.77 |   .72   .54 | 119 119              |
|   28      10       2.8   2.88|   -.21   .37 |  .36 -1.8   .37 -1.8| 1.72 |   .56   .54 | 121 121              |
|   28      10       2.8   2.86|   -.24   .37 |  .94   .0   .94   .0| 1.01 |   .05   .48 | 445 445              |
|   29      10       2.9   2.84|   -.26   .36 |  .57 -1.0   .57 -1.0| 1.41 |   .63   .48 | 624 624              |
|   31      10       3.1   2.83|   -.27   .36 | 2.06  2.0  2.06  2.0|  .02 |   .60   .44 | 588 588              |
|   26      10       2.6   2.83|   -.28   .37 |  .38 -1.8   .39 -1.7| 1.76 |   .84   .45 | 808 808              |
|   26      10       2.6   2.83|   -.28   .37 |  .37 -1.8   .37 -1.8| 1.74 |   .32   .45 | 815 815              |
|   30      10       3.0   2.82|   -.29   .36 |  .30 -2.1   .30 -2.1| 1.76 |   .65   .45 | 593 593              |
|   28      10       2.8   2.80|   -.32   .37 | 2.08  2.0  2.05  2.0| -.22 |   .35   .46 | 603 603              |
|   28      10       2.8   2.80|   -.32   .37 | 4.33  4.5  4.33  4.5|-2.69 |   .43   .50 | 497 497              |
|   28      10       2.8   2.80|   -.32   .37 |  .94   .0   .95   .0| 1.07 |   .34   .50 | 545 545              |
|   28      10       2.8   2.80|   -.32   .37 | 1.07   .3  1.08   .3|  .92 |   .22   .50 | 548 548              |
|   28      10       2.8   2.79|   -.33   .37 |  .29 -2.2   .29 -2.2| 1.77 |   .82   .50 | 467 467              |
|   24      10       2.4   2.78|   -.34   .38 |  .50 -1.3   .51 -1.3| 1.64 |   .49   .46 | 514 514              |
|   24      10       2.4   2.78|   -.34   .38 |  .50 -1.3   .51 -1.3| 1.64 |   .49   .46 | 516 516              |
|   28      10       2.8   2.76|   -.37   .37 |  .66  -.7   .66  -.7| 1.37 |   .09   .48 | 646 646              |
|   27      10       2.7   2.75|   -.38   .37 |  .40 -1.7   .40 -1.6| 1.66 |   .56   .52 | 639 639              |
|   27      10       2.7   2.75|   -.38   .37 |  .69  -.6   .71  -.6| 1.34 |   .62   .52 | 641 641              |
|   26      10       2.6   2.74|   -.40   .37 | 1.15   .4  1.16   .5|  .78 |   .21   .44 | 290 290              |
|   25      10       2.5   2.73|   -.41   .37 |  .38 -1.8   .37 -1.8| 1.69 |   .63   .45 | 461 461              |
+---------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit     Outfit    |Estim.| Correlation |                        |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num Examinees          |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|    25      10       2.5   2.73|   -.41   .37 | 2.64  2.8  2.62  2.8|-1.01 |   .45   .45 | 803 803                |
|    29      10       2.9   2.73|   -.42   .36 | 1.35   .8  1.36   .8|  .63 |   .48   .44 | 489 489                |
|    29      10       2.9   2.73|   -.42   .36 |  .41 -1.6   .41 -1.6| 1.59 |   .83   .44 | 493 493                |
|    29      10       2.9   2.73|   -.42   .36 | 1.08   .3  1.07   .2|  .97 |   .53   .44 | 560 560                |
|    29      10       2.9   2.73|   -.42   .36 |  .64  -.8   .64  -.8| 1.33 |   .56   .44 | 594 594                |
|    24      10       2.4   2.72|   -.43   .38 |  .64  -.8   .66  -.7| 1.38 |   .53   .48 | 520 520                |
|    31      10       3.1   2.72|   -.43   .37 |  .33 -2.0   .32 -2.0| 1.78 |   .87   .65 | 167 167                |
|    31      10       3.1   2.71|   -.44   .37 |  .66  -.7   .63  -.8| 1.39 |   .67   .66 | 190 190                |
|    27      10       2.7   2.70|   -.45   .37 | 1.00   .1  1.00   .1| 1.08 |   .83   .50 | 495 495                |
|    27      10       2.7   2.70|   -.45   .37 |  .32 -2.0   .34 -1.9| 1.82 |   .83   .50 | 540 540                |
|    27      10       2.7   2.70|   -.45   .37 | 1.10   .3  1.11   .3|  .88 |  -.01   .50 | 544 544                |
|    25      10       2.5   2.70|   -.46   .37 |  .74  -.5   .73  -.5| 1.31 |   .44   .44 | 638 638                |
|    27      10       2.7   2.69|   -.46   .37 | 5.20  5.2  5.17  5.2|-3.81 |  -.41   .49 | 465 465                |
|    28      10       2.8   2.69|   -.47   .38 |  .73  -.5   .74  -.5| 1.49 |   .97   .77 | 179 179                |
|    23      10       2.3   2.67|   -.49   .38 |  .20 -2.7   .22 -2.7| 2.02 |   .86   .46 | 509 509                |
|    26      10       2.6   2.66|   -.51   .37 | 1.31   .8  1.34   .8|  .62 |   .48   .47 | 440 440                |
|    26      10       2.6   2.65|   -.52   .37 |  .85  -.2   .87  -.1| 1.22 |   .70   .52 | 643 643                |
|    27      10       2.7   2.64|   -.53   .37 | 1.13   .4  1.14   .4|  .86 |   .41   .47 | 262 262                |
|    27      10       2.7   2.64|   -.53   .37 |  .16 -3.0   .16 -3.0| 1.95 |   .87   .47 | 620 620                |
|    24      10       2.4   2.62|   -.56   .38 |  .72  -.6   .74  -.5| 1.39 |   .70   .45 | 460 460                |
|    26      10       2.6   2.60|   -.59   .37 | 1.91  1.8  1.96  1.9| -.13 |  -.19   .50 | 543 543                |
|    32      10       3.2   2.59|   -.60   .36 | 1.08   .3  1.06   .2| 1.00 |   .79   .54 | 220 220                |
|    27      10       2.7   2.58|   -.62   .38 |  .70  -.6   .70  -.6| 1.54 |   .96   .76 | 173 173                |
|    22      10       2.2   2.56|   -.64   .39 |  .57 -1.0   .59 -1.0| 1.58 |   .58   .45 | 521 521                |
|    25      10       2.5   2.54|   -.67   .38 |  .18 -2.8   .18 -2.9| 1.94 |   .74   .55 | 615 615                |
|    25      10       2.5   2.54|   -.67   .38 |  .22 -2.6   .21 -2.7| 1.90 |   .69   .55 | 618 618                |
|    27      10       2.7   2.53|   -.69   .37 | 1.07   .3  1.09   .3|  .85 |   .47   .44 | 353 353                |
|    27      10       2.7   2.53|   -.69   .37 | 1.54  1.2  1.58  1.2|  .31 |   .46   .44 | 551 551                |
|    27      10       2.7   2.53|   -.69   .37 |  .85  -.2   .87  -.1| 1.16 |   .48   .44 | 561 561                |
|    27      10       2.7   2.51|   -.70   .37 | 1.04   .2  1.05   .2|  .93 |   .27   .44 | 581 581                |
|    27      10       2.7   2.51|   -.70   .37 | 2.44  2.5  2.41  2.5| -.76 |   .49   .44 | 582 582                |
|    25      10       2.5   2.50|   -.73   .38 |  .45 -1.4   .46 -1.4| 1.66 |   .74   .49 | 496 496                |
|    27      10       2.7   2.48|   -.75   .37 |  .82  -.3   .81  -.3| 1.14 |   .26   .44 | 574 574                |
|    27      10       2.7   2.48|   -.75   .37 |  .48 -1.3   .49 -1.3| 1.61 |  -.01   .44 | 576 576                |
|    21      10       2.1   2.48|   -.75   .40 |  .40 -1.7   .46 -1.4| 1.75 |   .73   .51 | 450 450                |
|    26      10       2.6   2.47|   -.77   .39 |  .32 -2.0   .33 -1.9| 1.68 |   .87   .76 | 178 178                |
|    25      10       2.5   2.44|   -.81   .37 |  .56 -1.1   .56 -1.1| 1.53 |   .64   .47 | 538 538                |
|    27      10       2.7   2.44|   -.81   .37 | 1.01   .1  1.04   .2|  .96 |   .30   .44 | 589 589                |
|    26      10       2.6   2.43|   -.82   .37 | 1.38   .9  1.37   .9|  .63 |   .64   .44 | 554 554                |
|    28      10       2.8   2.42|   -.83   .37 |  .50 -1.3   .48 -1.4| 1.52 |   .86   .65 | 168 168                |
|    22      10       2.2   2.41|   -.85   .39 | 2.02  2.0  1.97  1.9| -.19 |   .40   .44 | 457 457                |
|    22      10       2.2   2.40|   -.86   .40 | 1.23   .6  1.19   .5|  .96 |   .80   .60 | 483 483                |
|    22      10       2.2   2.40|   -.86   .40 | 1.23   .6  1.19   .5|  .96 |   .80   .60 | 486 486                |
|    22      10       2.2   2.40|   -.86   .40 | 1.23   .6  1.19   .5|  .96 |   .80   .60 | 490 490                |
|    26      10       2.6   2.40|   -.87   .37 |  .45 -1.5   .45 -1.4| 1.63 |   .95   .45 | 109 109                |
|    24      10       2.4   2.39|   -.87   .38 |  .95   .0   .95   .0| 1.08 |   .47   .49 | 534 534                |
|    23      10       2.3   2.36|   -.91   .39 |  .59  -.9   .55 -1.1| 1.37 |   .07   .53 | 128 128                |
|    23      10       2.3   2.36|   -.91   .39 | 1.19   .5  1.18   .5|  .79 |   .62   .53 | 129 129                |
|    24      10       2.4   2.36|   -.92   .38 |  .89  -.1   .92   .0| 1.17 |   .27   .47 | 642 642                |
|    20      10       2.0   2.33|   -.97   .41 |  .54 -1.1   .54 -1.1| 1.66 |   .75   .44 | 199 199                |
|    20      10       2.0   2.33|   -.97   .41 |  .66  -.7   .62  -.9| 1.30 |   .25   .44 | 207 207                |
|    20      10       2.0   2.33|   -.97   .41 |  .40 -1.7   .40 -1.7| 1.60 |   .30   .44 | 208 208                |
|    27      10       2.7   2.32|   -.97   .37 | 1.56  1.2  1.57  1.2|  .32 |   .56   .64 | 160 160                |
|    21      10       2.1   2.30|  -1.00   .40 |  .73  -.5   .72  -.6| 1.44 |   .84   .44 | 812 812                |
|    21      10       2.1   2.30|  -1.00   .40 |  .52 -1.2   .51 -1.3| 1.62 |   .84   .44 | 832 832                |
|    21      10       2.1   2.30|  -1.00   .40 | 1.30   .8  1.29   .7|  .72 |   .53   .44 | 840 840                |
|    23      10       2.3   2.29|  -1.01   .38 | 1.18   .5  1.15   .4|  .79 |   .66   .45 | 453 453                |
|    22      10       2.2   2.26|  -1.06   .39 |  .74  -.5   .74  -.5| 1.38 |   .77   .52 | 118 118                |
|    24      10       2.4   2.25|  -1.08   .40 |  .93   .0  1.01   .1|  .87 |   .55   .75 | 492 492                |
|    25      10       2.5   2.25|  -1.08   .37 | 2.08  2.0  2.11  2.1| -.34 |   .11   .44 | 432 432                |
|    25      10       2.5   2.25|  -1.08   .37 |  .71  -.6   .70  -.6| 1.31 |   .71   .44 | 426 426                |
|    18      10       1.8   2.24|  -1.09   .43 |  .26 -2.3   .29 -2.1| 1.82 |   .85   .44 | 475 475                |
|    24      10       2.4   2.23|  -1.10   .38 |  .91   .0   .90   .0| 1.12 |   .73   .44 | 565 565                |
|    26      10       2.6   2.21|  -1.13   .38 |  .69  -.6   .80  -.3| 1.20 |   .39   .65 | 197 197                |
|    24      10       2.4   2.20|  -1.15   .38 |  .78  -.4   .79  -.4| 1.28 |   .84   .44 | 107 107                |
|    24      10       2.4   2.20|  -1.15   .38 |  .16 -3.1   .15 -3.2| 2.01 |   .73   .44 | 112 112                |
|    24      10       2.4   2.20|  -1.15   .38 |  .58 -1.0   .59 -1.0| 1.53 |   .36   .44 | 114 114                |
|    24      10       2.4   2.20|  -1.15   .38 |  .58 -1.0   .58 -1.0| 1.49 |   .62   .44 | 115 115                |
|    24      10       2.4   2.19|  -1.16   .38 |  .75  -.5   .75  -.5| 1.27 |   .42   .44 | 579 579                |
|    21      10       2.1   2.15|  -1.22   .40 |  .22 -2.6   .25 -2.5| 2.00 |   .91   .51 | 117 117                |
|    21      10       2.1   2.14|  -1.24   .40 | 2.60  2.8  2.61  2.8|-1.08 |  -.29   .45 | 444 444                |
|    22      10       2.2   2.14|  -1.24   .39 | 1.07   .3  1.03   .2|  .85 |   .30   .46 | 535 535                |
|    23      10       2.3   2.13|  -1.25   .38 | 1.05   .2  1.05   .2| 1.01 |   .59   .43 | 567 567                |
|    19      10       1.9   2.12|  -1.26   .42 |  .66  -.7   .72  -.5| 1.23 |   .02   .41 | 474 474                |
|    18      10       1.8   2.09|  -1.32   .43 |  .60  -.9   .60  -.9| 1.12 |  -.30   .42 | 206 206                |
|    22      10       2.2   2.03|  -1.40   .39 | 1.36   .9  1.40   .9|  .52 |   .25   .43 | 477 477                |
+------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                     |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Num Examinees       |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|    22      10       2.2   2.03|  -1.40   .39 | 1.90  1.8  1.78  1.6| -.10 |   .49   .43 | 569 569             |
|    24      10       2.4   2.03|  -1.41   .39 | 1.00   .1   .93   .0| 1.02 |   .76   .63 | 165 165             |
|    24      10       2.4   2.03|  -1.41   .39 | 1.48  1.1  1.56  1.2|  .35 |   .19   .63 | 166 166             |
|    24      10       2.4   2.03|  -1.41   .39 | 1.21   .6  1.13   .4|  .73 |   .66   .63 | 709 709             |
|    22      10       2.2   2.02|  -1.41   .42 | 1.59  1.2  1.67  1.3| -.04 |  -.24   .74 | 181 181             |
|    22      10       2.2   2.01|  -1.44   .39 |  .60  -.9   .57 -1.1| 1.44 |   .56   .44 | 100 100             |
|    19      10       1.9   1.99|  -1.47   .42 | 1.58  1.2  1.34   .8|  .64 |   .78   .47 | 601 601             |
|    28      15       1.9   1.99|  -1.47   .35 |  .70  -.7   .66  -.9| 1.40 |   .75   .51 | 452 452             |
|    17      10       1.7   1.96|  -1.51   .45 |  .31 -2.0   .39 -1.6| 1.54 |   .53   .41 | 201 201             |
|    18      10       1.8   1.96|  -1.51   .43 |  .21 -2.7   .24 -2.4| 1.75 |   .64   .41 | 800 800             |
|    22      10       2.2   1.96|  -1.52   .39 | 1.09   .3  1.11   .4|  .94 |   .53   .43 | 425 425             |
|    23      10       2.3   1.93|  -1.56   .39 | 1.25   .6  1.21   .6|  .71 |   .58   .63 | 163 163             |
|    23      10       2.3   1.92|  -1.58   .39 | 1.40   .9  1.27   .7|  .60 |   .68   .63 | 188 188             |
|    23      10       2.3   1.92|  -1.58   .39 |  .63  -.8   .60  -.9| 1.25 |   .57   .63 | 194 194             |
|    21      10       2.1   1.91|  -1.59   .40 |  .59 -1.0   .57 -1.0| 1.52 |   .76   .43 | 106 106             |
|    19      10       1.9   1.88|  -1.64   .42 |  .79  -.3   .90   .0|  .94 |  -.04   .50 | 813 813             |
|    19      10       1.9   1.88|  -1.64   .42 | 1.54  1.2  1.44  1.0|  .68 |   .65   .50 | 842 842             |
|    21      10       2.1   1.86|  -1.67   .40 |  .37 -1.9   .37 -1.9| 1.73 |   .47   .42 | 420 420             |
|    17      10       1.7   1.84|  -1.70   .45 |  .38 -1.7   .40 -1.6| 1.68 |   .84   .40 | 837 837             |
|    20      10       2.0   1.84|  -1.71   .40 |  .42 -1.6   .43 -1.6| 1.71 |   .62   .42 | 436 436             |
|    20      10       2.0   1.82|  -1.75   .41 |  .59 -1.0   .60  -.9| 1.60 |   .70   .42 | 111 111             |
|    20      10       2.0   1.77|  -1.83   .40 | 1.20   .5  1.15   .4|  .84 |   .58   .42 | 421 421             |
|    18      10       1.8   1.76|  -1.85   .43 | 2.51  2.6  3.08  3.2| -.97 |  -.46   .45 | 466 466             |
|    19      10       1.9   1.75|  -1.88   .42 | 1.70  1.5  1.88  1.8|  .23 |   .32   .41 | 564 564             |
|    21      10       2.1   1.73|  -1.90   .41 |  .96   .0   .87  -.1| 1.12 |   .79   .62 | 187 187             |
|    19      10       1.9   1.71|  -1.94   .42 | 2.00  1.9  1.74  1.5|  .02 |   .56   .41 | 577 577             |
|    15      10       1.5   1.71|  -1.95   .51 |  .61  -.7   .71  -.4| 1.15 |   .26   .37 | 511 511             |
|    15      10       1.5   1.71|  -1.95   .51 | 1.76  1.4  2.26  1.9|  .40 |  -.01   .37 | 515 515             |
|    19      10       1.9   1.70|  -1.97   .45 |  .46 -1.2   .51  -.9| 1.36 |   .78   .70 | 172 172             |
|    19      10       1.9   1.70|  -1.97   .45 |  .40 -1.4   .36 -1.4| 1.59 |   .89   .70 | 177 177             |
|    20      10       2.0   1.66|  -2.05   .42 |  .75  -.4   .90   .0| 1.04 |   .35   .60 | 161 161             |
|    21      10       2.1   1.61|  -2.14   .40 | 1.24   .6  1.21   .6|  .93 |   .72   .51 | 192 192             |
|    16      10       1.6   1.60|  -2.16   .48 |  .40 -1.5   .40 -1.4| 1.61 |   .86   .45 | 120 120             |
|    18      10       1.8   1.59|  -2.18   .46 |  .90   .0  1.32   .7|  .47 |   .21   .69 | 184 184             |
|    18      10       1.8   1.59|  -2.18   .46 |  .40 -1.4   .37 -1.2| 1.67 |   .89   .69 | 491 491             |
|    19      10       1.9   1.56|  -2.25   .43 |  .64  -.7   .69  -.5| 1.20 |   .65   .60 | 185 185             |
|    19      10       1.9   1.56|  -2.25   .43 |  .71  -.5   .69  -.5| 1.34 |   .67   .60 | 189 189             |
|    19      10       1.9   1.56|  -2.25   .43 | 1.21   .5  1.03   .2| 1.08 |   .85   .60 | 196 196             |
|    17      10       1.7   1.50|  -2.40   .48 |  .42 -1.3   .60  -.5| 1.16 |   .74   .66 | 175 175             |
|    13      10       1.3   1.43|  -2.57   .62 | 1.76  1.2  1.85  1.2|  .62 |  -.09   .31 | 510 510             |
|    17      10       1.7   1.40|  -2.65   .46 | 1.47  1.0  1.92  1.5|  .27 |   .08   .56 | 191 191             |
|    17      10       1.7   1.40|  -2.65   .46 |  .78  -.3  1.08   .3|  .90 |   .33   .56 | 195 195             |
|    18      10       1.8   1.40|  -2.66   .44 |  .81  -.3   .85  -.1| 1.02 |   .15   .48 | 223 223             |
|    16      10       1.6   1.34|  -2.86   .49 |  .49 -1.1   .61  -.6| 1.20 |   .50   .54 | 164 164             |
|    16      10       1.6   1.33|  -2.88   .49 |  .79  -.3   .77  -.2| 1.27 |   .67   .54 | 198 198             |
|    14      10       1.4   1.29|  -3.02   .55 |  .89   .0   .95   .1|  .90 |  -.09   .33 | 102 102             |
|    12      10       1.2   1.29|  -3.04   .74 | 1.03   .2  1.14   .4|  .88 |  -.10   .26 | 209 209             |
|    12      10       1.2   1.29|  -3.04   .74 | 1.09   .3  1.59   .8|  .78 |  -.34   .26 | 210 210             |
|    16      10       1.6   1.28|  -3.07   .48 |  .84  -.1   .89   .0| 1.15 |   .45   .44 | 213 213             |
|    14      10       1.4   1.21|  -3.40   .56 | 1.30   .6  2.00  1.3|  .32 |  -.35   .47 | 169 169             |
|    14      10       1.4   1.20|  -3.43   .57 |  .44 -1.1   .56  -.5| 1.27 |   .63   .48 | 193 193             |
|    12      10       1.2   1.19|  -3.49   .75 | 1.20   .4  2.09  1.2|  .65 |  -.44   .30 | 122 122             |
|    13      10       1.3   1.14|  -3.78   .63 | 1.15   .4  1.18   .4|  .97 |   .42   .43 | 186 186
|    10      10       1.0   1.02|( -5.80  1.84)|Minimum              |      |   .00   .00 | 101 101             |
|    10      10       1.0   1.01|( -6.12  1.84)|Minimum              |      |   .00   .00 | 180 180             |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|    30.7    10.0     3.1   3.07|   -.03   .40 |  .99  -.3  1.00  -.2|      |   .46       | Mean (Count: 358)   |
|     8.7      .5      .9    .91|   1.37   .12 |  .76  1.6   .76  1.6|      |   .34       | S.D. (Population)   |
|     8.7      .5      .9    .91|   1.37   .12 |  .76  1.6   .76  1.6|      |   .34       | S.D. (Sample)       |
+---------------------------------------------------------------------------------------------------------------+
   With extremes, Model, Populn: RMSE .41  Adj (True) S.D. 1.31  Separation 3.17  Strata 4.56  Reliability .91
   With extremes, Model, Sample: RMSE .41  Adj (True) S.D. 1.31  Separation 3.17  Strata 4.56  Reliability .91
Without extremes, Model, Populn: RMSE .39  Adj (True) S.D. 1.24  Separation 3.18  Strata 4.57  Reliability .91
Without extremes, Model, Sample: RMSE .39  Adj (True) S.D. 1.24  Separation 3.18  Strata 4.58  Reliability .91
With extremes, Model, Fixed (all same) chi-square: 3262.9  d.f.: 357  significance (probability): .00
With extremes, Model,  Random (normal) chi-square: 304.0  d.f.: 356  significance (probability): .98
-----------------------------------------------------------------------------------------------------------------
```