2005

A dynamic simulation assessment of english as a second language students' academic readiness

S. "Sha" G. Balizet

University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the American Studies Commons

Scholar Commons Citation


http://scholarcommons.usf.edu/etd/2970

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
A Dynamic Simulation Assessment of
English as a Second Language Students' Academic Readiness:
An Initial Validation of a University Admissions Tool

by

S. ("Sha") G. Balizet

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of World Language Education
College of Arts and Sciences
and
Department of Secondary Education
College of Education
University of South Florida

Co-Major Professor: Jeffra Flaitz, Ph.D.
Co-Major Professor: Wei Zhu, Ph.D.
John Ferron, Ph.D.
Carine Feyten, Ph.D.

Date of Approval:
May 27, 2005
Keywords: IEP, EAP, testing, content-based, CALP, academic competence, academic achievement

© Copyright 2005, S. G. Balizet
DEDICATION

I dedicate this dissertation
to my mother, Carol,
with appreciation for her endless encouragement and prayers, and
to my father, Thomas,
with gratitude for his boundless support and confidence.
Thank you.
I could not have done this without you.

I also dedicate this work
to my fellow students
in the Second Language Acquisition/Instructional Technology program,
hoping you find joy and satisfaction in your dissertation research as did I.
May you, too, benefit from this word of comfort
by St. Teresa of Avila (1515-1582)
as you undergo the sojourn called the dissertation.

Nada te turbe. Let nothing disturb you.
Nada te espante. Let nothing frighten you.
Todo se pasa. All things are passing.
Dios no se muda. God only is changeless.
La paciencia todo lo alcanza. Patience gains all things.
Quien a Dios tiene, nada le falta. Who has God lacks nothing.
Solo Dios basta. God alone suffices.
ACKNOWLEDGEMENTS

Although this work bears my name, it would not have been possible without the kind assistance of many wonderful colleagues. Although I cannot possibly thank everyone, I am pleased to recognize the following individuals.

Very special thanks are due to my mother, Carol, and my father, Thomas. Without your support and encouragement, I could not have accomplished this. May it reflect well upon you.

Thank you to my splendid committee for the priceless experience of working with you all. I benefited by your contribution to my research, and I am grateful.

To Dr. Jeffra Flaitz I extend my deepest gratitude. Thank you for inspiring me, for entrusting me with challenges and your confidence. I am honored to have conducted this validation study with you, whom I respect so highly. I appreciate your splendid contributions to the dissertation and to my professional growth. Thank you for drawing out the best in me.

To Dr. Wei Zhu I express my greatest thanksgiving. How you have engaged me by your insight and wisdom. Thank you for your confidence in me and your unflagging encouragement. I am grateful to have investigated the CLEAR with you. I appreciate your high expectations and your insightful guidance. Thank you for leading me to excel.

To Dr. John Ferron, I am pleased to state my appreciation. Years ago, you asked me, "How will you respond to those who ask if the CLEAR is just a covert intelligence test?" Your questions, comments, and insight have shaped this investigation. Along the way, you have guided me into a higher level of professionalism. Thank you.

To Dr. Carine Feyten, I am honored to convey my tribute of thanks. You are, as your very name suggests, caring. For many years I have been privileged by your confidence in me and your stalwart encouragement for me to work to the extent of my potential. For all that, and for your special insights into this dissertation, I thank you.

To Dr. Robert Dedrick, I am grateful for your helping foster in me a love of research. Thank you for your kind commitment to me and my investigation.

To my sister, Dr. Cynthia Parshall, I express my deep obligation. When we were young, you piqued my interest in reading. Years later you evoked in me increasing enthusiasm for measurement, research, and writing. You continued to encourage me through these many years of graduate study. In these ways, and so many more, you are a great blessing to me. Thank you.

My acknowledgements go also to the members of the CLEAR test development team, Dr. Jeffra Flaitz and Dr. Wei Zhu along with Debbie Mitchell, Amna Mohamed, and Barbara Smith-Palinkas. What an outstanding team! Thanks for the numberless hours, and the many meetings, and work assignments you gave with such good cheer and fine quality. The CLEAR reflects brightly on you all.

My sincere thanks to colleagues at the language institute herein called the "English for Academic Purposes Institute." Your assistance over the years has helped bring the CLEAR to life. Your interest and encouragement in the CLEAR spurred the project to greater achievement.
Thanks in particular are due to Patti Baker, for her countless favors, and to Barbara Smith-Palinkas and Ann Morris, for logistical brilliance.

I am very grateful to the second language institute who participated in this validation study. My warm thanks to Myrna Creasman and the splendid teachers and staffers of the institute. I appreciate your gracious welcome and collegial hospitality. Thank you.

I am humbled by the kind participation of all the CLEAR examinees who contributed to this validation investigation and to the three pilot studies. Thank you for your help.

I truly appreciate the sage advice offered by the experts whom I consulted in this study. The two psychologists and the psychometrician contributed enormously to the CLEAR and my understanding. I am grateful to you, and regret that I must not acknowledge you by name.

To colleagues in the World Language Education department, I extend my thanks for your encouragement and support over the years. In particular, I am pleased to recognize Karen Kime Wilson, Tom Rodgers, Norma Rodriguez, and Osiris Albrecht. I likewise thank my colleagues in the College of Education, particularly Judith Thomas, Marilyn Washington, and Ursula White. Thank you for the many times you have helped me.

I am pleased to acknowledge the faculty of my doctoral program in Second Language Acquisition and Instructional Technology. My thanks to the authors of this program, Dr. Cole, Dr. Flaitz, and Dr. Feyten, and appreciation to all the faculty for leading me in scholarly growth.

It gives me particular joy to recognize my good friends, the students in the doctoral program. My warm appreciation to Michelle Macy, for her collegiality and kind commiseration. My thanks to Annmarie Zoran and Iona Sarieva for their friendship and camaraderie. The members of my cohort are particularly special to me: Martha Casteneda, Ruth Roux-Rodriguez, Sabine Siekmann, Phil Smith, Eun-Wook Park, and most especially my dear friend, Frank Chaich. How lovely that we traveled together in this journey. Thank you.

Finally, it is an honor to recognize Monsignor John Vargas and of my prayer companions at the Church of the Incarnation. Thank you for your gracious support. The Lord knows how greatly you helped me.
# TABLE OF CONTENTS

LIST OF TABLES....................................................................................................................v

LIST OF FIGURES....................................................................................................................vii

ABSTRACT....................................................................................................................................viii

INTRODUCTION...........................................................................................................................1

  Issues Concerning the TOEFL.................................................................................................20

  History and Description of the CLEAR Dynamic Simulation Assessment..........................16

  The Test Purpose of the CLEAR ..........................................................................................19

  The Construct of Academic Readiness (AR)..........................................................................24

  CLEAR Dynamic Simulation Assessment Description.........................................................28

Rationale........................................................................................................................................31

Statement of the Problem...........................................................................................................33

Purpose of the Study....................................................................................................................34

Research Questions...................................................................................................................36

Delimitations and Limitations....................................................................................................39

Definitions....................................................................................................................................41

LITERATURE REVIEW.............................................................................................................45

Introduction..................................................................................................................................45

Assumptions Underlying Language Tests: The Know-Use Framework...............................46

Modeling Language Proficiency...............................................................................................52

  Oller, language, and the ability measured by intelligence tests............................................55
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Linguistic Threshold Hypothesis of L2 Reading</td>
<td>64</td>
</tr>
<tr>
<td>The Construct of Academic Readiness (AR)</td>
<td>68</td>
</tr>
<tr>
<td>Dynamic Simulation Assessment</td>
<td>70</td>
</tr>
<tr>
<td>Academic Skills: L2 Students' Reading and Writing</td>
<td>77</td>
</tr>
<tr>
<td>Academic Reading</td>
<td>79</td>
</tr>
<tr>
<td>The Nature of Reading</td>
<td>79</td>
</tr>
<tr>
<td>The Component Skills Approach to Explaining Reading</td>
<td>80</td>
</tr>
<tr>
<td>The Metaphoric Approach to Explaining Reading</td>
<td>82</td>
</tr>
<tr>
<td>Human Factors in Reading: Good and Poor Readers</td>
<td>84</td>
</tr>
<tr>
<td>Factors in Reading: Vocabulary</td>
<td>86</td>
</tr>
<tr>
<td>Factors in Reading: Content/background Knowledge</td>
<td>88</td>
</tr>
<tr>
<td>Factors in Reading: the Threshold of Linguistic Proficiency</td>
<td>91</td>
</tr>
<tr>
<td>Factors in Reading: Reader Purpose</td>
<td>93</td>
</tr>
<tr>
<td>Reading-to-Write</td>
<td>94</td>
</tr>
<tr>
<td>Academic Writing</td>
<td>114</td>
</tr>
<tr>
<td>Characteristics of discipline-based academic writing</td>
<td>99</td>
</tr>
<tr>
<td>Knowledge in Discipline-based Academic Writing</td>
<td>99</td>
</tr>
<tr>
<td>Writing for a Discourse Community</td>
<td>101</td>
</tr>
<tr>
<td>Discussion</td>
<td>102</td>
</tr>
<tr>
<td>METHODOLOGY</td>
<td>105</td>
</tr>
<tr>
<td>Introduction</td>
<td>105</td>
</tr>
<tr>
<td>Participants</td>
<td>108</td>
</tr>
<tr>
<td>Comparison of Participants' Contexts</td>
<td>109</td>
</tr>
<tr>
<td>Participant Profile</td>
<td>110</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>113</td>
</tr>
<tr>
<td>General Description of the CLEAR</td>
<td>113</td>
</tr>
<tr>
<td>CLEAR Dynamic Simulation Knowledge Subtest</td>
<td>116</td>
</tr>
<tr>
<td>CLEAR Essay Subtest</td>
<td>119</td>
</tr>
</tbody>
</table>
Essay Scoring by Expert Judges.................................122
CLEAR Essay Scoring Tool.................................123
Materials...........................................................................................................128
  Description of Text Used in the CLEAR.................................128
  Description of Videotaped Lecture Used in the CLEAR..........129
Research Instruments: Examinee Feedback.................................131
Research Instruments: Rater Feedback.................................131
Research Instruments: Teacher Ratings.................................132
Procedures..................................................................................................................132
Issues Concerning Test Duration and Timing.................................133
Role of Specialists in Validating the CLEAR.................................135
Examinee Motivation in Taking the CLEAR.................................139
Analyses..................................................................................................................139
Reliability......140
Validity...............................................................................................................141
Research Questions and Analyses..................................................142
RESULTS........................................................................................................147
Overview of Research Questions..................................................147
Preliminary Analyses..................................................148
Research Question 1: What is the consistency of the scores derived from
the CLEAR?......................................................................................................161
Research Question 2: What is the validity of inferences based on scores
derived from the CLEAR?........................................................................167
  Content-related Evidence for Claims of Validity in the
  Knowledge Subtest ..................................................................................168
  Content-related Evidence for Claims of Validity in the
  Essay Subtest........................................................................................171
  Concurrent Measures Yielding Evidence for Claims of Validity......173
LIST OF TABLES

Table 1: Contrasting Orientations to Perceiving and Testing Language..........................48
Table 2: Academic Reading and Writing Skills (Flaitz & Zhu, 2001)...............................78
Table 3: Proficiency Levels According to MTELP-CELT Scores at Site #1...............112
Table 4: CLEAR Testing Events..................................................................................133
Table 5: Ratings Assigned During Initial Practice Session.................................138
Table 6: Recapitulation of CLEAR Analyses.................................................................148
Table 7: Mean Student CLEAR Scores by Test Site....................................................149
Table 8: Mean CLEAR Scores by ESL Proficiency Level............................................150
Table 9: Examinees' First Language...........................................................................152
Table 10: Mean CLEAR Scores by Gender.................................................................153
Table 11: Knowledge Subtest Item Analysis...............................................................156
Table 12: Discrimination Values of Knowledge Subtest Items..................................159
Table 13: CLEAR Knowledge Test Scores Descriptive Statistics................................161
Table 14: Knowledge Subtest Reliability Improvement by Removing Poor Items.....163
Table 15: Essay Rating Correlation in Range-finding Session.................................165
Table 16: Essay Assessment Rater Statistics...............................................................166
Table 17: Rater Mean Score t-test..............................................................................167
Table 18: Essay Assessment Inter-rater Score Correlations........................................168
Table 19: Correlations of CLEAR Knowledge Subtest and Teacher Ratings.............173
Table 20: Correlations of CLEAR Essay Subtest and Teacher Ratings......................175
Table 21: Correlations of CLEAR Knowledge Subtest and External Measures..........177
Table 22: Correlations of CLEAR Essay Subtest and External Measures...............178
Table 23: Intercorrelation Between CLEAR Subtests........................................178
Table 24: CLEAR Intercorrelations.....................................................................179
Table 25: Feedback Item 1a: Examinee Ratings of Difficulty of Video Lecture......181
Table 26: Feedback Item 1b: Examinee Ratings of Difficulty of Reading.............183
Table 27: Feedback Item 1c: Examinee Ratings of Difficulty of Knowledge Test.....183
Table 28: Feedback Item 1d: Examinee Ratings of Difficulty of Essay Assessment.................................................................184
Table 29: Feedback Item 1e: Examinee Ratings of Difficulty Overall...............185
LIST OF FIGURES

Figure 1: Sample Test Specifications from an Early Version of the CLEAR..............18
Figure 2: Know-Use Continuum..............................................................................47
Figure 3: Series Completion Task in Chinese.............................................................59
Figure 4: Elements Hypothesized to Contribute to Academic Readiness...............69
Figure 5: CLEAR Essay Scoring Tool......................................................................125
Figure 6: Results from Feedback Item 3, "From Taking the CLEAR, I
   Learned . . . "........................................................................................................186
Figure 7: Results from Feedback Item 4, "I Did Best on the ___ Section,
   Because . . . ".......................................................................................................187
Figure 8: Results from Feedback Item 5, University Admissions
   Test Preference....................................................................................................189
Figure 9: Results from Feedback Item 6, "Taking the CLEAR Over a 48-hour
   Period . . . "..........................................................................................................191
Figure 10: Results from Feedback Item 7, "If I Could Change
   One Thing . . . ".................................................................................................192
Figure 11: Results from Feedback Item 8, "The Test Gave Me a Feel
   for What I Will Have to do in the University.".....................................................194
Figure 12: Results from Feedback Item 9, "I Would Recommend Taking the
   CLEAR to a Friend.".........................................................................................196
Figure 13: Results from Feedback Item 2, "Comment on Any Section
   of the Test."........................................................................................................198
Figure 14: Results from Feedback Item 10, "Other comments"…………………200
Figure 15: Graphical Representation of CLEAR Intercorrelations………………216
ABSTRACT

This study investigated the measurement quality of the Content Learning Experience: Academic Readiness (CLEAR) test, a new measure for university admissions decisions regarding English as a Second Language (ESL) applicants. The CLEAR test measures ability through dynamic simulation: learning opportunities are followed by testing how well students learned the academic content, all modeled on university instructional experiences.

Measured by the CLEAR is academic readiness (AR), the direct, present evidence of ability to learn academic content via the second language as demonstrated during the dynamic simulation. AR is hypothesized to comprise above-threshold academic language proficiency, personal characteristics, topical knowledge, academic skills, and academic auxiliaries (motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies).

The participants were 36 international adults, studying pre-university academic English at intensive institutes in Florida who volunteered to take the CLEAR during the summer of 2004. Data were collected via the CLEAR multiple-choice knowledge test and essay test, teacher ratings, examinee feedback, and external measures.

Results showed the CLEAR knowledge test functions well at the item level although overall scores are only moderately consistent. The essay scoring consistency was satisfactory, perhaps partly due to the purpose-built scoring tool. Good support for content-related validity claims was found for the dynamic simulation overall, for the stimulus materials, for the knowledge test items, for the essay prompt, and for the essay scoring tool. The concurrent measure of teacher ratings correlated with the knowledge
test, but not with the content-based essay. Concerning construct-related claims of validity, support was evinced through the literature review as well as through inter-subtest correlation. External measures suggested some discriminant evidentiary support. Examinees perceived that the CLEAR closely resembled the target environment, they judged the CLEAR quality to be a key feature, and they would recommend the CLEAR to a friend for the growth experience. In conclusion, the CLEAR dynamic simulation assessment appears to offer potential to university admissions testing for non-native English speakers, particularly regarding the utility of the learning potential measurement, the essay scoring tool, and the examinee perceptions of the test.
CHAPTER 1
INTRODUCTION

International students have come to the United States for university study in increasing numbers. In the 2002-2003 academic year, over 586,000 international students attended college or university in this country, representing a 74% increase over the past 20 years (Institute of International Education, 2003).

A key tool for university study in the U.S. is advanced language proficiency, so the student is able to learn academic content delivered in English, the language of instruction. Many international students enroll in Intensive English Programs (IEPs). Although they often have some experience with formal English language study in their home country, these international students require more English language improvement or language instruction for their specific need, English for academic purposes (EAP). Indeed, 51,179 students enrolled in IEPs in 2002 (Institute of International Education, 2003). In the IEP, students learn English vocabulary and syntax, as well as skills in listening, speaking, reading, and writing (Turner, 2004). IEPs focus instruction on building students' second language (L2) academic language proficiency skills and
academic practices (Garner & Borg, 2005, p. 119). Such includes reading for meaning and summarizing; distinguishing main and secondary ideas in reading different text types; understanding and expressing complex, abstract ideas; organizing ideas and synthesizing information using appropriate rhetorical patterns. In the target environment, these competencies may be required orally or in writing; the abilities are needed in class at times, perhaps for homework, or during a test.

Advanced language proficiency is a university pre-requisite for admitting L2 students in North America (as will be discussed below) yet, "the relationship between language proficiency and academic success is complex and unclear" (Graham, 1987, p. 516). Requirements of university students vary considerably across different academic fields of study, and so do language skill demands (Elder, Erlam, & von Randow, 2002; Garner & Borg, 2005; Turner, 2004). Language proficiency is necessary, for students are "unlikely to cope with academic study" if their L2 proficiency is below threshold level (Elder, Erlam, & von Randow, 2002, p.1). Nevertheless, language is not the only skill required for academic success, even in linguistically intensive disciplines: Language proficiency explains perhaps 10% of GPA variance, but other factors account for 90% of the GPA (p. 1).

Evidently, other abilities play a compensatory role for the university students who are non-native speakers of English (NNSEs). What might these abilities be? In addition to language ability, other factors have been identified in the L2 research literature
(Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Huong, 2001). These four factors are topical knowledge, affective schemata, metacognitive strategies, and personal characteristics. Topical knowledge, an individual's background knowledge can certainly benefit learning when the L2 is sufficiently advanced so the student can comprehend L2-mediated instruction (Clapham, 1996). Affective schemata are exemplified by motivation, determination, and attitude toward the learning situation (cf. Elder, Erlam, & von Randow, 2002, p. 1). Metacognitive strategies might encompass the student's goal-setting, planning, and pacing study. Personal characteristics includes such attributes as age, sex, linguistic background, and educational experience.

A different model has been recently investigated by educational psychologists Elliot and DiPerna and their colleagues (S. Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, S. Elliott, & DiPerna, 2002). Working with native-speaking students (learning disabled, "at risk," and general education students), these scholars classify factors contributing to academic achievement into two groups, skills or behaviors. "Academic skills" is the first ability area, comprising competency in reading/language arts, mathematics, and critical thinking. The second ability group they call "academic enablers," which skills include motivation, study skills, engagement, and interpersonal skills. Most of these underlying academic skills and enabling behaviors contribute to academic outcomes at the university level. Other contributing factors are the dimensions called "work drive" and "emotional stability" (Ridgell & Lounsbury, 2004; cf. J. Turner, 2004). Work drive concerns
motivation to be productive, to complete projects, and realize success (Lounsbury, Sundstrom, Loveland, & Gibson, 2003, in Ridgell & Lounsbury, 2004, p. 608).

Informed by this literature from scholars in second language acquisition (SLA) and general educational psychology is the construct under examination in the Content Learning Experience: Academic Readiness ("CLEAR") dynamic simulation assessment. Academic readiness ("AR") is this construct. Quite likely AR is comprised of a constellation of abilities, auxiliaries, and resources, varying in complexity according to the complexity of the academic task (cf. Bachman, 2002; Garner & Borg, 2005). Certainly, above-threshold language proficiency is one such foundational quality. The threshold level of language proficiency is a necessary condition, but is not sufficient for learning via the L2. Other factors in AR are personal characteristics and topical knowledge. Important contributions to AR are supporting academic traits and skills. Termed "academic auxiliaries" herein, these enablers are the dimensions of motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies; combining these together reflects research on academic competence and L2 learners (Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Garner & Borg, 2005; Huong, 2001) and investigations into academic achievement among native-English-speaking university students (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002; Ridgell & Lounsbury, 2004).
AR is defined as the direct, present evidence of ability to learn academic content via the L2 as demonstrated during the dynamic simulation, learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions similar to the same. As a test of AR, the CLEAR does not seek to tap causal factors, nor elicit past learning achievement. The CLEAR seeks direct evidence. The dynamic simulation gives examinees a university learning experience in microcosm to see how well they can acquire new content knowledge. Content knowledge is "arguably the primary attribute assessed in many college courses" (Elliot & DiPerna, 2002, p. 10). Furthermore, a direct test offers powerful evidence, state Haywood and Tzuriel (2002), for "the best test of any performance is a sample of that performance itself (e.g., Cronbach, 1970; Freeman, 1950); therefore, assessment of learning abilities can be accomplished effectively with the use of learning tasks, especially those involving teaching--a condition that characterizes school learning" (p. 41).

Notwithstanding the importance of other abilities in academic success, university admissions maintain linguistic criteria for admissions. Setting language qualifications is generally a matter of admissions policy in U.S. universities. For applicants whose first language is not English, university admission usually includes a demonstration of the applicant's competency in English. This often entails taking the Test of English as a Foreign Language (TOEFL, Educational Testing Service [ETS]); these applicants might have no other option to demonstrate English language proficiency. Applicants sometimes
take this test repeatedly until they achieve the institution or program minimum score.

(The common minimum score for undergraduates is 550 on the paper-based test or 213 on the computer-based test; for graduate students, the typical minimum score is pegged at 567 on the paper-and-pencil test, 227 on the computerized version [Templer, 2004, para. 19].) The TOEFL is the pre-eminent English language proficiency test for university admissions purposes in the U.S. Such great acceptance has the TOEFL with college admissions officers that alternative measures have a difficult time gaining acceptance (Des Brisay, 1994; Roemer, 2002).

*Issues Concerning the TOEFL*

A new alternative to the TOEFL is the CLEAR test. The CLEAR tests through dynamic simulation, an innovative approach to assessing ESL students for their readiness to enter a university. Since validating this alternative to the TOEFL is the purpose of the study, some discussion of the TOEFL is warranted in order to provide context for the new assessment.

First appearing in the early 1960s, the TOEFL now is employed globally. Over 42,000 institutes and agencies use this test in more than 80 countries (ETS, 2000). The TOEFL is notable as probably the most highly studied language test (Barnwell, 1996, p. 125). TOEFL scores are extremely reliable: the most recently published data, for 1998-1999, report an overall reliability coefficient of .95 (ETS, 2000, p. 28).
Despite extremely high reliability, the TOEFL is not without flaws (Canadian Psychological Association, 1999; Raimes, 1990; Roemer, 2002; Stansfield, 1986). One crucial concern is the validity of the measurement approach taken in the TOEFL, for troubling discrepancies exist between the TOEFL and the real world conditions for university learning and testing. More specifically, the TOEFL is a static test, measuring what examinees have learned about academic English structure, phonology, and vocabulary. Indeed, about 80% of the TOEFL is comprised of selected-response items (Roemer, 2002); the selected-response format is used in half of the Internet-based TOEFL (iBT) sections (E. Tyson, personal communication, July 14, 2005). Using passive recognition items is not unreasonable for the skills of listening and reading: receptive language skills can only be assessed indirectly. Less acceptable, however, is using indirect measurement for productive language skills. Speaking and writing should be measured directly, advises language testing expert Hughes (1989, p. 16), especially in a proficiency test (emphasis mine). Proficiency as measured in the existing TOEFL is, therefore, largely inferred through responses on multiple-choice vocabulary and structure items. Writing is now required, in the form of a 30-minute timed essay. Actual writing performance was optional until very recently, with the advent of the computer-based version of the test in 1998 (ETS, 2000). Moreover, the productive skill of speaking is not considered at all in the present TOEFL (although the revised, Internet delivered TOEFL, to be launched in phases starting in late 2005, will require speaking and integrated skill
use). Thus, on the surface level, the traditional TOEFL approach to measuring language is problematical. More importantly, the TOEFL does not seek to measure AR, the construct of interest to the CLEAR developers.

**Viewing language knowledge versus language use: a CLEAR metaphor**

Beneath this surface manifestation lies the model of language proficiency upon which the TOEFL purports to be built. TOEFL research asserts (Stansfield, 1986) that the test is intended to measure “communicative competence” (Canale, 1983; Canale & Swain, 1980; Stansfield, 1986). The model of communicative competence is language very broadly conceived: not only grammatical competence, but also strategic competence, and the elements of sociocultural and discourse competence comprising sociolinguistic competence. Each of these aspects is further divided into smaller elements, then further subdivided: language is broken up into ever-smaller bits in this taxonomic approach.

The TOEFL does not measure strategic competence, nor indeed much in terms of sociolinguistic competence, but concentrates on grammatical competence. The TOEFL could thus be deemed an incomplete realization of the communicative competence model (cf. Stansfield, 1984). Alternately, one might critique the model itself: Despite attempts of the communicative competence model to identify comprehensively the components of language, this model falls short of capturing the fluid experience of language in use, and
in a specific context (cf. Garner & Borg, 2005). Indeed, the problem might reside in perceiving language as an object of knowledge rather than a tool for use.

The underlying view of language holds important consequences for the type of test that will be constructed. The CLEAR is built on a different view of language from that underlying the TOEFL: the CLEAR treats language as a tool to use, whereas the TOEFL seems based on language as an object of knowledge. Some elaboration may prove helpful here. For example, from a "Know" orientation, a test builder would likely follow a norm-referenced measurement approach, constructing a conventional static test. "Know"-type tests are typically derived from theory; they may suffer from surface irrelevance in attempting to tap latent abilities. The TOEFL reflects this orientation. Alternatively, viewing language as a tool to "Use" would likely lead to distinctly different outcomes, quite possibly performance assessment. The functionally oriented tester would more likely employ criterion-referenced measurement. Those criteria are normally derived, not from abstract theory, but from empirical evidence such as job or task analysis, generating crucial data to build a test adequate in domain representation and criterion sampling. These performance assessments typically reflect some resemblance to the target environment or activity: a driver's license test, for example, requires performance of target criterion activity. The CLEAR flows from this orientation. This topic will be further considered in Chapter 2.
The CLEAR development team (comprised of two professors of applied linguistics, two ESL professionals, and this researcher) found a major problem resulting from the analytical, taxonomic representation of language: language-related elements are not of equal importance. Indeed, some elements might not even be needed in certain occasions: context is crucial in assessing language in use. For example, essays deficient in targetlike language seem not to dismay discipline-based faculty as much as do essays deficient in subject-matter knowledge covered in the lectures (Weigle, 2002, p. 189; cf. Leki & Carson, 1997; Horowitz, 1986b; Johns, 1991; Santos, 1998; Schleppegrell, 2002; Weigle & Nelson, 2001). Here becomes evident the importance of context (cf. Garner & Borg, 2005) for the focus is upon learning academic content in the subject disciplines, and those instructors will look for such evidence in scoring their students' essays; the focus of learning in the L2 classroom is language-related elements, and these instructors seek evidence for learning such elements in their learners' compositions. The analytical, taxonomic approach of the TOEFL entails a close scrutiny of numerous small elements; it assumes that every element has been identified, that no element is omitted, and that elements are represented in appropriate proportions. Indeed, Turner (2004) charges that the TOEFL (and its British-based competitor, the IELTS) "underspecifies the complexity of language issues in the academic context" (p.97). The CLEAR does not seek to construct a replacement model of language proficiency. Academic readiness is the focus of interest for the CLEAR, which seeks to elicit the direct, present evidence of ability to
learn via the L2 as demonstrated during the dynamic simulation. How examinees learn subject-matter knowledge is in accord with the CLEAR perception that language is a functional tool to "Use" in learning academic content. In measuring content acquisition, not language, the CLEAR test builders' goals are akin to that of general educational or psychological measurement, compared to purposes of standard language tests (Bachman, 2002, p. 6). The CLEAR perception of language is not a matter of abstract knowledge, but something to use. The CLEAR seeks to capture the use of language in a dynamic experience, wherein examinees are tested on how well they learn content knowledge through the medium of English (Garner & Borg, 2005).

Metaphorically, the TOEFL stands inches away from an object, myopically examining chips of glass and ribbons of solder. The CLEAR stands back, gaining a perspective of the entire stained glass window of language; more importantly, the CLEAR attempts to look through the language window and view the target environment beyond. Thus, the CLEAR may reveal language proficiency, but primarily it aims to illuminate what examinees can do through their proficiency by assessing dynamic language use for learning in context.

Measuring language knowledge, as is done in the TOEFL, is not the same as revealing ability to do things through language, the task of the CLEAR. In the TOEFL, examinees are limited from demonstrating their full abilities in action (Asher, 1990; Roemer, 2002). ESL professionals can recount anecdotes of individuals who earned
TOEFL scores incommensurate with reasonable expectations. That is, some ESL students earn higher TOEFL scores than would be expected from ESL class performance in the judgment of ESL experts; more anecdotes are told of individuals strong in classwork who earn poor TOEFL scores (Larsen-Freeman, 1986). High scores on the TOEFL are usually associated with test-takers who can pass their university classes (although this is not assured as seen in Dunkel & Davis, 1994 and in Mason, 1994). This might be because the TOEFL functions well in the cases of highly proficient ESL candidates, or it might be that TOEFL is tapping a competency that overlaps with other abilities useful in university study, whether such competency is intelligence or another ability (Larsen-Freeman, 1986).

The Fallacy of TOEFL Score Interpretations

The logical fallacy of this interpretation of the TOEFL score is that, if high-scoring examinees tend to manage in university study, then low-scoring examinees will not (Hinofotis, 1986). Such a conclusion imputes to the TOEFL two abilities. First, the TOEFL is presumed to measure the relevant construct or constructs. Yet, such an assumption is not necessarily in evidence simply because high-scoring examinees also do manage to pass university courses (Larsen-Freeman, 1986). The second assumed ability imputed to the TOEFL is the similarity of measuring the aforesaid construct across proficiency levels. That is, TOEFL scores are interpreted as offering a particular quality of measurement at not only high levels of proficiency, but maintaining this measurement
quality at lower proficiency levels (Hinofotis, 1986).

These assumptions may do a disservice to university applicants with TOEFL scores below the recommended minimum. Since the language proficiency test does not measure AR, university admissions officers have little or no evidence for applicants' qualifications in terms of their academic skills (reading, writing, and reasoning) and academic auxiliaries (motivation, study skills, engagement, work drive, emotional stability, affective schemata, metacognitive strategies). These qualities are revealed in the CLEAR dynamic simulation, and contribute to university success. Thus, university applicants with TOEFL scores below the recommended minimum might prove their abilities better in the dynamic assessment of the CLEAR simulation. That is, many applicants might be able to demonstrate learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions for the same. This might be possible even if these students cannot attain common cut-point scores of 550 on the paper-based TOEFL or 213 on the computer-based TOEFL (Taller, 2004).

Successful knowledge acquisition requires language proficiency beyond a basic threshold (Adamson, 1993; Alderson, 1984; Clapham, 1995; Clarke, 1980/1988; Cummins, 1979; Elder, Erlam, & von Randow, 2002; Garner & Borg, 2005; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997; Taillefer, 1996). However, seeking to measure the linguistic threshold is not as simple as might appear. The
linguistic threshold might vary according to topic, task, or reader background, in the
view of many scholars (Alderson, 1984, p. 20, Alderson, 2000, p. 39, Grabe, 2000, p. 243,
Urquhart & Weir, 1998, p. 72). Yet Laufer (1997) argues that the threshold is stable and
essentially a function of lexis. However, some research (Clapham, 1995) suggests the
presence of multiple thresholds. Practical issues add to the problem of seeking to measure
the linguistic threshold: the presence and location of the threshold (or thresholds) are
generally inferred from other data, and no commercial tests of threshold proficiency exist.

Moreover, threshold proficiency is simply a necessary condition, but it is not a
sufficient one. Knowledge acquisition also draws upon a complex constellation of other
examinee attributes. Whereas the below-threshold learner must focus all energies on
message decoding and encoding, such is not the case for the linguistically proficient
learner. This student is free to deploy such strengths as topical knowledge and L1-based
academic skills, as well as academic auxiliaries such as motivation, work drive,
metacognitive strategies, affective schemata, and personal characteristics (Adamson,
1993; Bachman, 2002; Bachman & Palmer, 1996; Elder, Erlam & von Randow, 2002;
Huong, 2001). This bears much in common with the model of academic competence by
educational psychologists S. Elliott and DiPerna, student academic achievement is
influenced by "academic skills," such as skills in reading/language arts, mathematics, and
critical thinking skills and influenced by "academic enablers" of motivations, study skills,
and engagement (S. Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, S. Elliott, &
DiPerna, 2002). For example, the student who can comprehend the L2 lecture can employ critical thinking skills, can become engaged with a class discussion or reading the text, so these qualities contribute to the learner's ability to benefit from L2-mediated content instruction. Demonstrating ability to learn new subject-area material to a level appropriate for native-speaking students, and having learned under conditions for the same, can provide evidence of AR. (Although it should be noted that the level deemed appropriate would likely vary by discipline and academic level. Establishing this level of acceptability would be a matter for additional research setting CLEAR passing scores.)

AR will be examined through the dynamic assessment of the CLEAR wherein test-takers learn new material and are later tested on their knowledge acquisition. The CLEAR represents the university experience in microcosm, for examinees are assigned a textbook chapter to read, they take notes while following a university professor's lecture on videotape, they are allowed hours of study, then they are tested on content knowledge acquisition through multiple-choice and essay tests. This micro-world provides a context, an environment in which the examinees can demonstrate their AR. The CLEAR presents a simulation approach that bears some fidelity to the actual learning experience (although some verisimilitude might be eroded by conditions to prevent breaches of security or attempts to standardize testing conditions). Domain sampling and representation are key qualities in a criterion-referenced measure; with evidence of these, support is provided for claims for validity of CLEAR score interpretation.
History and Description of the CLEAR Dynamic Simulation Assessment

In an unconventional move, university admissions authorities in 1999 granted authorization to the host institute of this research (called by the pseudonym "English for Academic Purposes Institute," and hereafter called "the EAPI"). The EAPI became authorized to recommend ESL students for admission. This large southeastern research university (pseudonym "LSRU") acknowledged the expertise of the professionals at the EAPI, and committed to admitting institute-recommended applicants into degree programs without requiring them to submit TOEFL scores.

The EAPI leadership could have based this recommendation upon ESL class performance, or could have shifted to ESL teachers the responsibility for predicting student success or failure in the university. Also questioning the validity of inferences based on TOEFL scores, the EAPI leadership sought to diverge from the tradition of using and imitating this test. Instead, they resolved to develop a new assessment procedure, one that would provide an empirical foundation for the admissions recommendations. Out of this decision was born the battery now known as the Content Learning Experience: Academic Readiness test, the CLEAR.

A test development team has been working on the CLEAR since the turn of the millennium. The committee is comprised of two university professors in applied linguistics (of whom one was the EAPI Director) and two ESL experts holding M.A. degrees. The fifth committee member, the CLEAR Coordinator, is the writer of this
dissertation. Although the present study is focused upon test score validation, the test
development history is not unrelated to this purpose. The historical notes may support
claims for test quality, while helping contextualize the CLEAR for readers.

Development work initially proceeded down familiar paths. The goal was to write
a program exit test. As language can be analyzed into components, it seemed important to
measure every possible language-related element. Early test plans comprised assessing
numerous component skills relevant to academic study: reading comprehension, reading
notes, lecture listening comprehension, writing lecture notes, library research notes,
research paper drafts, formal in-class oral presentations, informal spontaneous oral
interaction, and knowledge of academic culture. (To illustrate, Figure 1 presents an
extract from an early version of the CLEAR skills specifications for assessing study
skills, research, and vocabulary.) These measures were not intended to measure AR, but
were constructed from familiarity with teaching such component skills in the EAPI, and
the early intention to build a test measuring achievement in learning what is taught at the
EAPI.
This atomistic approach essentially reflected assumptions about the nature of language and about the purpose for testing. The CLEAR team initially assumed that measuring every discrete language element would add up to the “whole” of measuring language ability in use. Yet, not all language skills are equally important. Speaking, for example, is less important at the undergraduate level and in certain disciplines, whereas writing seems important across academia (Carson, 2001; Horowitz, 1986b; Rosenfeld, Leung, Oltman, 2001; Waters, 1996; Weigle & Nelson, 2001). Moreover, it became increasingly obvious that the CLEAR could never tap every component of language and
academic skills. Trying to do so yielded a cumbersome assessment product. Pilot testing, committee discussions, and reviewing the research literature helped the team return to the fundamentals, and ask once more, “What do we need to answer through this test?”

The Test Purpose of the CLEAR

Identifying the test purpose is fundamental in test development, and indeed is a pivotal determination. The test purpose helps the developers construct an instrument shaped for its intended use, analogous to structuring a curriculum along instructional objectives. The developers, reconsidering this matter, determined that the test purpose of the CLEAR did not fit into any of the standard language test categories (Hughes, 1989, p. 9; NAFSA, 2003, p. 2-4) of achievement, diagnostic, placement, and proficiency test types. That is, the CLEAR is not meant to measure second language (L2) learning of a set curriculum, as is the case with achievement tests. Nor is the CLEAR intended as a diagnostic test to identify instructional needs. Finally, the CLEAR does not seek to quantify the examinee's language level without respect to a curriculum, as is true of placement and proficiency tests.

The test developers began to realize that, precisely speaking, the CLEAR does not need to measure the language development nor attainment of these ESL examinees. If admitted to college-level classes, non-native English speakers (NNSEs) will need to process information they read and hear, take notes, and write, all based on the new information. The NNSEs will need to analyze, synthesize, summarize, and investigate in
the university-level content class. However, they will perform few activities that are the norm in ESL classes. They will not take listening comprehension quizzes, nor will they turn in lecture notes for a grade, they will rarely need to submit multiple drafts of a paper for the content-area instructor. Indeed, many ESL teaching and testing activities will never occur in the college classroom. Thus, a simple question will be answered by CLEAR data: Can the L2 examinee demonstrate learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions for the same?

In other words, the CLEAR need not test language level. AR is the construct the test developers seek to measure, not language proficiency. The latter bears an indirect or unclear relationship to academic success (Adamson, 1993; Canadian Psychological Association, 1999; Elder, Erlam, & von Randow, 2002; Graham, 1987). AR seems much more direct evidence upon which to base university admissions recommendation: AR does not show accumulated knowledge but shows the examinees' ability to acquire new knowledge, and does so through a direct sample of student learning performance (Haywood & Tzuriel, 2002, p. 41; cf. Embretson, 2000). A crucial point to recall is that content knowledge is "arguably the primary attribute assessed in many college courses" (Elliot & DiPerna, 2002, p. 10).

In considering the what the new test should measure, the EAPI has no need to compare or rank examinee performance, nor to report fine gradations of performance.
Indicators therefore do not point to the usual norm-referenced testing approach, but rather to a using criterion-referenced measurement approach. Additionally, the Institute must merely classify examinees into two groups of AR: examinees who are likely and those who are not likely to succeed in deriving knowledge from content-area instruction delivered in English. Keeping in mind the simple question to be answered helped identify the CLEAR test purpose, a category found in general educational testing and evaluation, namely the certification test. Also known as the mastery test, this is a test that results in a pass/fail decision, and is often associated with the concept of “protection” (AERA et al., 1985). Certification tests restrict privilege and may protect public safety. A driving test, for example, is a certification test, as are medical board examinations and teacher certification examinations. Certification test results are a simple pass/fail, not gradations or rankings of examinee performance to others. A driver's license, for example, does not report how the driver performed compared to other examinees. The CLEAR fits this category because examinees need only be distinguished into pass/fail groups for undergraduate and graduate levels. Furthermore, the CLEAR is instrumental in protecting the university and examinees from recommending unready students and in protecting the professional credibility of the EAPI.

In rethinking the test function and test type, the CLEAR developers focused again on the motivation for a TOEFL-less recommendation. Attention returned to issues of validity, the central quality of good measurement. Validity lies at the heart of whether the
TOEFL measures what it is supposed to measure and whether inferences made on the basis of TOEFL scores can be supported. Questions about the validity of the meaning of TOEFL scores formed the basis for many of the EAPI objections to using the test for admissions purposes. Therefore, the CLEAR team paid careful attention to developing a contextualized assessment with claims to verisimilitude with the target environment. The validity claims of the CLEAR, furthermore, are central to the present research investigating the measurement quality of this dynamic simulation assessment.

Authenticity in testing is claimed by many proponents of alternative assessment. Yet, the CLEAR team considered that, in a test, complete authenticity is not possible (cf. Bachman, 1990, p. 10). Because a test has stakes attached, an activity performed for a test cannot be completely like an activity done without any stakes. A test will be more or less obtrusive, not according to the test format, but according to the importance of the test outcomes. The developers set their sights on enhancing the quality of the test by improving the congruence between the real world and the test, building a contextualized test, and eliciting representative performance data from the examinees. Upon consideration, the team determined that the CLEAR was not concerned with measuring products of prior learning, but would consist of a series of linked content learning and performance tasks. Embretson (2000) expresses well the conundrum experienced by the CLEAR developers: "The dominant paradigm for measuring abilities remains performance on problem-solving tasks that depend heavily on prior knowledge or
developed skills. A disadvantage of tests developed in this paradigm is that the learning ability construct is measured indirectly. A more direct approach would simulate the learning situation" (p. 505). That direct, simulation model is key to appreciating the learning potential of ESL university applicants, and is a central innovation granted in the CLEAR.

This simulation approach was chosen for the CLEAR, as the developers focused upon a means of discerning evidence for candidates' university readiness. (The dynamic simulation approach employed in the CLEAR will be explored in more depth later in this chapter, but an introduction may prove helpful here.) A simulation approach seemed appropriate for the goal of the CLEAR, i.e., to identify among ESL examinees those who are likely to be academically ready for university study. Using a simulation may yield data supporting claims of validity, specifically that the CLEAR “ecology” or context allows generalizing to target situations (Garner & Borg, 2005). This connection is especially important in certification-type tests (cf. AERA et al., 1985). The simulation may also be a good approach to tap examinees' abilities. Academic success evidently relies upon a complex constellation of skills and enabling behavior (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, J. Elliott, & DiPerna, 2002), not merely linguistic prowess. Giving examinees a simulation of the university learning/testing cycle appears to offer an effective technique for testing AR.
The Construct of Academic Readiness (AR)

As the CLEAR team has defined this construct, AR is the direct, present evidence of ability to learn via the L2 as demonstrated during the dynamic simulation assessment. In operational terms, "learning" implies understanding new subject-area material, to a level appropriate for native-speaking students, and having learned under conditions for the same.

Underlying this evidence of AR may exist a shifting array of abilities that work in compensatory harmony according to the particular task demands (cf. Bachman, 2002; Garner & Borg, 2005). Cognitive/academic language proficiency is a fundamental quality, and so is language proficiency above a threshold level in order to learn by means of the L2. (Adamson, 1993; Alderson, 1984; Clapham, 1995; Clarke, 1980/1988; Cummins, 1979; Elder, Erlam, & von Randow, 2002; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997; Taillefer, 1996; Turner, 2004). Other enabling qualities are motivation, study skills, engagement, and academic skills such as reading/writing skill, math/science skill, and reasoning skills (Elliot & DiPerna, 2002). The SLA literature identifies other contributing resources such as topical knowledge, affective schemata, metacognitive strategies, and personal characteristics (Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Huong, 2001).

Latent traits, however, are not the target constructs of the CLEAR: As a test of AR, the CLEAR seeks direct evidence. The dynamic simulation gives examinees a university
learning experience in microcosm to see how well they can learn new academic content. Thus, AR indicates the examinee can acquire new content knowledge, of central interest in university assessment (Elliot & DiPerna, 2002).

In the extended dynamic simulation of the CLEAR, examinees have the opportunity to test their mettle in a context very much like what they will experience in the university (Garner & Borg, 2005). Their performance, and their perceptions of the experience, may extend the discussions concerning the Linguistic Threshold Hypothesis, or LTH (Alderson, 1984) examined below, for the CLEAR developers believe that AR will be evident only in examinees whose language proficiency is above this threshold. AR research might also illuminate our understanding of the dimension of cognitive/academic language proficiency, commonly called "CALP" (Cummins, 1980), in adult ESL students preparing for university learning through English.

A brief discussion of CALP may prove helpful to illuminate the connection between CALP and the CLEAR. Cummins (1980) distinguished two types of language (as will be further elaborated in Chapter 2). Briefly, Cummins termed the language of everyday conversation "Basic Interpersonal Communicative Skills" (BICS); the language found in classrooms he named "Cognitive/Academic Language Proficiency" (CALP). Cummins (1983) situated BICS and CALP within two intersecting continua, the degree of contextual support and the level of cognitive demand. CALP, which can be empirically distinguished from BICS (Cummins, 1979, 1980, 1983) is associated with literacy and
formal education. Cummins (1980) attests that the L1 CALP dimension transfers for use in L2 contexts. CALP theory and research on L1 transfer are particularly germane to academic L2 reading, a critical skill area for students of English for Academic Purposes (EAP), such as those who will take the CLEAR. Since CALP is the product of literacy and formal education, this type of language development is a necessary foundation for success in academic learning.

CALP is a key component in L2 reading, which comprises both literacy and language. What is the relative importance of each component? Alderson (1984) phrased the issue as follows: Is a foreign language reading problem a problem based in L1 literacy or in L2 proficiency? Some scholars (Bernhart & Kamil, 1995; Lee & Schallert, 1997) have found evidence that both factors contribute. Especially important to L2 reading is the contribution of L2 proficiency, as is posited by the Linguistic Threshold Hypothesis (LTH). The LTH, which holds that fluent L2 reading is possible only when the learner has achieved a “threshold” of L2 proficiency, is generally well supported (Alderson, 1984, 2000; Alderson & Urquhart, 1985; Bernhart & Kamil, 1995; Clarke, 1980/1988; Cummins, 1979; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997; Taillefer, 1996; Urquhart & Weir, 1998). This theory is not out of line with the Linguistic Interdependence Hypothesis (LIH) which posits that, underlying academic competence in the L1 and L2 is a shared ability (Cummins, 1979; Lee & Schallert, 1997).
Relevant to this trend in L2 reading research is the line of inquiry into background knowledge or content schemata. Scholars (Carrell & Eisterhold, 1983; Clapham, 1996) have considered how L2 reading performance is moderated by background knowledge. The relationship between subject-area knowledge and improving L2 reading comprehension, however, is not straightforward (Alderson & Urquhart, 1984). Background knowledge of the subject area might help L2 readers with some degree of proficiency, but such knowledge does not help beginners (Ridgway, 1997). Beginners' comprehension is “short-circuited” (Clarke, 1980/1988 p. 119) by inadequate L2 ability. Background information does seem to help those who have crossed beyond a threshold of some proficiency level, hence the LTH (Alderson, 1984, 2000; Alderson & Urquhart, 1985; Bernhart & Kamil, 1995; Clarke, 1980/1988; Ridgway, 1997; Taillefer, 1996). LTH support has also been found in the L2 vocabulary research by Laufer (1997): analyses revealed a threshold of roughly 5,000 words, as will be discussed in further detail elsewhere in this study.

In light of this research, it is expected that the CLEAR examinees will perform likewise, e.g., by proficiency differences, they will be differentiated into two contrasting groups. Examinees with sufficient proficiency to have crossed the linguistic threshold will be able to cope with the instruction delivered in English whereas those below the threshold will not be able to benefit from the instruction. The two groups might not
correspond with class placement, since instructional placement differs by language institute.

CLEAR Dynamic Simulation Assessment Description

The CLEAR can be described as a simulation, since it is a microcosm of a real-world instructional cycle with embedded testing events that take place over a 48-hour period. The term "simulation" has been adopted from the fields of instructional technology, although its roots can be traced to engineering, work study, and military strategy (Romiszowski, 1993, p. 170). In instructional technology design, a simulation can replicate a real world system (Gibbons & Fairweather, 1998, p. 24). This micro-world is useful in measuring task performance (Romiszowski, 1993, p. 172). The simulation "is widely appreciated as a powerful tool for instructing higher-level principles, procedures, and cause-effect relationships. The ability of simulations to grab and hold the attention of learners is an effect so common that it is seldom questioned," noted Gibbons and Fairweather (1998, p. 297). Although valued, the simulation is a challenge to the instructional designer, state Gibbons and Fairweather (p. 297), because it is expensive to develop, it requires "non-sequential and non-linear thinking" in designers, and it may be difficult to implement within a curriculum. Instructional simulation thus presents strengths and weaknesses; these advantages and drawbacks apply in a similar manner to simulation for assessment, rather than for instruction.
Simulation describes the format of CLEAR because examinees participate in a university micro-world. That is, the assessment measures task fulfillment of learning new information through tasks that link learning with performance. Test events unfold progressively, as in real university experiences. Thus, the CLEAR examinees read a university-level textbook chapter, then they view a university professor's videotaped lecture on a related topic, all the while taking lecture notes. Examinees may opt to study during a supervised study hall session using their own notes and texts. Later, examinees are assessed on how well they learned subject matter knowledge by means of a selected-response test and an essay task. Of the two dynamic simulation test components, one is delivered in selected-response format (the knowledge subtest) and one employs constructed-response format (the essay subtest). The knowledge subtest was scored by the researcher. The essay rating was conducted by an expert judge and the researcher using rubrics constructed by the CLEAR developers. By simulating in microcosm the world of university learning and testing, the CLEAR test potentially offers more fairness to examinees (Grigorenko & Sternberg, 1998; Haywood & Tzuriel, 2002; Sternberg & Grigorenko, 2002) and more information to test score users (J. Elliott, 2003; Embretson, 2000), yielding data about examinees' ability to learn content taught via English.

Learning new information lies at the heart of the CLEAR. The CLEAR is a simulation, one comprising learning opportunities then content-knowledge testing, with the intent to measure novel learning. This description mirrors Grigorenko and Sternberg's
(1998) statement that "Testing thus joins with instruction and the test taker's ability to learn is quantified while she or he learns" (p. 75). The CLEAR team, therefore, developed a dynamic test (J. Elliott, 2003; Embretson, 2000; Grigorenko & Sternberg, 1998; Haywood & Tzuriel, 2002; Laing & Kamhi, 2003; Sternberg & Grigorenko, 2002; Swanson & Lussier, 2001). In such a measure, the test examiner provides "active and direct teaching precisely in order to produce change. Thus, the basic datum in dynamic assessment is a change variable: How do examinees learn new things?" (Haywood & Tzuriel, 2002).

Haywood and Tzuriel (2002) recount the concepts fundamental to dynamic assessment. First, achieved knowledge is not the best predictor of ability to acquire new knowledge (although the two are highly correlated). Second, every person functions at below-optimal capacity, so all persons have the potential for improvement. Third, a performance sample is the best test of a performance, so measuring learning ability can be tested by using learning tasks, particularly with instruction. Finally, there exist blocks to a person's accessing and effectively using his or her intelligence. These blocks can be, among other things, ignorance, poor motivation, flawed self-concept, inadequate study habits, ineffective learning styles (p. 41).

The CLEAR may arguably be characterized as a dynamic test, Yet, the CLEAR is distinct from other dynamic tests: the CLEAR provides examinees with a contextualized experience (Garner & Borg, 2005), one that bears verisimilitude with the target
environment. The CLEAR is a simulation (Gibbons & Fairweather, 1998, p. 22). Herein, the university learning cycle is simulated by the series of linked learning and performance tasks (e.g., textbook reading, observing a lecture, studying, and taking multiple-choice and essay tests). Blending the dynamic testing approach with the simulation model results in the dynamic simulation approach of the CLEAR. The dynamic simulation approach is poised to offer richer information about the IEP students' abilities and readiness for university-level study, knowledge of interest to examinees, the language institute, and researchers.

**Rationale**

The CLEAR takes an innovative approach to testing whether ESL students are academically ready for university study. After directly sampling performance in university tasks, instruction through lecture and textbook followed by an encounter with a college instructor, the CLEAR simulation culminates in testing examinees by multiple choice and essay instruments. While this simulation of the college learning experience offers at least an intuitive appeal, questions remain concerning the quality of the present instrument and, on a larger scale, the usefulness of the simulation approach.

The rationale for the present study is the possibility that the CLEAR might provide an assessment that is potentially more informative and more fair than measures used heretofore. Such a tool might also shed some light on measuring AR in NNSEs who are prospective university students. As well, the CLEAR approach may help prompt other
EAP institutes to consider the implications of the college admissions testing program for their curricula. All these are impossible without assessing the CLEAR measurement quality, the purpose of the present research.

The validation of the scores derived from the CLEAR is important because this stage is requisite in good measurement and because the EAPI decision-makers need to know the quality of data collected by the CLEAR. The CLEAR, moreover, may help improve university admissions testing of EAPI students at LSRU, broadening the options for the applicants and the university, and other institutions that may adopt this tool. The CLEAR might also be useful in improving our understanding of the issues relevant to L2 speakers studying in U.S. universities.

This alternative measure may, eventually, prove helpful for other universities' admissions decisions, as the CLEAR is potentially more fair towards examinees and more informative for all test users. More probable, however, is that any impetus for change would come from English for Academic Purposes (EAP) institutes seeking an instrument that elicits how examinees actually use academic English, as the CLEAR attempts to do. Claiming that the CLEAR may offer benefits of fairness and rich data is a claim which is grounded on the CLEAR task conception that uses the testing cycle; perhaps the CLEAR's innovative contribution of simulation, adopted from instructional technology, will be useful to second language testing.
Test construction is highly pertinent to teaching and research; CLEAR pilot sessions already produced initial findings of interest to the field and to EAPI students and teachers. More solid data would strengthen confidence in initially tentative findings. Such might help in making decisions concerning EAP curriculum, instruction, and evaluation. Finally, examining the construct of AR may prove a contribution to ongoing discussion of the LTH.

Statement of the Problem

The EAPI has been authorized by its host university to make admissions recommendations for NNSEs and opted to base recommendations on empirical evidence. As mentioned earlier, North American universities usually screen NNSE applicants for linguistic proficiency through the TOEFL, a test consisting primarily of multiple-choice items. Deeming the TOEFL inadequate in measuring the NNSEs' ability to actually use English in actual academic contexts, the EAPI leadership chose to construct a new instrument, the CLEAR dynamic simulation. This assessment answers the EAPI's need for evidence of ESL students' ability to function in academic settings. Seeking evidence of AR, the CLEAR does not follow in the tradition of the TOEFL. The CLEAR employs an innovative approach of “dynamic assessment” (Sternberg & Grigorenko, 2002). That is, the test first gives examinees learning experiences, and these lead to testing events. In the CLEAR dynamic simulation of the university-level learning cycle, the language skills of listening, speaking, reading, and writing are not isolated. Rather, the skills are
naturally interwoven throughout academic activities. These unfold in an integrated experience, wherein work products are graded by norms employed in subject disciplines. The EAPI has the confidence of the university, and has invested years into the CLEAR development. After numerous improvements and three pilot rounds, the CLEAR has been shaped into an innovative tool. The measurement quality of the test is yet unproven, however. The ensuing step in the test development process is to test the CLEAR under stringent and consistent conditions and with a sufficient sample in order to evaluate the test measurement quality. Every test score reflects both the examinee's true ability and some measurement error, according to classical test theory (Crocker & Algina, 1986, p. 106). The error, which is related to reliability, must be estimated to support good decision-making. The present validation work has yielded information in order that the CLEAR team is better equipped to interpret the meaning of CLEAR test results.

**Purpose of the Study**

Responsible test owners know the meaning of scores derived from their tests; thus, the institute that commissioned the CLEAR needs the present research to illuminate where the dynamic simulation is effective and accurate, and where it needs improvement. Similarly, examinees who take the CLEAR are entitled to a fair test from which reasonable interpretations may be drawn. The test user needs the CLEAR in order to make sound educational decisions concerning examinees' AR. Information to help answer these questions is provided by the present investigation.
The purpose of this study is to estimate the measurement quality of the CLEAR dynamic simulation and analyze the test-takers' perceptions. The research will first conduct fundamental analyses, impossible to calculate under the limitations of pilot conditions, to ascertain the sample equivalence of the two testing sites, then compare performance across different subgroups, then to calculate knowledge subtest item statistics.

The investigation next examines the consistency of the scores derived from the CLEAR. For the knowledge subtest this is estimated via Cronbach's alpha, while descriptive statistics and the standard error of measurement help answer this question. In the essay subtest, consistency is reported in terms of essay scorers' ratings. Validity concerns are also addressed: the study investigated evidence for claims of validity based on content, criterion, and construct grounds. Content-related evidence for claims of validity is provided by a content expert evaluation. Criterion-related evidence for claims of validity are reported by correlating concurrent measures of ability (placement test scores, proficiency test scores, and teacher ratings) with the two CLEAR subtests. The third aspect of validity to be investigated was construct validity, examined by inter-correlation of scores on the two subtests. Completing this research is the study of the examinee perception of the CLEAR, that is, how well test quality and effectiveness balance with efficiency in this dynamic simulation.
Whether this test is solid or riddled with measurement error cannot be known without test validation. Validation is best approached as a series of studies that will help support claims of test quality. The present study is crucial, as the first step in this ongoing validation process.

Research Questions

Examining the measurement quality of CLEAR is necessary for three aspects of the test, and will be organized into three major research questions. These questions are reported, then each is briefly discussed. Prior to investigating the actual research questions are some preliminary analyses to estimate test functioning.

Research Question 1: What is the consistency (i.e., reliability) of the scores derived from the CLEAR?

Research Question 2: What evidence exists for claims for validity of inferences based on CLEAR scores?

Research Question 3: What is the examinee perception of the CLEAR?

The preliminary matters investigated sample equivalence across testing sites, then equivalent test functioning across subgroups. Knowledge test item analyses were conducted to estimate the p-value, item discrimination, and distractor analyses for this subtest.

The first question considers, What is the consistency (i.e., reliability) of the scores derived from the CLEAR? The knowledge subtest consistency was analyzed in terms of
Cronbach's alpha, whereas the essay subtest consistency was measured by essay scorers' inter-rater consistency. Descriptive statistics for the overall knowledge test were reported, as was the standard error of measurement (SEM).

The second research question is, What evidence exists for claims for validity of inferences based on CLEAR scores? To answer this, evidence was collected, from the perspective of content, criterion, and construct aspects of validity. Content-related evidence for claims of validity was provided by a content expert evaluation. Criterion-related evidence for claims of validity was reported by correlating concurrent measures of ability (placement test scores, proficiency test scores, and teacher ratings) with the two CLEAR subtests. The third aspect of validity to be investigated was construct validity, analyzed by inter-correlation of knowledge subtest and essay subtest scores. One aspect of validity is the predictive utility of the CLEAR, investigating any relationship between CLEAR dynamic simulation scores and examinees' future success. This predictive aspect of validity-related evidence would be most helpful in demonstrating the merit of the CLEAR. Limitations of time and funding do not permit including this predictive aspect of validity claims into the the scope of the present work. However, findings from this initial study suggest that the CLEAR may offer unique insight into examinees' learning potential.

The third research question asks, What is the examinee perception of the CLEAR? This question seeks the assessment of test quality in the judgment of the
examinees, answered through qualitative data analysis of examinee feedback, in order to help estimate the relative quality of information from the dynamic simulation.

These data have illuminated the present state of measurement quality of the scores derived from the CLEAR. The present study, it is hoped, may contribute to further improvement for a future version of the test.
**Delimitations and Limitations**

As a validation study of the CLEAR Test, the present research is critically concerned with the test quality and functioning. Delimiting the bounds of the research is important, as is detailing the limitations of the study. These factors can be grouped into those related to the examinees and to the present study itself. (Limitations exist in terms of the instrument, but such are discussed in the instrumentation section of the Methodology chapter.)

The participants for this validation study were drawn from administrations at two different EAP institutes in Florida. The dual-site testing might better yield results representative of typical pre-university ESL students at the intermediate or advanced level. The sample obtained in the summer semester might not be congruent with the population in attendance during the regular academic year, but such would need a separate investigation. The examinees taking the CLEAR test were adults from an international population: the CLEAR was not designed for a population of North American residents, whether native speakers or not.

In terms of the present study, time and dissertation requirements have limited the scope of coverage. It was not feasible, for example, to make and set the passing score as a function of the present research. Standard-setting research demands a later investigation; this would be constructed on the knowledge base established in the present initial validation study. With information about the quality of the test items and the meaning of
CLEAR simulation scores, future standard-setting research can help ensure that pass decisions will be made upon sound data and reasoned judgment. Other studies remained beyond the reach of the present research. For example, of great benefit would be discerning how well CLEAR results can predict examinees' success in university studies. Investigating predictive utility is important research that can follow the groundwork laid by the present study. Another useful future research project that might ensue from the present work would be examining consequential validity; such a study might consider the impact of the CLEAR on examinees, teachers, curriculum and instruction, and on university admissions.

Despite the limitations, the present validation study can contribute to future investigations. This dissertation has helped construct a knowledge base that may someday support a CLEAR test manual. The present work, by commencing with fundamentals, might later prove to have laid a foundation for research into predictive utility or consequential validity.
Definitions

**Academic readiness**: The ability which the CLEAR dynamic simulation attempts to measure, AR is defined as the direct, present evidence of ability to learn academic content via the L2 as demonstrated during the dynamic simulation, learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions similar to the same. AR indicates that ESL examinees' second-language (L2) proficiency is above the "threshold" level, and therefore sufficiently advanced to allow the examinees to allocate resources to learning new subject matter information transmitted in the L2. The construct of AR is hypothesized to comprise the elements of CALP (i.e., above-threshold academic language proficiency), personal characteristics (age, educational background, etc.), topical knowledge, academic skills (reading and writing skills, mathematics and science skills, reasoning skills such as analysis, synthesis, and investigation), and academic auxiliaries (motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies).

**BICS**: Basic Interpersonal Communication Skills, or context-supported and cognitively undemanding language or everyday conversation.

**CALP**: Cognitive/Academic Language Proficiency, or context-reduced and cognitively demanding language of the classroom.

**Certification test**: a test that reports decisions as simple pass/fail; usually associated with
the concept of “protection,” as driver licenses protect public transportation safety.

CLEAR: The Content Learning Experience: Academic Readiness Test

Communicative competence: a very broad view of the abilities that comprise human communication in language, including linguistic, sociolinguistic, and strategic nodes in the Canale and Swain (1980) model.

Consistency: similar to the measurement quality "reliability," this term is employed in criterion-referenced measurement approaches. Cronbach's alpha is a measure of internal consistency, or how consistent is response behavior across the items within a test.

Dynamic assessment: measurement which incorporates a learning experience, so the tool measures fluid learning, rather than previously learned knowledge.

EAP: English for Academic Purposes, a field within the general ESL discipline. EAP instruction is typically delivered to adults preparing for university study in English.

ESL: English as a Second Language. This acronym will be used to identify those persons who are still studying English, though many intend to enter an English-medium university program.

ESP: English for Specific Purposes. Rather than general language development, ESP programs are more oriented to the students' vocational or academic purposes for L2 English.

IEP: Intensive English Program. This is an institute, typically situated in an English-speaking country, offering English instruction for 20 to 25 hours a week to adult learners.
**L1:** First language, or mother tongue, acquired naturally by all normal humans.

**L2:** Second language. Any language acquired after the L1, whether learned in a formal setting or naturalistically, in the environment of the L2 or in a foreign language setting.

**LSP:** Language for Specific Purposes. Instruction and testing tailored to the occupational or academic focus of the L2 learner.

**LIH:** Linguistic Interdependence Hypothesis. According to Cummins, L1 skills are hypothesized to transfer to the second language. Although an L1 and L2 have different surface features, below the surface are abilities that transfer across language. This hypothesis has been graphically represented by a "dual iceberg".

**LTH:** Linguistic Threshold Hypothesis, proposed by Cummins, posits that L2 proficiency is crucial in L2 reading. At low L2 proficiency levels, L1 literacy skills will not benefit the student, but above the threshold of language proficiency, these skills are engaged. Scholars disagree on whether the threshold is stable or variable.

**NNSE and NNSEs:** non-native speaker(s) of English; in this investigation, the NNSE acronym will be used referring to former ESL students, now in English-medium university study.

**Reliability:** the measurement quality which indicates how stable are test scores across examinees, administrations, raters, or from one part of a test to another. This quality is termed "consistency" in criterion-referenced tests. The reliability within a dichotomously-scored test may be measured by Cronbach's alpha or the Kuder-Richardson 20 formula.
**SME:** Subject Matter Expert, a person qualified in a particular field or discipline to judge the merits of content-relevant issues in a test.

**TOEFL:** Test of English as a Foreign Language, the pre-eminent test demonstrating English L2 proficiency for ELS students seeking university admissions in North America.

**Validity:** the fundamental quality typically associated with the idea that “the test should measure what it is supposed to measure.” Recent theory (AERA *et al.*, 1985) holds that validity is not inherent to the test itself, but is a matter of supporting inferences based on test scores. Three broad approaches exist to support validity claims: content-related evidence, criterion-related evidence, and construct-related evidence.
CHAPTER 2

LITERATURE REVIEW

Introduction

Language testing is intrinsically connected to a broad range of topics. Certainly this is the case of the CLEAR, a dynamic simulation assessment of AR for ESL university applicants. Examining the range of topics relevant to the CLEAR will proceed in this literature review as follows. After a brief review of the test under scrutiny, the chapter will address assumptions and approaches in language and in testing, then proceed to discussing issues of relevance to AR. The problem of defining language proficiency will be considered, next a treatment of intelligence testing, and then theories on language development. Having examined broader, more theoretical issues, the chapter then continues with examining the research describing the target domain activities of academic reading and vocabulary, leading into the literature on academic writing. Measurement issues of reliability, validity, and the validation process are covered later parts of this study.
Recall that the CLEAR, under development since the turn of the millennium, was commissioned by the EAPI. Constructed with the intent to offer good congruence with the target situation, the CLEAR provided data about ESL speakers' performance in learning subject-matter knowledge through instruction delivered in English. Rather than following the static testing model that others employ, the CLEAR presents examinees with an opportunity to demonstrate their ability to learn new material in a simulation of the university learning/testing cycle. On this evidential basis, the EAPI authorities will be able to make informed university admissions recommendations without resorting to the use of the TOEFL.

Assumptions Underlying Language Tests: The Know-Use Framework

Educational tests are intended to permit informed decisions. Tests are not created in a vacuum, however, and this includes tests for L2 speakers. Any test for L2 speakers is fundamentally shaped by, and proceeds out of, the test builder's beliefs about the nature of language. It is useful to examine different approaches to perceiving and testing language.

How to test L2 speakers is critically affected by whether the test builder perceives language as something one knows or as something one uses. This will be examined in light of the Know/Use matrix, inspired in part by Spolsky (1989; cf. Widdowson, 1989). The “Know language” view indicates language as a subject of knowledge, and is evidenced by expressions such as “She knows some Chinese” or “She has not learned the
subjunctive in Spanish.” The “Use language” perspective considers language in functional terms, and is seen in phrasing such as “He can speak French” or “He gets by in Russian.” This Know/Use distinction can be seen as two endpoints along a continuum, as illustrated in Figure 2. A perspective closer to the Know end of the continuum leads the test developer towards particular decisions, just as a position nearer the Use point will incline that developer to other choices.

Figure 2
Know-Use Continuum

These two perspectives can be considered in light of many other parameters, yielding a matrix that comprises contrasting approaches to language testing. This Know/Use matrix of contrasting positions offers orientation helpful in constructing or describing a language test (summarized in Table 1). Some resemblance exists between the Know/Use matrix presented in this study and that of other researchers (cf. Chomsky, 1965; Hymes, 1970) who have contrasted abstract language proficiency of ideal speaker-hearers with communicative language performance by real language users. Such, however, is not the distinction of interest in the present study. The Know/Use
matrix was constructed for a different purpose, to contrast approaches to language and the consequences for language testing. By deliberately employing different terminology, this matrix may better distinguish itself from connection to other models.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Contrasting Orientations to Perceiving and Testing Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Know Language” Orientation</td>
<td>“Use Language” Orientation</td>
</tr>
<tr>
<td>Usual characteristics of test:</td>
<td></td>
</tr>
<tr>
<td>norm-referenced</td>
<td>criterion-referenced</td>
</tr>
<tr>
<td>based on theory</td>
<td>based on real-world practicalities</td>
</tr>
<tr>
<td>underlying construct derived from theory</td>
<td>tasks derived from job analysis</td>
</tr>
<tr>
<td>indirect testing: inferences based on construct</td>
<td>direct testing: performance as representative</td>
</tr>
<tr>
<td>static “snapshot” of crystallized knowledge</td>
<td>dynamic “video” of fluid ability in action</td>
</tr>
<tr>
<td>Example tests:</td>
<td></td>
</tr>
<tr>
<td>psychological tests (general measurement)</td>
<td>certification (general measurement)</td>
</tr>
<tr>
<td>language proficiency test</td>
<td>performance assessment language tests</td>
</tr>
<tr>
<td>language achievement test</td>
<td>task-based language testing</td>
</tr>
<tr>
<td>Testing problems:</td>
<td></td>
</tr>
<tr>
<td>support for theoretical assumptions</td>
<td>domain sampling for selecting task</td>
</tr>
<tr>
<td>all factors in appropriate proportion</td>
<td>construct underrepresentation</td>
</tr>
<tr>
<td>lack of correspondence with the real world</td>
<td>time-consuming, expensive test model</td>
</tr>
<tr>
<td>Potential measurement strength:</td>
<td></td>
</tr>
<tr>
<td>reliability</td>
<td>validity</td>
</tr>
</tbody>
</table>

Consider, for example, a test builder who takes the view that language is something to Know. This person would be perhaps more inclined to construct a theory-based proficiency test or achievement test. This Know-oriented tester would follow
norm-referenced measurement, seeking to find how much each examinee knows on a zero-to-infinity scale while maximizing differences among the examinees. This test constructor would likely be content with a traditional test, a static snapshot of what the examinee has learned. Such a snapshot is satisfactory in a theory-based test if the test taps the underlying construct. Of critical importance, therefore, is the quality of the underlying theory. This theory must account for all important factors, in appropriate proportions, without neglecting any needed aspect. The Know-type test often is indirect or abstract, and can suffer from problems with seeming irrelevance to the measurement target.

Know-type tests are common in educational and psychological testing; examples include many standardized tests of intelligence and the TOEFL.

By contrast, a different test would be the outcome if constructed by one who views language as something to use. This test is likely to include performance assessment, or take a task-based approach. Tests developed from the Use perspective are more likely derived from a task analysis or job analysis than a test from the Know viewpoint. For the Use-oriented tester, identifying criterion performance is important, since examinees are measured against this standard. It is critical to sample enough tasks and tasks that are representative. More like a dynamic video than a snapshot, the Use-type test features examinees' ability in action. Problems of the Use-type test include the narrow sampling: fewer tasks can be sampled in comparison to the quantity of items on a Know-type test. Use-type tests are typically time-consuming and expensive relative to
Know-type tests. Use-type tests are frequently used for certification and licensure needs; examples include driver license tests and medical board examinations.

The Know orientation was initially adopted by the CLEAR team, as the developers originally sought to formulate a conventional test based on theories of communicative competence. The shift towards Use orientation came about as the developers reflected upon the model for setting passing standards, defining required or optional test elements and setting pass scores. Theory of communicative competence permits a very broad conceptualization of language, accounting for constructs of grammatical, discourse, strategic, and sociolinguistic competence. Theoretical application to the present test need is, however, limited: current theory certainly attempts to list language abilities extensively. Theory falls short of accounting for the relative importance of skills. Moreover, a solid description of the compensatory nature of language is lacking: despite the "strategic competence" node, a clear, systematic description is lacking concerning how certain strengths compensate for other deficiencies. These aspects are crucial in determining passing scores of a test, for these factors are central to the realities of NNSE students in English-medium university programs. Thus, while the communicative competence model may be lauded for its attempt to comprehensively account for every aspect related to language proficiency, that model did not provide the CLEAR developers sufficient guidance for their test development needs.
The fundamental issue, as the CLEAR developers perceive it, is whether ESL students can function in the university learning/testing context (cf. J. Elliott, 2003; Garner & Borg, 2004; Haywood & Tzuriel, 2002; J. Turner, 2005). University success is generally based on learning knowledge (Elliot & DiPerna, 2002) not on measures of linguistic competence (Santos, 1988). For example, in the target setting, NNSE students will not take listening comprehension quizzes, but they will take tests of whether they learned from the professors' lectures. Thus, the CLEAR team gradually moved away from the Know orientation and, with growing awareness of the functional orientation of the test purpose, came to adopt a Use orientation for the CLEAR.

The Use orientation is reasonable and appropriate for pass/fail tests of ability, such as driver's license tests or professional board examinations. This kind of test, called certification or mastery testing, is relatively uncommon in educational settings, but may be found in some graduate comprehensive examinations. Theory may contribute to a certification test. However, the real foundation of such a test is practical, resulting from a job, task, or domain analysis. Thus, in the case of the CLEAR, the test is grounded in the realities of the university learning/testing experience, although the developers were informed by theories relevant to academic language proficiency. While such theories helped illuminate the test construction process, CLEAR was not derived from a theoretical model. The CLEAR was built upon observations and evidence about the target language use domain, the university setting. The present research does not seek to test
theory, yet might contribute to building theory, by revealing the performance of learners in a dynamic simulation of the university experience. Perhaps by offering such data to the knowledge base, a better profile may be depicted of EAP students' developing expertise.

Even in practical tests from the Use perspective, a conception of the nature of language is requisite. Such is the first step in shaping understanding of the tasks in the university experience and identifying the needed components in a test of AR. Far easier, however, is it to call for a definition of language than to agree upon one.

*Modeling Language Proficiency*

Scholars do not agree on a definition of language proficiency nor even on the constituent elements (cf. Bachman, 1990; Canale & Swain, 1980; Carroll, 1993; Oller, 1976/1983; Spolsky, 1989; Widdowson, 1989). Indeed, quantifying the precise factors comprised in language proficiency has been called the “philosopher's stone” in psychometrics (Jensen, 1987, p. 110 cited in de Jong and Verhoeven, 1992, p. 10). Just as psychologists have argued for over a century without resolving the issue of constituent factors in intelligence, so have linguists continued to debate the nature of language ability. Although the debate is far from resolved, and although the CLEAR does not seek to measure latent traits, language proficiency nevertheless plays an important role.

Language proficiency underlies achievement in educational settings. One type of language ability, CALP, may be seen as a product of education. Language ability is
traditionally tested by means similar to intelligence measures. Verbal reasoning and intelligence measures bear much in common. Among L2 speakers, threshold L2 proficiency is a necessary, though not sufficient, condition for AR. Teasing apart language proficiency and intelligence is by no means a straightforward matter, but such examination could help shed light on issues foundational to AR.

The differing models of language proficiency can illustrate the debate. Consider first perhaps the most prominent contemporary model of language proficiency, the “communicative competence” model. As interpreted by Canale and Swain (1980), this model identifies three high-level components—grammatical competence, sociolinguistic competence, and strategic competence. Sociolinguistic competence is further analyzed into two elements, sociocultural and discourse competence. This communicative competence model accounts for linguistic elements below, above and at the sentence level; it considers non-linguistic factors including facets of appropriacy and compensatory strategy use. The Canale and Swain model, and Canale's (1983) adaptation, both hold considerable weight with ESL teachers. Yet validating this model remains problematical (Cziko, 1984); criticisms (Skehan, 1998, p. 159) and alternative views persist (Bachman, 1990; Carroll, 1993; Cummins, 1980; Cummins, 1983; Oller, 1976/1983; Skehan, 1998; Spolsky, 1989; Widdowson, 1989).

For example, although the Canale and Swain conception is extensive, Widdowson (1989) suggested further modification, dividing each component into a competence
aspect and a performance aspect, yielding an eight-factor model. Widdowson, however, does not account for every contributory dimension in language. Bachman's (1990) construction of language competence is extremely complex. At the bottom of three levels of hierarchy exist 14 different factors. Subsumed under grammatical competence are vocabulary, syntax, morphology, phonemes/graphemes; under textual competence are found cohesion and rhetorical organization; within illocutary competence are ideational function, manipulative functions, heuristic functions, and imaginative functions; finally, under sociolinguistic competence are sensitivity to dialogue or variety, sensitivity to register, sensitivity to naturalness and cultural references/figures of speech. While the Bachman model may be useful for descriptive purposes, a checklist is of limited contribution with testing concerns (Skehan, 1998, p. 164). Extending the fissiparous trend even farther, Spolsky (1989) sought to account for every factor relevant to language learning. He identifies scores of factors, and identifies whether the criterion is, for example, typical or not, and necessary or not, for language learning. Similarly, Carroll's (1993) analysis identified 70 different factors that bear on language proficiency. This broad array of different views of language proficiency and the quantity of contributing elements illustrates the scholarly disagreement about a fundamental issue in language research. Building a theory-based language test is problematical with such scholarly disagreement on the constituent components of language. Interestingly, the range of opinions is not unlike the different models of intelligence, wherein Cattell's model offers
two components and Guilford's model includes about 150 factors (Brody, 2000, pp. 20-1).

At the opposite end of the spectrum are found adherents of a general language ability. Perceiving language proficiency in terms of a global factor is an approach identified with Oller (1976/1983) and with Cummins (1980, 1983), which will be examined closely in the following pages. Interestingly, Spolsky (1989) does not dispute the possible merits of the general language ability perspective. In the process of enumerating factors contributing to an overarching theory, Spolsky takes an unexpected stance, declaring, “To say that linguistic and communicative competence are divisible does not necessarily rule out the claim that there is a core of common knowledge of a language underlying the specific abilities of a speaker” (p. 71).

Studies by Oller (1976/1983, 1997) and by Cummins (1980, 1983) are particularly germane to the quality measured by the CLEAR, AR. It is fitting, therefore, to consider their theories closely. First to be examined are Oller's arguments relevant to language and intelligence. Subsequently considered will be Cummins' positions on everyday language versus academic language.

Oller, language, and the ability measured by intelligence tests

Oller is the leading proponent of the position that language ability is unified (Spolsky, 1989, p. 71). Results from a principal component factor analysis led Oller (1976/1983) to formulate the “unitary competence hypothesis,” which is to say that most
variance in language proficiency scores can be explained by one sole factor. This hypothesis is not unlike the models of intelligence that identify one major ability factor, the “g factor.” One such model is that of Spearman (1927, cited in Board of Scientific Affairs of the American Psychological Association, 1995). Indeed, Oller hypothesized that this global factor accounting for language ability is the same factor underlying much of the variance in educational and intelligence test scores. Later studies (Bachman & Palmer, 1981; Carroll, 1983) indicated that the findings were an artifact of the statistical technique. Oller consequently announced (1983, p. 352) that he had abandoned the strong version of the unitary competence hypothesis of one singular ability accounting for language and intelligence.

He continues, nevertheless, to assert the importance of general language ability. Oller (1997) argues that intelligence tests are actually measuring the reasoning ability filtered through the examinee's primary language; he explicitly charges verbal intelligence tests as language proficiency instruments (p. 465). The relationship between language testing and intelligence testing is relevant to the CLEAR validation, particularly since the CLEAR may be perceived as a covert test of intelligence.

Intelligence, in the view of mainstream psychologists, is innate and immutable (Board of Scientific Affairs of the American Psychological Association, 1995). Intelligence tests (IQ tests) purport to measure innate cognitive ability that underlies all reasoning and problem-solving (Board of Scientific Affairs of the American Psychological Association, 1995).
Psychological Association, 1995). The trait called “g” has been defined as “an ability that is common to all intellectual tasks and is measured to some degree by all tests of intelligence” (Grigorenko, Sternberg, & Ehrman, 2000; cf. Board of Scientific Affairs of the American Psychological Association, 1995).

Supposedly, g is tapped whether a verbal or non-verbal intelligence test is used. Oller (1997) argues that both test types require language proficiency, and both measure language proficiency. The importance of language seems evident in measures of verbal ability. Verbal intelligence test items would include defining words, identifying a word that does not belong with others in the set, and completing analogies (Brody, 2002; Board of Scientific Affairs of the American Psychological Association, 1995; Gottfredson, 1998). In the case of NNSEs, intelligence tests may actually be measuring English proficiency when given to speakers of another dialect or language (AERA et al., 1985). NNSEs are recommended to take IQ tests that have been translated into their L1 or to take a nonverbal intelligence test (AERA et al., 1985; Mainstream science, 1994).

Nonverbal tests of intelligence include such tasks as classifying and completing a series; similar activities are employed in the “culture-fair” tests of intelligence, such as that constructed by Cattell (Brody, 2002, p. 21; Cattell, 1940; Oller, 1997, p. 489). Nonverbal and “culture-fair” IQ tests are claimed to be more fair than tests obviously relying on language. “Culture-fair” tests are intended to measure cognitive potential without regard to verbal, cultural, or educational experience. Test-takers include NNSEs,
those with poor literacy skills, and those from disadvantaged backgrounds. Indeed, developing a culture-fair test wherein the disadvantaged score highly has “proved to be a chimera, for the obvious reason that educational tests measure a comfortableness in the classroom that usually doesn't go with growing up in difficult circumstances,” states Lemann (1995, section 2, para. 9). Indeed the very term “culture-fair” is an ideal which is not actually realized: every test is influenced by culture (Sternberg 2000; Sternberg & Grigorenko, 2002).

Assuming that nonverbal tests are language-free is logical, but erroneous, argues Oller (1997). Nonverbal intelligence measures ultimately depend on language, and this charge can be laid upon not only the early attempts to measure mental ability without dependence on language or literacy skills, but also prominent contemporary tests, such as the Raven's Progressive Matrices (1965) and Cattell's Culture-Fair Intelligence Test (1933-1973). As evidence, consider that nonverbal tests typically require predicting the next item in a series or in a matrix.

However, as will be illustrated, nothing is innately evident about completing a series, a row, or a column. Moreover, completing a series is not necessarily free of language, as is evidenced by Figure 3. Sequencing numbers one through five is a simple task; monolingual test developers might not consider the need for familiarity with the linguistic code. The item in Figure 3 merely asks for the test-taker to finish “1, 2, 3, 4
“__, ___” with “5” and “6.” Yet, this simple task is easy only for those persons who know the Chinese language numbering system.

Matrix-completion, like series completion, is without inherent transparency. Humans do not have the innate, untutored ability to complete matrices. Because the items are not self-evident, the tests require explanations and directions, which necessarily rely upon language or conventional gestures. Using pantomime would only be of help with a repertoire of universal gestures or symbols, but such a repertoire does not exist (Oller, 1997, pp. 484-5). Knowing the task, therefore, requires language—even in a purported nonverbal test—and thus, nonverbal tests depend on language.

The above examples help illuminate the importance of language in items used in nonverbal tests of intelligence. Oral directions are also problematical in such measures. Consider the instructions, quoted below, that were given in an investigation with matrix-completion activities with six-year-old children (Siegler & Svetina, 2002). The complex
verbal instructions for a nonverbal test task may account for the ensuing attrition of 11% of the sample of 90 children. To better appreciate the importance of language, imagine that the following explanations are delivered in a language you do not know.

To ensure that children understood the matrix completion task, eight example items were presented at the beginning of Session 1. Each example involved a 2 X 2 matrix with objects in three of the four squares; the bottom right square was empty. The items in a given matrix varied along only one dimension, and the experimenter described each matrix in a way that conveyed this fact. For example, one introductory matrix included a large green fish facing left in each of the two squares on the left and a small green fish facing left in the top right square. The experimenter said, “Look: This is a big fish and this is a little fish [pointing to the top left and top right squares]. Here there is also a big fish [pointing to the bottom left square], and here there is one missing [pointing to the bottom right square]. What do you think this one should look like [pointing to the empty square]?” After the child answered, the experimenter asked, “OK, which of these do you think is correct,” while pointing to the six alternative answers, which were arranged in three rows and two columns. The 10 children who did not correctly answer at least half of these simple items were assumed not to understand the task and were not tested further. (p. 797)

Language could be one issue among the possible causes for the children's lack of comprehension. Even when the speakers and hearers share the same language, this does not guarantee conveying meaning. Perhaps the children could not follow the directions because their linguistic development was not sufficient in terms of abstract, decontextualized language. The attrition would not have occurred in a dynamic test or assessment, because teaching and practice are integral to the test (J. Elliott, 2003;
The language of instruction and directions is a formal variety of language. Such differs greatly from the type of language used in everyday social interaction, is less commonly encountered in low-socioeconomic (SES) homes and more common in middle-SES or higher-SES ones. Decontextualized, abstract language is associated with formal education, and the language of the classroom (Cummins, 1979). McLaughlin goes so far as to say, “The use of decontextualized language in the home is highly predictive of educational success” (1990, p. 163).

The issue is worth examining, not merely to better understand why the children could not participate. Considering this matter permits an exploration of the relationships among intelligence tests, educational tests, language proficiency, and language tests. The CLEAR Test might be thought to be any one of these instruments. Moreover, this discussion may suggest some advantages to the direct sampling techniques employed in dynamic simulation assessment of the CLEAR.

**Cummins, BICS/CALP, and reading.** Cummins (1980) identified two distinct language dimensions, Basic Interpersonal Communication Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). These names indicate the very different nature of discourse; these are situated along two intersecting criteria, the extent of contextual support and the level of cognitive demand. BICS identifies everyday
conversational language, typically conducted in a setting with contextual cues and generally with minimal cognitive demand. BICS is acquired by every normal human as part of the natural first language (L1) acquisition process. Chatting about options for lunch while standing in a cafeteria line is an example of BICS language. CALP, by contrast, is a very different matter. It is a function of literacy development, learned through formal education (Cummins, 1980, 1983). CALP indicates the kind of language used when speaking in abstract terms: the discourse found in classroom settings is usually CALP, marked by little contextual support and increased cognitive demand.

Although issues may seem straightforward, the picture becomes complicated when variables of education and second language acquisition are included (Collier, 1989; Cummins, 1983, p. 123). Minority language children may acquire L2 BICS in only a few years, and that without formal education (Collier, 1989). CALP, however, is a different matter. L2 children require 5 to 7 years of schooling in the L2 for CALP development (Cummins, 1983, p. 123). Evidence for the lag between BICS and CALP development is observed among minority-language children who can communicate acceptably in general conversation yet perform poorly on tests (Cummins, 1980). CALP skills from the L1 may transfer to the L2, posited Cummins, (1983) both drawing on a common underlying proficiency. This cross-linguistic relationship he hypothesized as the Linguistic Interdependence Hypothesis (LIH), wherein a literacy base in the L1 supports academic competence manifested in the L2. Cummins' work has illuminated why minority L2
students benefit from instruction conducted in the L1. Children who are formally educated via the L1 will benefit not only in L1 CALP: their expertise in literacy and reasoning in their L1 are believed to transfer to the L2 and support learning conducted in the L2. This hypothesis connects with the Linguistic Threshold Hypothesis, which will be soon be explored in detail. CALP theory and research on L1 transfer are important issues in L2 testing, for such abilities underly AR. The LIH holds that (with a threshold proficiency in the L2) the learner is free to employ in the L2 tools developed in his or her L1, abilities such as reasoning and L1 literacy skills. The LIH is particularly germane to academic L2 reading.

Academic reading is a critical skill area for students of English for Academic purposes (EAP), such as those who will take the CLEAR. While reading research will be explored more fully later in this literature review, a brief treatment is helpful here, to provide background for discussion of AR. “L2 reading” is an ambiguous term (Alderson, 2000, p. 23), variously comprising both reading ability and language ability. How much L1 ability contributes to L2 reading has been a topic of research interest. One position is the LIH, that L1 literacy will transfer to the L2 (Cummins, 1980; Verhoeven, 1992). By implication, notes Alderson (2000, p. 23), the L2 student will need no reading instruction. Yet, even if literacy training need not be re-taught to the L2 student, the student will need some level of proficiency in the L2 to make sense of the reading (Grabe, 1991). This is
the question Alderson (1984) raised, whether reading in a foreign language is a reading problem or a language problem.

*The Linguistic Threshold Hypothesis of L2 Reading*

Investigating such issues has been the concern of Linguistic Threshold Hypothesis (LTH) research (Alderson, 1984, 2000; Alderson & Urquhart, 1985; Bernhart & Kamil, 1995; Clarke, 1980/1988; Cummins, 1979; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997; Urquhart & Weir, 1998). The LTH asserts that, until an learner has reached a particular threshold level of L2 proficiency, L1 strengths (such as literacy, background knowledge, or metacognitive skills) do not compensate for L2 deficits when the L2 is at a low level. An early examination of this concept was in Clarke's (1980/1988) study. Transfer of the participants' good reading skills in their native Spanish was “short circuited” (p. 119) and did not benefit English reading ability.

The CLEAR does not propose to measure the LTH itself. (The CLEAR, in fact, does not attempt to test reading comprehension in the conventional sense used by L2 or native-speaking populations, as will be addressed in the sections on text-responsible reading and writing.) The LTH is relevant to CLEAR validation research because of its explanatory potential. The test developers posit that, once ESL students have progressed beyond a threshold proficiency level, they will be able to benefit from instruction conducted in English. Below this level, ESL students likely will struggle with the
language of the instruction to such an extent that they cannot process the subject-matter knowledge.

The LTH has found good empirical support (Alderson, 1984, 2000; Bernhart & Kamel, 1995; Grabe, 2000; Laufer, 1997; Laufer & Sim, 1985; McLaughlin, 1990; Ridgway, 1997; Taillefer, 1997). Investigations (Carrell & Eisterhold, 1983; Clapham, 1996) into the moderating variables of background knowledge or content schemata interacting with the threshold is one particularly illuminating line of research. (Although only briefly considered here, the topic is examined more fully in the section on L2 reading research, below.) Content knowledge can help L2 readers comprehend better than would be expected from their proficiency. The relationship between subject-area knowledge and improving L2 reading comprehension, however, is not straightforward. Background knowledge of the subject area might help L2 readers with some degree of proficiency, but such knowledge does not help rank beginners: their comprehension is “short-circuited” (Clarke, 1980/1988) by inadequate L2 ability. The background information does seem to help those who have crossed beyond a threshold of some proficiency level (Clapham, 1996; Ridgway, 1997). Indeed, the threshold is thought to vary not only as a function of the reader's background knowledge, but also according to the topic, or the task demands, in the view of many scholars (Alderson, 1984, p. 20, Alderson, 2000, p. 39, Grabe, 2000, p. 243, Urquhart & Weir, 1998, p. 72).
The possibility of a more stable linguistic threshold is suggested by a study by Perkins, Brutten and Pohlmann (1989, cited in Laufer, 1997, p. 21). Having tested the participants' Japanese L1 reading and English L2 reading, then correlating the comprehension scores with TOEFL scores, the researchers indicate that the L2 linguistic competence threshold might exist at the 375-429 score interval of the paper-based TOEFL. Above this threshold, the researchers suggest, perhaps L2 reading might begin to resemble that in the L1.

If a stable threshold exists, defining it might be possible in terms of vocabulary (Laufer, 1997). Laufer points out (p. 20) that vocabulary (along with understanding the subject matter of a text) is a key component in L2 reading. So, she avers, “The threshold for reading comprehension is, to a large extent, lexical” (p. 21). Evidence for centrality of lexis is empirically documented by a threshold effect at the level of 3,000 word families level (1991, cited in Laufer 1997, p. 23). In another study (1992, cited in Laufer 1997, p. 24), she found additional support for the lexical nature of threshold. Based on data from a university admissions test, L1 reading comprehension, vocabulary, and an English as a Foreign Language test, results showed that even examinees who are able readers in their L1 cannot read well in the L2 if their vocabulary is below a threshold of 3,000 word families. This seems to indicate a distinct point at which L2 reading becomes different.

The research above indicates that above a level of threshold proficiency, the quality of the learner's L2 changes distinctly. Above this point it is arguably possible for
the L2 user to employ individual strengths; the academically ready L2 users are ready to

The L2 users' individual resources may be topical knowledge, affective schemata,
metacognitive strategies, and personal characteristics (Adamson, 1993; Bachman, 2002;
Bachman & Palmer, 1996; Huong, 2001). Another view is that academic success relies
upon academic skills and academic enablers such as motivation, study skills, and
engagement (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna,
2002). Related to academic enabling behaviors is the construct called "work drive,"
which has attracted recent research attention. Ridgell and Lounsbury (2004) note the
"growing interest in the role of a dispositional construct developed by Lounsbury et al.
that appears to be predictive of academic and job performance-work drive: 'an enduring
motivation to expend time and effort to finish projects, meet deadlines, be productive,
and achieve success' "(p. 608). Ridgell and Lounsbury found intelligence and work drive
significantly positively related to grade point average among native-speaking
undergraduate students. Employing resources such as work drive or, more generally,
motivation is helpful for learning academic content for L2 users above the linguistic
threshold. Below that threshold, learners must still strive to attain sufficient proficiency
so that learning may take place.
The Construct of Academic Readiness (AR)

The lexical base evidently is a major factor in the linguist threshold, which in turn marks a distinct difference in the quality of language proficiency above and below this area. Above the linguistic threshold, L2 speakers are able to capitalize on their individual strengths and deploy their particular resources. These students are likely to demonstrate AR, that is, ability to demonstrate content learned acquired through the L2, to a level appropriate for native-speaking students, and having learned under conditions for the same. This achievement relies partly on the linguistic foundation, but also upon academic skills developed in their L1, that is, reading and writing skills, mathematics and science skills, reasoning skills such as analysis, synthesis, and investigation (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002), the importance of particular skills varying according to the instructional event. Also contributing to AR are supporting academic traits and skills. Termed "academic auxiliaries" herein, these enablers are the dimensions of motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies; combining these together reflects research on academic competence and L2 learners (Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Garner & Borg, 2005; Huong, 2001) and investigations into academic achievement among native-English-speaking university students (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002; Ridgell & Lounsbury, 2004). Topical knowledge is another individual resource the above-threshold
L2 user can deploy, demonstrating AR. No longer "short-circuited" (Clarke, 1980/1988) by inadequate L2 ability, the individual's topical knowledge can benefit learning content in the L2 (Clapham, 1996; Ridgway, 1997).

Figure 4, below, diagrams the elements hypothesized to contribute to AR.

Figure 4  
*Elements Hypothesized to Contribute to Academic Readiness*

The contribution of the linguistic threshold to the state of AR appears important. Targeted research is needed to quantify precisely this contribution, however. The present literature review found only studies employing traditional language testing approaches, using static tests of achieved knowledge. The literature search did not reveal any other
LTH research employing a simulation model or dynamic testing, nor any other studies conceptualizing AR as do the CLEAR developers. Despite the utility of the LTH hypothesis in illuminating differential academic experiences among EAP students, the point should be stressed that the present study is not investigating the LTH. The present research is a validation study of the CLEAR, a dynamic simulation to assess AR.

**Dynamic Simulation Assessment**

Dynamic simulation is a hybrid testing approach created by the CLEAR developers. Melded together are the simulation model, which provides verisimilitude and context (Garner & Borg, 2005), and "dynamic testing" (Elliott, J. 2003; Embretson, 2000; Grigorenko & Sternberg, 1998; Grigorenko, Sternberg, & Ehrman, 2000; Haywood & Tzuriel, 2002; Laing & Kamhi, 2003; Swanson & Lussier, 2001; Sternberg & Grigorenko, 2002), wherein instruction is a key part of the test itself. The literature on dynamic testing research can provide context for this CLEAR validation study.

"Wouldn't it be nice if researchers could test people's ability to learn new things rather than just people's ability to demonstrate the knowledge they already have acquired?" ask Grigorenko and Sternberg (1998, p. 75). Offering this information is dynamic assessment, a type of interactive assessment which comprises deliberate teaching and the assessment of the results of that teaching (Haywood & Tzuriel, 2002, p. 40).
A précis of the conceptual basis of dynamic assessment is provided by Haywood and Tzuriel (2002). First, achieved knowledge is not the best predictor of ability to acquire new knowledge (though the two are highly correlated). Second, every one functions at below-optimal capacity, so all have the potential for improvement. Third, a performance sample is the best test of a performance; so measuring learning ability can be tested by using learning tasks, particularly with instruction. Finally, there exist blocks to a person's accessing and effectively using his or her intelligence. These blocks can be, among other things, ignorance, poor motivation, flawed self-concept, inadequate study habits, ineffective learning styles (p. 41). These points summarize the rationale for employing dynamic measures.

Many different varieties of dynamic assessment and dynamic testing exist. A basic distinction may be made between assessment and testing: feedback intervention is, strictly speaking, common with dynamic assessment but not with dynamic testing. Grigorenko and Sternberg delineate the distinction thusly: "Broadly defined dynamic assessment is naturally linked with intervention. In essence, the goal of dynamic assessment is to evaluate, to intervene, and to change. The goal of dynamic testing, however, is much more modest: It is to see whether and how the subject will change if an opportunity is provided" (p. 76). In the present study, however, "dynamic assessment" and "dynamic testing" will be used interchangeably.
With growing criticisms of static testing has come increased attention to dynamic testing research, yet dynamic testing is not new. Indeed, Grigorenko and Sternberg (1998) sketch the connections between notions of dynamic testing and theoreticians of the early 1900s, including Binet and Thorndike, although credit is usually assigned to Vygotsky for introducing dynamic testing to the field of psychology (p. 77).

This measurement approach is applicable to needs in different fields, particularly as an alternative means of ability testing (cf. Ebretson, 2000). Testing learning potential has been conducted with individuals who typically under-perform on static tests of ability, such as members of minority groups, people from disadvantaged backgrounds, people with mental retardation, individuals with brain damage or psychiatric disorders (J. Elliott, 2003; Grigorenko & Sternberg, 1998; Haywood & Tzuriel, 2002; Sternberg & Grigorenko, 2002). Dynamic tests have also been used with new immigrants and with culturally and linguistically diverse students (Gonzales, Castellano, Bauerle, & Duran, 1996; Grigorenko & Sternberg, 1998; Haywood & Tzuriel, 2002; Laing & Kamhi, 2003).

With these special test-takers, dynamic approaches seem effective. Seemingly, the opportunity to practice helps reduce construct-irrelevant variance with these examinees. "Static or traditional approaches to assessment of aptitude typically provide little feedback or practice before testing, and therefore performance on such measures often reflects the student's misunderstanding of instructions more than their ability to perform"

Yet, beyond these special groups, dynamic testing has not been very widely disseminated to a broader general educational audience. Haywood and Tzuriel (2002) state that dynamic assessment is "still virtually unknown to many psychologists and educators" (p. 41). This seems surprising, considering the potential of these measures for educators, clinicians, and researchers. Indeed, Sternberg and Grigorenko (2002) state simply, "If dynamic testing is successful, then it is also revolutionary. We at last have a way to reduce the effects of all the environmental variables that can color performance and hence distort estimates of latent capacity. Dynamic testing may give us a means to quantify a person's true potential for growth, from wherever he or she may happen to be cognitively at any given moment" (p. 29).

Several reasons have been proposed to account for the paucity of attention to dynamic testing. Grigorenko and Sternberg (1998, p. 76) noted that published reports lack sufficient methodological detail to permit replication. They furthermore stated that the novelty of dynamic testing is a reason it has been under-reported (Sternberg & Grigorenko, 2002, p. 31). Elliott (2003) decries the "dearth of systematic and controlled studies that compare the differential impact of interventions based on static approaches" (p.24). Grave weaknesses are that the dynamic assessment literature provides inadequate reliability and validity data (Haywood & Tzuriel, p. 58; Grigorenko & Sternberg, 1998, p.
76), and that psychometric problems are inherent with the basic metric, the pretest-posttest change score (J. Elliott, 2003, p. 24; Snow, 1990 in Grigorenko & Sternberg, 1998, p. 104).

Despite the gaps in the present literature, dynamic testing appears to be a promising measurement approach for the purposes of the CLEAR, providing a "direct approach [that] would simulate the learning situation" (Embretson, 2000, p. 505). The direct simulation of the CLEAR was selected for its "ecological" properties, providing examinees with rich context (Garner & Borg, 2005; Hyland & Hamp-Lyons, 2002; Sternberg & Grigorenko, 2002, p. 121; J. Turner, 2004). The context, moreover, was extended over time and across tasks to become a simulation of the target university experience, where learning opportunities unfolded into content-knowledge testing. This approach is closely connected to the principle wherein "Testing thus joins with instruction and the test taker's ability to learn is quantified while she or he learns" (Grigorenko & Sternberg, 1998, p. 75). In such a measure, the test examiner provides "active and direct teaching precisely in order to produce change. Thus, the basic datum in dynamic assessment is a change variable" (Haywood & Tzuriel, 2002, p. 41).

The dynamic approach does not test the examinee on static, achieved, decontextualized knowledge as is the norm in conventional tests for university admissions for NNSEs. Rather, the assessment is contextualized through the simulation of a university class and AR is tested in active, purposeful activity by learning novel
academic content. Dynamic testing permits investigating whether examinees can perform their central activity in the target environment of the university, demonstrating content learning. Content learning is key, "arguably the primary attribute assessed in many college courses" (Elliot & DiPerna, 2002, p. 10). The nature of academic content learning will be detailed, after a discussion concerning the nature of AR in the ensuing section of this review.

Before progressing to the next section, a brief recapitulation is presented to summarize salient themes reviewed above. Evidence has been presented showing the prominence of language in intelligence tests, even nonverbal measures. Though intelligence is commonly assumed to be innate, research showing the contribution of language proficiency to IQ scores casts doubt on this innatist view. CALP, that ability honed through literacy and formal education, appears to play a part in both IQ tests and language proficiency tests. (Indeed, the factor called “global language proficiency” by Oller and “CALP” by Cummins is the same trait [Cummins, 1980, p. 177]). The language of the classroom, CALP, is developed with years of formal education and can be empirically distinguished from the language of everyday social interaction. The unique nature of classroom discourse also contributes to the demands of the university learning/testing cycle. Since the CLEAR test requires examinees to learn information through instruction delivered in English, a reasonable assumption is that the CLEAR therefore requires CALP. Development of CALP may be that which fundamentally
contributes to crossing the linguistic threshold, and a key factor in this might be extensive
lexical knowledge. ESL students with these abilities may tend to perform well in the
CLEAR, though the simulation model also allows the examinee to bring in individual
strengths and skills. These underlying traits are crucial in the L2 user's development and
ultimate AR, although AR is hypothesized to be demonstrated only above the linguistic
threshold.

AR is posited to be demonstrable through dynamic testing. This alternative means
of measuring learning potential incorporates a learning experience into the test situation.
Dynamic assessment has been effective in measuring ability in individuals not well
served by normative measures. This measurement technique has important potential for
researchers and educators, despite under-representation in the literature and lingering
methodological and psychometric concerns.

AR is the construct that the CLEAR seeks to measure. AR has been
operationalized as the direct, present evidence of ability to learn via the L2 in the
dynamic simulation, learning new subject-area material to a level appropriate for native-
speaking students, and having learned under conditions for the same. AR is thought to be
enabled by a constellation of abilities and behaviors in persons who have above-threshold
L2 proficiency: CALP, academic skills, academic auxiliaries, topical knowledge, and
personal characteristics likely all contribute to the L2 speaker's AR.
**Academic Skills: L2 students’ reading and writing**

The above discussion of constructs may be helpful in improving understanding of issues, and may provide supporting understanding of the construct the CLEAR team has termed “academic readiness,” the construct that the CLEAR seeks to tap. A key point is that, while latent abilities are preeminent in psychological tests, job analysis matters more for certification-type testing (AERA et al., 1985, p. 9, 63). Thus, examining university tasks is germane to university admissions testing of ESL students and to validating the scores from the CLEAR dynamic simulation assessment. This section of the literature review will therefore undertake a job analysis by describing the academic skills of central concern to L2 university students, reading and writing. These two skill areas are important to university students across disciplines at both graduate and undergraduate levels, although the same does not hold for the skills of listening and speaking (Carson, 2001; Rosenfeld, Leung, & Oltman, 2001; Waters, 1996). This research helped shape the CLEAR into an assessment wherein reading and writing are prominent and integrated (Carson, 2001; Esmaeili, 2002; Johns, 1990; Grabe, 2001; Waters, 1996), albeit integrated with lecture listening. In keeping with the focal skill areas of the CLEAR, this review will closely inspect the nature of reading and of writing, and how these skills are required in university coursework in North America.

Seeking evidence of university demands of students, the CLEAR team has relied on earlier research conducted by Flaitz and Zhu (2001). The needs analysis led to the
definition of AR as well as the CLEAR dynamic simulation itself. Underlying AR are what J. Elliott and DiPerna term "academic skills" of reading and writing, wherein these language abilities are used for knowledge acquisition. The reading and writing skills list, reprinted in Table 2, identifies language skill objectives for students in an intensive English for Academic Purposes program.

Table 2  
*Academic Reading and Writing Skills (Flaitz & Zhu, 2001)*

In writing, the student should be able to ...
- synthesize information and write a well documented five-page research paper
- express complex ideas on a topic fluently so that they can be understood by the academic community
- write short and long essay exam answers in an academic testing situation
- organize ideas using rhetorical patterns acceptable in academic settings

In reading, the student should be able to ...
- distinguish fact from opinion and use information to discuss and write
- read extensively for general meaning to summarize for discussion
- demonstrate appropriate use of vocabulary guessing strategies
- distinguish main ideas and supporting ideas
- demonstrate familiarity with the organization of texts and other information sources

The core, requisite, academic language skills comprise academic reading with the purpose of learning and academic writing with the purpose of demonstrating that learning has occurred. The definition continues with addressing other skills (speaking, listening, technology, study, and academic preparation). Covering these topics lies beyond the
scope of the present review, which will focus discussion to areas relevant to the CLEAR dynamic simulation.

**Academic Reading**

Considerable research effort has been devoted to reading, both in L1 and L2 populations (see Alderson, 2000, or Grabe, 2000 for an overview), but much of reading remains murky. Confusion might stem partly from the many factors that contribute to reading; lack of clarity is also due to varied operationalizations of “reading,” as will be discussed later in this section.

**The Nature of Reading**

Little agreement exists concerning the number and identification of factors comprising fluent reading (Alderson, 2000; Urquhart & Weir, 1998). This debate echoes similar disputes, previously reviewed, among linguists (concerning the nature of language proficiency) and among psychologists (regarding the nature of intelligence), whether these constructs are comprised of singular underlying abilities or multiple underlying abilities. As research in this field is extensive, a comprehensive treatment of reading lies beyond the scope of this review. Nevertheless, a discussion of some particulars relevant to the CLEAR dynamic simulation will provide context for the present validation study. Discussion will first consider the different approaches to describing the nature of reading. This review will next sketch a description of the fluent reader, the target for L2 instruction, then proceed to a description of the weak reader, so typical of the L2 reader
trapped in a vicious cycle. Discussion will then turn to three key factors in learning through reading and in reading assessment. These factors are vocabulary, topical familiarity, and reader purpose. Throughout, the review will seek connections between the literature and the present assessment, the CLEAR dynamic simulation assessment.

Establishing the nature of reading is a fundamental issue in the field. Explaining reading tends to separate scholars into two different theoretical approaches. The earlier tradition, which Grabe (1991, p. 379) labels the Components Approach, took a skill orientation to analyzing reading. The Components Approach will be discussed before an examination of the second, more recently developed theoretical approach. What Grabe (1991, p. 383) called the Metaphoric Approaches actually encompasses several perspectives, all related to underlying psychological processes.

The Component Skills Approach to Explaining Reading

Reading is perceived, in the Components Approach, as an ensemble of microskills (Grabe, 1991, p. 379). Identifying these component skills is far from established agreement: some taxonomies account for fluent reading in six or eight skills (Davis, 1968, cited in Alderson, 2000 p. 9), while others identify dozens of skills (Munby, 1978, in Alderson, 2000, p. 10-11). “Recalling word meanings” and “following the structure of a passage” are part of the early list developed by Davis (1968, cited in Alderson, 2000, p. 9). One influential list was that developed by Munby (1978, cited in Alderson, 2000, p. 10). The taxonomy seems a comprehensive framework capable of supporting
instructional materials development and test construction. It includes such items as “skimming,” “recognizing indicators in discourse,” and “understanding conceptual meaning.” Grabe (1991, p. 379) synthesizes identified components and groups them into related categories. His list includes (1) automatic recognition skills; (2) vocabulary and structural knowledge; (3) formal discourse structure knowledge; (4) content/world background knowledge; (5) synthesis and evaluation skills/strategies; and (6) metacognitive knowledge and skills monitoring.

Problems exist with taxonomies of reading components (Alderson, 2000, p. 11). The taxonomy might be based on assumptions rather than data. Another difficulty arises in operationalization, when categories overlap or are confused. For example, “recalling word meanings” is actually knowledge, not a skill. Operationalizing taxonomy items into test items is problematical: judges disagree on what skill the item measures. Furthermore, despite the general popularity of this orientation, the evidentiary support for taxonomies of reading skills is now perceived as dubious. Alderson (2000) summarizes the research in reading skills thusly:

Analyses of test performance do not reveal separability of skills, nor implicational scales nor even a hierarchy of skill difficulty. Thus there are statistical and judgemental reasons for doubting whether skills can be measured separately, or whether subskills of reading can be shown to exist and be related to the ability to answer particular sorts of test questions. indeed whether test questions can unambiguously be said to be testing particular skills is quite unclear .... (p. 49).
The Metaphoric Approach to Explaining Reading

This lack of evidence for the Components Approach may contribute to the interest in the second major theoretical approach to explaining fluent reading processes, the Metaphoric Approach (Grabe, 1991, p. 383). This approach encompasses more recent perspectives of cognitive and educational psychologists; herein, proponents tend to perceive reading as processes of learning wherein people develop these componential skills or knowledge. The Metaphoric Approach is not monolithic, but an assortment of perspectives on reading, where the reading process is exemplified by a metaphor for comprehension. Three major Metaphoric Approaches exist: the bottom-up approach, the top-down approach, and the interactive approach.

The bottom-up approach tends to be text-based, and sees the worth of decoding skills in the reading process. Developing automaticity is important to advocates of this approach (Eskey, 1998). The bottom-up approach in general and reading speed in particular have not received much research attention recently. Nevertheless, rapid, accurate word recognition is a hallmark of good readers (cf. Laufer, 1997, p. 21). Reading slowly (at speeds below 200-300 wpm) can cause comprehension problems by overloading working memory, and this is especially an issue for L2 students, notes Grabe (2000, p. 230). Nuttall (1996, p. 56) estimates that an L1 English speaker of average intelligence and education reads at roughly 300 words per minute (wpm), but this might range between 140 and 800 wpm. Nuttall (p. 56) also offers contrasting examples: A
university student in an EFL setting might read about 200 wpm; where there is little
tradition of reading, speeds might be as low as 40 wpm, even in the L1. Grabe (2000)
summarizes this line of research stating, “Good readers are efficient because they
recognize words automatically, form meaning propositions quickly, integrate
propositional information into a text model rapidly, and restructure the text model to
reflect the main ideas of the text being read” (p. 230).

The second of metaphorical approaches, the top-down approach, can be said to
view the human mind as paramount: understanding entails more than decoding what is
imprinted on the page, and reader background will influence interpretation. This
approach has been associated with the view of reading as a “psycholinguistic guessing
reading model, Grabe (2000) argues forcefully, “is clearly wrong and is not considered
seriously by current researchers. We do not sample texts and hypothesize meaning as the
basic reading comprehension process” (p. 227) (cf. Grabe, 1991, p. 376-8). The top-down
metaphor is informed by schema theory (Carrell, 1998, p. 4), which has been called a
“theoretical metaphor for the reader's prior knowledge” (Grabe, 1991, p. 384). Schema
theory falls short of explaining the process of activating and using prior knowledge from
long-term memory; moreover, it cannot yet predict nor define processes of understanding

83
Third among metaphoric approaches is the interactive one. The metaphoric label has been interpreted as the interaction of the reader with the text or, alternately, as the interaction of top-down comprehension skills with bottom-up decoding skills. The interactive approach is often espoused in contemporary L2 reading instruction programs (Alderson, 2000, Carrell, 1998, Grabe, 1991).

Human factors in reading: good and poor readers

Although these approaches look to generalities to explain reading, also important is sketching a description of the reader, for human factors play an important role in reading. An appropriate beginning is a description of the fluent L1 reader, who personifies “the end point of expertise that an L2 reader is aiming towards” (Grabe, 2000, p. 227), which will soon be contrasted with weak readers. In a fluent L1 reader, reading requires processing efficiency, strategic processing, sufficient language knowledge, knowledge of the world, and time on task; expert reading is rapid, interactive, and purposeful (Grabe, 2000, p. 229). The L2 reader faces additional challenges, as will be addressed below. Reading development programs should enable this student “to enjoy (or at least feel comfortable with) reading in the foreign language, and to read without help unfamiliar authentic texts, at appropriate speed, silently, and with adequate understanding“ (Nuttall, 1996, 31).
In actuality, many L2 readers are dysfluent; their reading is often slow, inefficient, or even self-defeating (Johns, 1991, 170; Nuttall, 1996, 56). Guessing from context is a mark of the weak reader (Grabe 2000, p. 227 and 237; cf. Urquhart & Weir, p. 44). “Context use does not distinguish good readers from poor readers as they are engaged in real-time reading processes, except in cases when poor readers overuse context resources” (Grabe, 2000, p. 237).

Nuttall (1996) views the weak reader as locked in a vicious cycle of infrequent reading, poor comprehension, slow reading speed, and little reading enjoyment. The different causes she dismisses, stating, “It doesn’t matter where you enter the circle, because any of the factors will produce any of the others. Slow readers seldom develop much interest in what they read let alone pleasure. Since they do not enjoy it, they read as little as possible. Deprived of practice, they continue to find it hard to understand what they read, so their reading rate does not increase. They remain slow readers. Somehow or other we must help them to get out of this cycle of frustration and enter instead the cycle of growth” (p. 127).

Worth specific attention are four particular factors in reading that can influence learning and assessment results, particularly as the CLEAR comprises both learning and testing. These four factors are reader vocabulary, content familiarity, linguistic proficiency threshold, and purpose.
**Factors in Reading: Vocabulary**

Vocabulary is key in reading comprehension and fluency development. “No text comprehension is possible, either in one's native language or in a foreign language, without understanding the text's vocabulary,” asserts Laufer (1997, p. 20). Estimates of L1 vocabulary size range from 10,000 to 100,000 words. For the L2 learner, approximating the immensity of educated native speaker vocabulary is a formidable task. In recent trends using communicative methodology for L2 teaching, this task has been largely left to the student: to acquire vocabulary on their own, through reading (Boyd Zimmerman, 1997). The low-L2-proficiency reader is confronted with what Coady (1997, p. 229) calls the Beginner's Paradox: How can students learn how to read when they do not know enough vocabulary to read well? Guessing meaning from context is the usual prescription for this diagnosis. Yet such is not easily achieved. Successful guessing, estimates Laufer (1997, p. 24), occurs only when approximately 95% of the words in a text are known. Such a level of lexical command will require considerable work for the L2 reader.

With an adequate lexical foundation, the reader is sufficiently equipped to guess meaning from context and thus to acquire more vocabulary items incidentally. The weak reader, without a good foundation, encounters opaque meaning and hence blocks to building vocabulary. That “the rich get richer, and the poor get poorer” is the conundrum of the Matthew Effect (Stanovich, 1986, p. 361). This phenomenon, named after the
passage in Matthew 25:14-31, accounts for the tendency of reading to benefit the
development of other cognitive abilities, including vocabulary growth. The Matthew
Effect illuminates the interplay of both language proficiency and reading skill
development in L2 reading. L2 reading is crucially stymied until the readers attain a
certain proficiency or threshold level. As discussed elsewhere in this review, this
threshold is believed to vary according to the individual, the topic, or the task demands
(Alderson, 1984, p. 20, Alderson, 2000, p. 39, Grabe, 2000, p. 243, Urquhart & Weir,
1998, p. 72). Laufer, as mentioned previously, argues that the threshold is crucially a
matter of lexis (1997, p. 21), and has empirically documented a threshold effect at the
1997, p. 24). The factor of vocabulary, then, is seen to be crucial for non-native readers.
Below the linguistic threshold, L2 reading is a struggle. Of little help are L1 literacy and
cognitive abilities; even guessing word meanings is difficult when vocabulary is limited.
Above the proficiency threshold, L2 readers' experiences are akin to those in L1 reading.
L2 readers can deploy their cognitive and metacognitive resources to acquire new
vocabulary, and to compensate for deficits. Students who have worked to exceed the
linguistic threshold have achieved a degree of linguistic richness and will find, according
to the Matthew Effect, that further work repays effort even more richly (cf. Gunderson &
Siegel, 2001).
**Factors in reading: content/background knowledge.** The next reading factor that affects learning and testing is familiarity with the text's content matter, commonly viewed as an important asset for the reader. A conventional belief is that such knowledge can facilitate understanding (Coady, 1997, p. 232), and conversely that knowing nothing on a topic would make text comprehension more difficult. Assessing an examinee's background knowledge is very challenging (Clapham, 2000), as is defining specificity (Alderson & Urquhart, 1984/1988). In test development, text selection is important and should address the aspect of text familiarity (Urquhart & Weir, 1998). Since the source reading is a central stimulus in the CLEAR dynamic simulation, discussing these matters in some detail may prove illuminating.

The degree to which individuals' prior knowledge can support comprehension has been a topic of interest over recent decades. The notion that general world knowledge and cultural knowledge, if activated, can support comprehension and recall is supported by research (see Alderson, 2000, for a review). The scope of the present study, however, does not permit a full discussion of this line of research.

Of particular relevance to the CLEAR validation are studies conducted in the fields of English for Academic Purposes (EAP) that illuminate the interplay of reader knowledge and language proficiency with the text in a testing situation. Several studies in the 1980s and 1990s (Alderson & Urquhart, 1984; Bernhart & Kamil, 1995; Clapham, 1996; Laufer & Sim, 1985; Lee & Schallert, 1997; Mohamed & Swales, 1984, cited in Urquhart
Weir, 1998, p. 63-4; Ridgway, 1997; Taillefer, 1996) investigated the relationship with students' academic discipline knowledge and how this may help comprehension in reading texts related to the test-takers' academic discipline. This was connected with the interest in developing Language for Specific Purposes (LSP) tests, e.g., a test of reading comprehension based on readings in economics for economists or in engineering for engineers.

Some studies found mixed or inconclusive results. Mohamed and Swales (1984, cited in Urquhart & Weir, 1998, p. 63-4) assigned native and non-native speaking postgraduate students in the arts and in science to read a pamphlet on how to set the time on a digital clock and how to set the alarm for the next day. The subjects' performance proficiency was measured in terms of speed, and ranked in the following order: native speaking Science, non-native speaking Science, native speaking Arts, non-native speaking Arts. That the Science students enacted the directions on the pamphlet was despite an overall lower English proficiency compared to that of the Arts students (as measured by the IELTS band scores). The authors considered that results could be due to a familiarity with the field, or with the “genre” of instructions; they interpreted findings to indicate that general English proficiency is seemingly not important above a certain threshold.

Alderson and Urquhart (1984) undertook studies examining reading performance in light of overall L2 proficiency, the readers' academic discipline, the text topic, and the
text difficulty. Although some effect was found for the examinees' background knowledge, the findings were not consistent. Test-takers were postgraduates from four broadly different fields: Engineers, Science/Math, Liberal Arts, and Development Administration and Finance/Economics (DAFE). Students read different types of texts, extracted from materials representative of the discipline areas. In different studies, examinees either completed cloze items, or a combination of cloze and writing short answers, or took components of the English Language Testing System (ELTS) test. Some background knowledge effect was noted. For example, the DAFE performed better comprehending the reading based on their own field in comparison with how engineers performed with the text related to that discipline; similarly, engineers' comprehension on the engineering text was better than that of the DAFE students. Yet the two groups performed similarly on the engineering text in the follow-up study. Their third study, using the ELTS as the dependent measure, found mixed results. On the technology-oriented reading test, the science/engineering students outperformed the business/economics students, as might be anticipated. Contrary to expectations, however, the business/economics students did not perform better than the science/engineering students on the social studies reading test. In other words, despite equivalent text difficulty, understanding was sometimes easier with texts outside the field of expertise.
Factors in Reading: The Threshold of Linguistic Proficiency

L2 proficiency seems an important component in L2 reading (Clapham, 1995; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997). For the low-proficiency Turkish students of Ridgway's (1997) research, for example, reading comprehension was not facilitated by academic discipline knowledge. In a major study of content knowledge interaction with linguistic proficiency, Clapham (1996) found evidence of two linguistic thresholds, not one. One occurred at approximately 60% on her grammar test, while the second one was at about 80% on her grammar measure. Below the lower threshold, examinees' L2 proficiency was so low that any content knowledge could not help them; above the higher threshold, students could read in other disciplines with little difficulty. For students at a proficiency between the upper and lower thresholds, content knowledge provided an advantage. Therefore, the construct-irrelevant variance of subject-matter knowledge contributed to the reading comprehension score, but did so unevenly across proficiency levels. In light of this evidence and other studies of LSP and discipline-specific knowledge, Clapham (2000, p. 514) concludes that language proficiency plays a crucial role in performance on reading comprehension tests, and proficiency seems as important as background knowledge.

In sum, then, research has confirmed the popular perception that the reading factor of familiarity with the text's content matter does affect learning and testing. Indeed, research indicates that the interaction of variables is far more complex than was hitherto
known. Moreover, the route of testing LSP is fraught with difficulties. Clapham (2000) notes that assessing an examinee's background knowledge is difficult, perhaps not evident from an examinee's present activities or field of study. Text selection is a crucial step in test development since examinee topical knowledge is not easily discerned yet it can, at given levels of proficiency, interact with reading comprehension. Specificity of topic is also problematical in reading texts: what may seem specific enough to an outsider is perceived as general knowledge to an insider (Alderson & Urquhart, 1984/1988). Thus, one great challenge facing the test builder is identifying a text of desirable specificity for the reading comprehension stimulus. Urquhart and Weir (1998, p. 143) suggest very practically that texts be selected which do not fall at either extreme of a familiarity continuum and this degree of familiarity be established in field testing the instrument. To control for examinee background knowledge, the test developer can either modify the text or the task (Urquhart & Weir, 1998, p. 115). The route chosen by LSP test builders is to seek a stimulus text particularly germane to the examinee. Adapting the task to include multiple readings is another option. Or, one could abandon the “specific purposes” route and present “general” reading texts.

These reading factors helped to shape the CLEAR team's choice of psychology as the academic subject area for the dynamic simulation. This field is generally little known to university-bound EAP students, in the experience of the CLEAR team. That the examinees are unfamiliar with psychology helps answer questions concerning examinee
prior knowledge and the interaction of such with proficiency. A crucial component of
dynamic testing is that the examinee undertake a novel learning experience; background
knowledge would seem to change the meaning of the scores. The lack of examinee
background with the stimulus helps avoid test bias, supporting fairness in the CLEAR
test, and reducing construct-irrelevant variance.

Factors in Reading: Reader Purpose

Reader purpose is the fourth factor under discussion affecting learning and
testing. Purpose is a central aspect of reading, and relevant to test validity (Urquhart &
Weir, 1998, p. 119). Many of the cited studies do not specify why the examinees were
taking the test, or if they had anything at stake based on the test results. It is unknown
how the test-takers approached the task, whether they had the intent to excel or whether it
was an empty exercise. NNSE students would seem to be motivated to excel, especially
upon admission to the university. Yet, even then, they might not know what it means to
read at the university level. Professors in one study reported that the NNSEs were
unfocused when reading during class, stopping to look up words and seemingly unable to
read strategically (Johns, 1991 p. 170). The laborious reading seems typical of lower-
level ESL students. Similarly, one CLEAR simulation pilot test participant, observed
exhaustively translating during the study session, reported her initial opinion of the text
as, “I can understand it!” The examinee changed her evaluation after taking the CLEAR
knowledge subtest, sadly saying, “I did not know how to read.” She had not prepared for
a knowledge-based test. Thus, even when content learning is at stake, students may continue to view reading as slow, ponderous work merely to develop ESL skills. After years of reading with a dictionary in hand, such students may find it difficult to change their purpose and read for learning.

Whether test-builders are testing reading ability or testing language ability might be confounded in conventional L2 reading tests (cf. Alderson, 2000, p. 112; Urquhart & Weir, 1998, p. 111-2). With the CLEAR assessment, however, the purpose for reading is more transparent: how well can the examinee learn from the text, the lecture, and in using the study sessions? CLEAR dynamic simulation examinees were given two motives to perform well. First, was the possibility of university admission, exempt from taking the TOEFL. Second, prizes were awarded by lottery among the 10 highest-scoring examinees at each site. These measures might have promoted extrinsic motivation to CLEAR volunteers. In addition, previous participants had reported gaining a sense of the university experience through the CLEAR dynamic simulation assessment.

**Reading-to-Write**

Purpose in reading might also be enhanced through the read-to-write task, as Weigle (2003) suggested. The CLEAR's central task of reading as the basis for writing is, in simulation, the university experience of instruction leading to assessment. This differs sharply from the norm in ESL and English-subject classes, which typically use readings
as a springboard for compositions, as different scholars have noted (Carson, 2001; Horowitz, 1986a, 1986b; Johns, 1991; Leki, 1995; Leki & Carson, 1997; Spack, 1988; Weigle, 2003; Weigle & Nelson, 2003). In ESL and English-subject courses, the stimulus is meant to inspire the student-writer or to frame the prompt for the composition. The reading passage need never be referenced. In assessing this kind of writing, including a reading stimulus should be a matter of forethought (Johns, 1991, p. 172). Fundamentally different are the requirements in the rest of academia (Carson, 2001; Horowitz, 1986a, 1986b; Weigle, 2002, 2003; Weigle & Nelson, 2003). Here, text is not a stimulus; text is a knowledge source. Moreover, in the subject disciplines, students are responsible to show reading-derived knowledge in their writing (Leki & Carson, 1997). Intrinsic to academic literacy is the integration of language skills, particularly reading and writing (Carson, 2001; Esmaeili, 2002; Grabe, 2001). Indeed, source texts such as readings and lectures form the basis for academic writing (Weigle, 2003), and such is common across tasks (Carson, 2001) and across different discourse communities (Spack, 1988). This “reading-to-write” is of particular importance for students in EAP programs, who seek preparation for English-medium university study. Although product-oriented approaches may be unfashionable in contemporary reading research (Alderson, 2000, p. 5), discipline-based faculty, nevertheless, expect students to learn from course readings and to demonstrate such in writing.
Academic Writing

The discussion will now focus more closely upon issues related to academic writing per se. The skill of writing holds considerable importance in university settings (Rosenfeld, Leung, & Oltman, 2001), and the importance increases as university students advance toward and beyond graduation (Waters, 1996, p. 18). In the freshman and sophomore years, undergraduates write less than in later years (Johns, 1990; Waters, 1996). At the lower level, university students are generally expected to write at lower levels of Bloom's (1956) taxonomy, demonstrating content in terms of knowledge, comprehension, application, or analysis (Rosenfeld, Leung, & Oltman, 2001). Much undergraduate student writing consists of responding to essay-type examination questions (Carson, 2001; Horowitz, 1986a, 1986b; Weigle, 2003; Weigle & Nelson, 2003).

Understanding the actualities of academic writing for the disciplines is important for validating the scores from the CLEAR dynamic simulation. Hamp-Lyons (1990) identifies four facets that bear upon writing assessment: variables of task, writer, scoring, and reader. These four variables intermesh. Thus, constructing a writing task which resembles its target cannot be achieved in a vacuum. The task must be scored by criteria used in the target situation by readers familiar with the topic and criteria. When these represent the examinees' future conditions, the factors work in harmony, and the examinee can perhaps better demonstrate ability.
To provide context for understanding the CLEAR essay writing task, this review will touch on issues in ESL writing instruction and task demands in academia. Since writing skill requires years to develop (Grabe & Kaplan, 1996), writing instruction for ESL students is critical. A brief overview of the four instructional trends in L2 writing pedagogy, based on Raimes' (1991) historical treatment, may help shed light on beliefs about ESL student needs. These are presented in order of appearance in the field, although new trends do not necessarily supplant earlier ones.

**Historical trends in writing instruction.** A focus on form was a major feature of ESL writing pedagogy dating back to the mid-1960s (Raimes, 1991, p. 408). Evolving out of behavioristic audiolingual approaches, this trend considered paramount well formed sentences within controlled compositions. The mid-1970s saw the appearance of a trend towards focusing on the writer, which seemed an outgrowth of emphases on expressionism and cognitivism (Emig, 1971 and Zamel, 1976, in Raimes, 1991 p. 409). The writer-focused approach gave less prominence to the product, and more attention to the process of writing. Writers might receive considerable time, produce multiple drafts, then collaborate with classmates. Two new movements arose in the mid-1980s, both concerned with content and academic disciplines. Reacting in part to the attention upon the writer, the focus upon the reader was a trend that appeared around 1986 (Horowitz, 1986a, 1986b). This movement was rather directive and considered that ESL students needed an apprenticeship into academic discourse communities (Johns, 1981 cited in
Raimes, 1991; Santos, 1988; Spack, 1988; Swales, 1990 in Grabe & Kaplan, 1996, p. 108). Good writing was considered that which imitated the features of texts found in particular disciplines (Grabe & Kaplan, 1996, p. 138; Spack, 1988). At about the same time a different trend appeared in EAP writing, focus on content. Concerned that instructional emphasis focus on EAP students' needs, Horowitz (1986a, 1986b) was a proponent of this trend, pointing out that university students will need to write essays for examinations far more than they will need to write multiple drafts of personal expression type essays. Although the process approach was not entirely dismissed, this movement reintroduced product into the focus, particularly relevant for writing for the disciplines.

These different approaches to writing instruction might indicate not only shifting concerns, but also some uncertainty about the nature of writing. Illustrated in Raimes' (1983) work are facets of writing to be addressed in instruction. Here appears to be the universe of features relevant to good writing, including writing for academic purposes. Many factors contribute to the central goal, which is effectively communicating ideas. Raimes' factors include the writer's process, audience, purpose, word choice, organization, mechanics, grammar, and syntax. Indeed, faculty across academia seem to value organization, logic, and development in writing (cf. Leki, 1995, p. 35). Nevertheless, despite the appearance of comprehensiveness, Raimes' diagram lacks any mention of subject-area knowledge. Even under the “content” rubric are found only relevance, clarity, originality, and logic. While these are fine qualities in writing,
discipline faculty will expect display of accurate content knowledge (cf. Elliot & DiPerna, 2002, p. 10).

Characteristics of discipline-based academic writing

Major distinctions exist between the demands and standards of writing for subject-English classes and those of writing for the disciplines (Horowitz, 1986; Leki, 1995; Weigle & Nelson, 2001). First, the discipline-based faculty might not perceive their role as teaching students how to write. Content-area faculty might not correct the language problems in NNSEs' writing or might employ more generous standards for these students (Santos, 1988; Schleppegrell, 2002; Waters, 1996). In Leki's (1995) study, subject-English faculty ratings of ESL student compositions were distinctly different from the judgments made by academic discipline faculty and by those of ESL students. For example, though both groups were concerned with rhetorical features of the essays being examined, only the discipline faculty also evaluated “quality of information or argument” (p. 38-9).

Knowledge in Discipline-based Academic Writing

Although discipline-based differences exist in writing tasks and writing styles (Grabe & Kaplan, 1996; Spack, 1988), perhaps more important is the shared value across different fields: student writing for the disciplines must display knowledge. Such is the fundamental characteristic distinguishing discipline-based writing from subject-English compositions. In the midst of the era of process writing for personal expression, Horowitz
(1986) called for attention to the factual component in student writing, stating, “Most academic writing tasks, at least at the university where I work, require students to present data, usually obtained through written sources, according to a fairly explicit set of instructions“ (p. 142). Evidence in students' essays that they have learned basic information is especially important to discipline-based faculty teaching at lower levels (Carson, 2001; Leki, 1999). These findings are echoed by Weigle and Nelson (2001) who noted that in their university's undergraduate classes across history, political science, and biology, “the purpose of most writing in the classes studied was to demonstrate mastery of course content (e.g., facts, theories, concepts) on tests” (p. 121).

*Topic in Discipline-based Academic Writing*

The topic holds a central role. The connection between topic and reading passages and the examinees' background knowledge was raised previously. In writing, also, topic is key (Johns, 1991. Topic is said to account for considerable variability in scoring, and indeed for the writers' response (Hamp-Lyons, 1990). An issue in subject-English writing, topic seems particularly relevant in cases of examinees studying specific-purposes English, such as English for science or business. Clapham (1996) investigated the influence of background knowledge on reading and writing performance of examinees taking the British proficiency test, International English Language Testing System (IELTS). As discussed elsewhere, an interaction was noted between background knowledge and L2 proficiency. High-proficiency students were sufficiently advanced that
their background knowledge was not needed; at very low levels, the inadequate L2 “short-circuited” (Clarke, 1980/1988, p. 119) students' domain background. Yet, at intermediate proficiency levels, seemingly above a certain threshold, students' background knowledge gave them an advantage. (Responding to this research, the IELTS no longer employs discipline-specific modules for testing reading and writing.)

**Writing for a Discourse Community**

Specific-topic testing becomes an interesting option when the test developer attempts to accommodate examinees' areas of specialization (Clapham, 1996; Bachman & Palmer, 1996). Within each academic discipline exists a singular discourse community: applied linguists, for example, abide by norms and rules for communication distinctly different from those of engineers and of computer scientists. Presenting examinees with a specific-purposes test, one constructed for the examinee's discourse community, appears to offer the test of ideal relevance. Yet operationalizing specificity might be impossible: what appears to be a specialization to outsiders may seem broad to insiders (Alderson & Urquhart, 1984). Equating different specific-purposes tests might also be difficult (Alderson, 2000; Clapham, 1996). Classifying the domain is another problem, as Davies (2001) notes: the English of chemistry overlaps with that of medicine, and medical English encompasses pediatrics, immunology, neurology, and so forth (p. 137). Identifying ideal material for each specific discourse community is likely to be problematical (Spack, 1988), particularly when this must also accommodate different
levels of knowledge and linguistic proficiency. The LSP testing route may, in effect, lead
the test builder down a garden path (Davies, 2001). In seeking to offer a test tailored to
the examinee's background knowledge and future goals, a test must be so specific that it
is not feasible: Clapham considers perfect specificity to be impossible because of the
interplay of proficiency level and domain knowledge (Clapham, 2000).

Discussion

The CLEAR dynamic simulation assessment represents a departure from many
language tests. The CLEAR developers employ a dynamic approach to testing,
incorporating both learning and testing in order to assess examinees' ability to benefit
from instruction delivered in English. The simulation aspect of the assessment reflects
that the CLEAR is intended to bear close congruence with the target environment.
Acknowledging that the CLEAR might be accused of functioning as a covert intelligence
test, this review has reported arguments that many intelligence tests covertly measure
language proficiency, at least in part. The real problem with many intelligence tests,
however, is the same problem as with proficiency tests: they measure static, achieved
knowledge. Language proficiency has been modeled in various ways, with different
numbers of elements postulated to have different importance. Likewise, models of
reading and writing have attempted to describe and explain these abilities. In these arenas
of overall proficiency, of reading, and of writing, testing can measure atomistic elements of language.

However, no matter how exhaustive the models of L2 proficiency and models of academic literacy skills, no theoretical framework seems to capture how human abilities operate together in fluid situations. Indeed, language performance is synergistic: the whole is greater than the sum of the parts. Dividing up language into more elements cannot explain the whole of language performance. The problem is similar to that experienced in attempts to model reading and writing: product-oriented approaches have limited utility in explaining processes. To account for the harmony created by diverse human abilities variously working together requires a different approach, describing language ability by capturing language in use. In this way, a description can take into account the compensatory nature of human abilities under real world demands. NNSEs in the English-medium lecture hall have an arsenal of weapons at their disposal. Students might variously deploy their personal characteristics, their topical knowledge, their CALP development. As well, students possess skills developed first in their L1, skills in reading/writing, in mathematics/science, and in critical thinking. NNSEs moreover possess academic auxiliaries, motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies. These abilities, elements which contribute to AR, might not be evident in a static test lasting a few hours.
A dynamic test offers different conditions, for the examinees learn new material as part of the assessment, then are tested on their knowledge acquisition. Dynamic assessment might offer examinees more fairness in the test experience. Students' performance might be more representative of their true abilities when the test has engaged a broader and more representative range of the examinees' abilities. Thus, although language proficiency tests measure on an indirect representation of what examinees know about language, the CLEAR dynamic simulation was intended to assess how examinees use language, in learning academic content. Investigating the measurement quality of the CLEAR, the present research, might allow the test developers to interpret and use scores appropriately for determining university admissions recommendations.
CHAPTER 3

METHODOLOGY

Introduction

The present research is intended as an initial validation study of the CLEAR dynamic simulation assessment. This measure is a criterion-referenced dynamic simulation intended to classify non-native English speakers in terms of AR. This construct is defined as the direct, present evidence of ability to learn academic content via the L2 as demonstrated during the dynamic simulation, learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions similar to the same. AR indicates that ESL examinees' second-language (L2) proficiency is above the "threshold" level, and therefore sufficiently advanced to allow the examinees to allocate resources to learning new subject matter information transmitted in the L2.

Sufficiently advanced proficiency appears necessary from L2 reading research (Alderson, 1984; Clarke, 1980/1988; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 105
1997; Ridgway, 1997; Taillefer, 1996). These studies report the importance of a foundational or "threshold" level of proficiency, particularly in terms of lexis, in order to comprehend L2-mediated content, especially that of an academic nature. (This line of research is considered in more detail elsewhere in this study.) One may say that AR in ESL students is constructed upon L2 proficiency, for with this sufficiently advanced proficiency, the examinee is able to allocate resources to the language use situation.

These supporting resources are thought to comprise the elements of personal characteristics (age, educational background, etc.), topical knowledge, academic skills (reading and writing skills, mathematics and science skills, reasoning skills such as analysis, synthesis, and investigation), and academic auxiliaries (motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies) (Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Cummins, 1979; Elder, Erlam, & von Randow, 2002; Huong, 2001; Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002; Ridgell & Lounsbury, 2004). L2 speakers who are academically ready thus possess above-threshold proficiency and are free to deploy these resources in the subject-area classroom, while the less-advanced ones would need to devote all resources to processing the linguistic input.

AR indicates that an examinee is able to benefit from content instruction conducted in English. That is, the L2 speaker demonstrates learning new subject-area material to a level appropriate for native-speaking university students, and having learned
under conditions similar to those in the university learning experience of native speakers. Although a threshold level of language proficiency is a necessary condition before the learner is able to learn via the L2, the language level itself is not the measurement focus, nor does the CLEAR seek to elicit L2 deficiencies or strengths. As a simulation of the university experience, the CLEAR dynamic assessment seeks the demonstration of learning achievement via English.

Recall that the CLEAR had been developed by the author of this study, coordinating the committee, along with two professors of applied linguistics and two teachers of pre-university ESL. The CLEAR test development committee had already undertaken steps appropriate to the initial stages of validation (cf. Urquhart & Weir, 1998). That is, the team had performed an initial examination of the target situation needs and of the literature, steps which comprise the stage of construct specification. The second stage consists of *a priori* validation, then *a posteriori* validation. In *a priori* validation, the CLEAR team spent two years in such activities as selecting the appropriate text, determining the test format, constructing rubrics, deciding on the timing of the test, constructing test items, ordering the items, producing and revising the initial test, anticipating the passing standards, and identifying score report information. These activities, conducted with two pilot administrations, laid the foundation for revising the CLEAR and for initial statistical analysis. Using the revised CLEAR and the improved
rubrics and administration procedure in the Fall 2003 pilot refined the test in preparation for the present research study.

This validation study entailed investigating three major questions, following preliminary concerns. The preliminary matters were determine sample equivalence across testing sites, test functioning across subgroups, and conducting item analyses. Below are listed the major research questions. Following thereafter is a description of the participants, instruments, materials, procedures, and analyses for the research.

**Question 1:** What is the consistency (i.e., the reliability) of the scores derived from the CLEAR dynamic simulation?

**Question 2:** What is the validity of the inferences based on the scores derived from the CLEAR dynamic simulation?

**Question 3:** What is the examinee perception of the CLEAR?

**Participants**

Volunteers were recruited from the English for Academic Purposes (EAP) institutes affiliated with two accredited state universities in the southeast. Before discussing the participants of the study, some examination is appropriate regarding the similarity of the two language institutes and the cities from which the participants were drawn. These programs, their host universities, and their cities, are arguably similar enough to warrant including students from both schools as samples drawn from a similar
population, as will be considered in the following section. The data below were extracted from the institutes' public internet pages and from a demographic data website (Sperling, 2003) for comparing U.S. cities.

**Comparison of Participants' Contexts**

The host institutes share much in common. Both language institutes are established programs: one was founded in 1978, and the other in 1987. Both institutes have earned the prestigious accreditation by Commission on English Language Program Accreditation (CEA). Such accreditation is optional, not mandated by an outside authority; the rigorous process means that CEA-accredited institutes are of the highest caliber. Both belong to a state consortium for EAP institutes. Membership in this association is for programs whose instructional focus is on teaching academic English and test preparation, preparing ESL students for English-medium university study. Students are divided into five levels at both Intensive English Programs (IEPs) based on placement test and essay scores. Both institutes report that they employ a communicative approach to English instruction, offering student classroom and language lab learning modes. The two institutes are similar in the cycle of instruction, both following a 14-week semester. The two schools offer an "intensive" curriculum. “Intensive” English can be defined as 20 or more hours of weekly instruction minimum. The two programs offer students a good variety of extra-curricular activities, including sports, cultural, service, and social events. Both schools provide students with academic counseling to assist them
in applying for university admission. In financial matters, the institutes are fairly similar. Tuition at both locations is about $3,000 per semester, which does not include books, housing, board, or medical insurance.

Examining the geographic area helps inform the argument for the student population similarity. The southeastern state where both IEPs are situated draws more IEP students than any other state except California (Open Doors, 2003). Climate may draw the students: one city averages 238 sunny days per year and the other, 242, while the national average is 213. In January, the average low temperature nationally is 26.8º Fahrenheit, yet the cities of interest experience an average of 50.1º and 50º, respectively. The population of the first metropolitan area is 276,027, larger than the second's population of 182,685. Compared to the national median age (35.5), both of the cities are somewhat more youthful (the first city, 34.08; the second, 31.96). These features are representative, but are not the only shared characteristics. Across many measures, the two cites are quite similar. A comparative report follows this study (see Appendix A).

Participant Profile

The students attending these EAP programs come from diverse home countries. They come to Florida from South America, especially from Brazil and Venezuela. Asian students hail from Japan, Korea, Taiwan, and Thailand. The Middle East is well represented with persons from Kuwait, Saudi Arabia, and United Arab Emirates. Many of these English language learners are Muslims; during Ramadan, the lunar month of
fasting, the religion variable could affect scores earned on the CLEAR.

Similarities may be noted in other background factors. The international students in the United States tend to be advantaged in socioeconomic terms. The majority of EAP students are men, but a substantial part of the population is female. Adults in their 20s are the most common, though other ages are also represented. Almost without exception, these English language learners are well educated: some have already earned a college degree or graduate degree in their home country. All of the students speak at least one other language, of course, and many of them are fluent in several languages.

Participants are similar in proficiency level as well. Since level of proficiency is a variable in this study, it is appropriate to clarify what is meant by proficiency levels. Institutions administer placement tests to new students, often using a commercially available placement test such as the Michigan Test of English Language Proficiency ([MTELP], University of Michigan, nd) or the Comprehensive English Language Test ([CELT], Harris, 1985). The resulting scores help in assigning students to particular levels. Some institutes have only four levels while others may have five or six. At the first IEP, for example, administrators formerly used CELT scores to group students into four levels; the same institute currently uses the MTELP and divides students into five levels. This can lead to confusion, for "Level IV" is the highest level at one institute but the second-highest level at the other. This study will employ cover terms, calling the highest proficiency level, "Level Z" and the penultimate level, "Level Y."
The current classification at Institute #1 is represented in Table 3, where the proficiency levels are divided according to MTELP and CELT scores. Of particular interest are students at roughly the same proficiency as those in the IEP's Level Y and Z. Level Y students have scored between 66 to 75 on the CELT or 60-79 on the MTELP; Level Zs earned 76-100 and 80-100, respectively. The second IEP, however, uses the institutional TOEFL for placement purposes. Their highest level is restricted to students earning 480 or higher. This institute did not detail the TOEFL score breakdown for other proficiency levels.

<table>
<thead>
<tr>
<th></th>
<th>MTELP</th>
<th>CELT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level A</td>
<td>0-34</td>
<td>0-29</td>
</tr>
<tr>
<td>Level B</td>
<td>35-47</td>
<td>30-44</td>
</tr>
<tr>
<td>Level C</td>
<td>48-59</td>
<td>45-65</td>
</tr>
<tr>
<td>Level Y</td>
<td>60-79</td>
<td>66-75</td>
</tr>
<tr>
<td>Level Z</td>
<td>80-100</td>
<td>76-100</td>
</tr>
</tbody>
</table>

Only individuals at higher levels of English proficiency were invited to participate in this study. Below a certain threshold of proficiency, students are probably not academically ready, unable to succeed in subject matter learning due to limitations in processing the L2. Indeed, according to results from the second pilot test in Spring 2003,
some Level Y learners have crossed this threshold and are able to benefit from instruction in English. Most of the Level Y participants, however, were not successful in pilot trials of the CLEAR, while a greater proportion of Level Z students succeeded. The trial administrations of the CLEAR were given to small groups. However, this pattern of distinction between the top two levels was not replicated in the present research. In results from the knowledge subtest and the essay subtest, mean scores between students at the highest level were not statistically significantly different from scores of students at the penultimate level. (However, such might be due to a lack of statistical power rather than to a lack of actual difference, and bears further study.) Finding similar performance from both proficiency levels lent support for the decision to offer the CLEAR at both of these levels, for these students are the target examinee group. In terms of sampling, drawing from both proficiency levels increased the sample size and greater heterogeneity, factors contributed to meeting the assumptions of statistical procedures. Drawing upon both levels and sites, 36 volunteers participated in this validation study, 16 females and 20 males. From Site 1 came 20 participants and from Site 2 came 16 volunteers.

Instrumentation

*General Description of the CLEAR.*

The CLEAR was developed for the purpose of university admissions recommendations. The test was never intended for placement, diagnostic, or achievement purposes, and very likely such information cannot be extracted from the present version.
Not a conventional admissions test for L2 speakers of English, the CLEAR attempts instead to measure ability to learn through the medium of English. This is done by sampling performance in a simulation environment. Thus, test-takers undergo a series of experiences that resemble real-world events. The test tasks replicate, within the constraints of test administration, actual academic student endeavors. The examinees read from a college textbook, attend a lecture on video, take lecture notes, study their notes and readings, then demonstrate subject-matter learning in a multiple-choice test and writing an essay. These subtests attempt to tap ability to learn in English. Both components require the examinee to demonstrate content-area learning. Nevertheless, a caveat should be noted concerning the limited range of academic difficulty in the CLEAR, and the limited quantity and breadth of tasks. This suggests a need, despite years of careful test development, for continuing test revisions. As a dynamic simulation of the university learning/testing cycle, the CLEAR assessment takes place over time. This feature may be viewed as a benefit to examinees, allowing them to employ in the test experience a fuller range of skills than they would be able to in a three-hour session. The same feature of extended time may, however, be considered a limitation by the participants who undergo a 48-hour test.

In this dynamic simulation assessment, testing events must be based on a particular instructional topic. Indeed, selecting the content area was an important issue for the test development team. Examinees might be advantaged if they have subject-matter
background knowledge (Clapham, 2000), yet knowing the extent of an examinee's subject familiarity is difficult. Awareness of the background knowledge factor motivated the CLEAR developers to select a field that is relatively little known to IEP students, psychology. This discipline is uncommonly studied in the Asian, Middle Eastern, and South American home countries of the IEP students. Indeed, no CLEAR examinee has ever reported a background knowledge of Psychology. Another advantage of psychology is the linguistically intensive nature of the discipline. Such may be usefully contrasting with fields where meaning might be conveyed symbolically as in Mathematics, or graphically, as in Art, or kinesthetically as in Dance. Recall that a threshold of linguistic proficiency underlies AR. If an examinee demonstrates AR, ability to learn via English, in a novel and linguistically demanding discipline such as psychology, that examinee would seem likely to have the resources to manage learning where linguistic demand is lower. Moreover, the CLEAR assesses AR by dynamic simulation, where examinees learn under conditions similar to those in the university learning experience of native speakers. The quality of verisimilitude provides the third rationale for selecting the field of psychology. Undergraduate university education in the United States typically requires that students take Social Science courses, among which psychology is quite popular. After consulting with a psychology professor at a research university, the test developers selected a topic concerning family relations for the first version of the CLEAR. (The precise test topic cannot be disclosed due to security constraints.) In time, the test
developers will construct alternate versions of the CLEAR, perhaps using other topics in psychology. The professor of psychology further assisted by recommending a textbook and delivering a lecture on videotape. (These stimuli are detailed below in the Materials section.)

**CLEAR Dynamic Simulation Knowledge Subtest**

The stimulus materials are the basis for two test components of the dynamic simulation, the knowledge subtest and the essay subtest. The combination of these two components, with selected-response and constructed-response item formats, was a combination intended to measure performance in formats typically encountered in university courses. The knowledge subtest contained 32 selected-response items, and examinees were allowed 45 minutes to complete this test. This represented a reduction from the initial knowledge test. The original knowledge subtest had about 50 items, based on two texts and the video lecture, and items included constructed response formats.

Written from the perspective of a content-area instructor, CLEAR items were developed to cover all content areas and fall cognitive levels according to Bloom's (1956) taxonomy. The items do not attempt to elicit L2 deficiencies, as this does not meet the test purpose. Items were written and vetted by team members, all of whom had participated in an item-writing workshop. The workshop was developed by the CLEAR Coordinator, based on her expertise in item writing and on principles derived from the
literature (Crocker & Algina, 1986; Osterlind, 1997). Basic statistical procedures were not possible during the pilot phase of test development, because of sample size and other restrictions. Thus, item analyses were conducted for this dissertation as preliminary analyses prior to the major research questions. (The essay stimulus, prompt, and scoring tool are discussed below.)

Two pilot trials made it evident to the test developers that changes were needed. The knowledge test, like the overall early CLEAR, needed streamlining. For example, scoring the eight constructed response items of the first CLEAR took almost as much time as scoring all of the 40 multiple-choice ones. This was quite inefficient, requiring 500% longer scoring time for the constructed-response knowledge items versus the selected-response items. Another eliminated factor was the secondary reading text, which stimulus did not yield many test items. Eliminating these components, the development team agreed, would simplify and streamline the CLEAR. The result yielded the present version of the knowledge test of 32 items. As stated elsewhere, these items do not seek to elicit linguistic deficiencies or strengths. The items test content knowledge of psychology, such as important persons and themes, taken from both the videotaped lecture and the university textbook. The items were written to represent a broad range of levels of Bloom's (1956) taxonomy. (Regrettably, the actual knowledge test cannot be appended. Since this dissertation is required to be posted on the World Wide Web, such access would breach security for a live test.)
The developers piloted this knowledge test in Fall 2003, scheduled for 45 minutes on the Friday of CLEAR Testing Week. (More information on the scheduling is in the section for Procedures.) Possibly this duration could be shortened for future administrations: most examinees had answered all 32 items within 20 minutes. An alternate interpretation is possible: perhaps the Fall 2003 examinees needed little time because of their below-threshold performance. This group generally performed poorly, with knowledge test scores ranging from a high of merely 67% to a low of 37% correct. This weak performance was congruent with the ability judgment made by the ESL experts, both the examinees' classroom teachers and the ESL professionals on the CLEAR team.

A test of only 32 items is suspect in terms of reliability. Even before the results of the present study were known, the item pool underscores the need for test improvement. A longer test will certainly improve the consistency of scores. The CLEAR Coordinator has called on team members to write new items for the knowledge test and rewrite abandoned problem items: these can be easily trialled in a future version of the test. With this work presently underway, the test cycle continues to proceed and does not suffer unnecessary delays. More statistical information about the CLEAR is found in Chapter 4, reporting results of data analyses.
CLEAR Essay Subtest

College students are not tested by multiple choice alone: they must also write essays. This is certainly true in English composition classes, the so-called "English 101" courses typically compulsory in basic general education requirements. Writing in freshman English courses perhaps no longer focuses on the classic five-paragraph essay, but the writing does tend to be personal, descriptive, or narrative; these rhetorical forms are also found in the ESL classroom. In such courses, “content does not have to be correct or accurate” (Leki & Carson, 1997; cf. Santos, 1988). The attention to language practice, divorced from truth and reality, can have disturbing implications. Smith (1999) recounts the consequences of teaching a foreign language with attention only to language practice, not factual accuracy: His students began to remark that it was “OK to lie in French” (p. 14).

After completing freshman English courses, university students continue to write, encountering assignments with very different demands in the content-area courses for electives and the major. The writing is now text-responsible: it must now show the student's knowledge of the subject, and do so with accuracy. Indeed, Weigle (2002) states plainly, “Accuracy of content is primary in text-responsible writing, and a good deal of research has shown that content area faculty are predominantly concerned with content accuracy rather than linguistic control” (p. 189). The rhetorical form required by the task may also differ in knowledge-based writing for the academy. In many disciplines, student
writers do not need to express personal feelings: rather, they must take a position and defend it. The CLEAR essay assessment is an attempt to replicate this content-bound writing.

The essay prompt, changed little since the first pilot administration in the Fall of 2002, required examinees take a stand on the importance an issue in family relations. The argumentation genre is so common across academic disciplines that Johns (1991) considers it one of the bases for an academic literacy test. The present version of the prompt (not reproduced to maintain security of a live test) improved the contextualization, following guidelines in the writing assessment literature (Blattner & Frazier, 2002). For example, the writers were asked to imagine that the lecturer on the video was the instructor and to write an essay as a midterm examination for this psychology course. Another change from the original administration was making more explicit certain elements of the essay guidelines. The examinees were told to show their new knowledge in organized writing, with appropriate citations from the stimulus materials. The directions were also more clear concerning issues for test administration: for example, examinees should wait at their desks while the assistant brings the essay printout to the writer. (CLEAR examinees have always been given the choice of composing in handwriting or word processing; to date, every examinee has opted to use a computer.)
Perhaps the most important changes were modifying the expectations for greater congruence with real world demands. The changes related, firstly, to time and, secondly, to writing length. The CLEAR standard originally had required three to four typed pages for an acceptable essay, written within two hours. The essay standard now required two typed pages (a 50% reduction), written within an hour and a half (a 25% reduction).

Concerning time, the CLEAR essay session had originally lasted two hours. The test developers shortened the time period based on observation and research. Observation of pilot volunteers showed that the last 30 minutes were of little benefit. Many examinees, in fact, had stopped writing by this point. Fatigue was perhaps a factor, for the essay follows the knowledge test and marks the culmination of 48 hours in a dynamic simulation assessment. Research findings also helped motivate the shorter time allotment: Using relatively short, in-class writing is important in discipline-based writing, a finding of earlier research that Weigle and Nelson (2001) confirmed. Time-limited writing represents a sharp departure from the leisurely process approach employed in many ESL programs (Horowitz, 1986a). Lack of experience with timed in-class writing may explain some of the complaints by CLEAR pilot participants that 90 minutes was too short to complete an essay.

Writing length is the second aspect of the CLEAR essay that had been modified. Requiring the two-page essays within an hour and a half offered congruence with the actual expectations of academic discipline examinations (very different from TOEFL
writing, where examinees have 30 minutes to respond to a decontextualized prompt). In the university context, though some assignments are “take home,” examinations are normally held during class meeting times (Carson, 2001; Horowitz, 1986a; cf. Esmaeili, 2002). Final examinations at the first university test site, for example, are limited to a two-hour period. (More details on the essay can be found in Appendix B, the essay subtest guidelines.)

*Essay Scoring by Expert Judges*

The essays were scored by an expert in the content area, along with the researcher, together using a purpose-built scoring tool. These three factors—SME raters, group rating sessions, and the CLEAR essay scoring tool—result from pilot experiences. Factor number one addresses the rater: the expert was recommended by colleagues in the psychology department. She was hired to consult with the researcher for an honorarium. The consultant and the researcher discussed the CLEAR scoring tool criteria and technique, prior to rating practice essays, then the actual research essays. This topic is covered in more detail later. Factor number two concerns the conditions for the essay scoring: Scoring essays as a group is key because of the disparate ratings in the Spring 2003 pilot, when CLEAR team members were allowed to rate the essays individually and at leisure. Problems emerged, for raters were basing judgments on different factors; some CLEAR members, experienced in grading ESL essays, struggled with scoring content-responsible essays as if discipline faculty (cf. Leki, 1995; Santos, 1988; Schleppegrell,
2002; Waters, 1996). The convenience of individual scoring was abandoned in the last pilot in Fall 2003; together, the team reached better agreement, for members assembled to score papers simultaneously. Group scoring was particularly important in the present study, since the consultant data contributed to the findings. Factor number three, the CLEAR essay scoring tool, will be discussed in detail in the ensuing section.

CLEAR Essay Scoring Tool

The raters' improved decision consistency can be credited not only to the group scoring process, but also to the new scoring tool. The original essay rubric (attached in Appendix C) was not without merit. Created with meticulous attention, the rubric covered numerous discrete elements of language at the word, sentence, and discourse levels. It was a good reflection of the original purpose of the CLEAR, i.e., to test students' linguistic and cultural knowledge. Over time, however, the CLEAR development team ultimately came to a different view of the test purpose: simply to distinguish between individuals who are likely or unlikely to succeed in university work. This likelihood of success is to be based upon demonstrated ability to learn new subject-area material to a level appropriate for native-speaking university students, and having learned such under conditions similar to those in the university learning experience of native speakers. In a simulation of this experience and in a classification test, it is no longer important to examine every possible discrete linguistic element. In brief, the CLEAR was served by the original essay rubric, but not optimally so.
The CLEAR essay scoring tool represents a sharp departure from the typical rubric used in holistic essay scoring. In such, the rater identifies the group into which the writing best fits. Various factors can help improve reliability: few bands in the rating scales, an even number of bands, rater training, rater monitoring, and rating speed. Apparently high reliability, however, can mask disparity. One rater might have classified a paper as "satisfactory" because the introduction and conclusion were strong; another rater might have assigned the same rating, but did so because the citations did not warrant "superior" status.

The CLEAR team was pointed in the direction of a two-stage scoring process, after a review of recent research (Haswell, 1998; Meiron & Schick, 2000; North & Schneider, 1998; Norton, 2000; Santos, 1988; Tedick & Mathison, 1995; Turner, 2000; Weigle, 2002). The scoring tool functions like a flow chart, channeling the rater's thinking in the process of making score decisions. A screen shot of the scoring tool appears in Figure 5. (More legible is the full-sized copy of the document found in Appendix D.)
First, an essay is scrutinized for evidence of meeting primary criteria. These criteria were identified by a literature review (Norton, 2000; Watanabe, 2001) as
important in information-based essays, and were moderated through committee
discussion. An effective CLEAR essay (1) responds to the prompt; (2) includes accurate
and sufficient subject matter; (3) demonstrates appropriate interpretation of the text; (4)
refers to source materials; and (5) takes a clear position, whether explicit or implicit. If
the essay clearly meets those criteria, it will ultimately be scored a 3 or a 4; if it blatantly
fails to meet the criteria, the essay will be marked a 1 or a 2. If the rater is unsure whether
the essay meets or does not meet the criteria, the essay is considered "marginal," and will
eventually be scored a 2 or a 3.

The decision to go with the higher or lower of the two scores depends on how
well the essay meets the secondary criteria. This is the second step of the two-step
classification. The secondary criteria, like the primary ones, were determined on the basis
of the literature and expert judgment. These factors are organization and development,
relevant support for position (whether examples, links, connections, arguments, reasons,
logic, or citations), coherence and unity, clarity and comprehensibility, focus on task, and
language control, variety, and register. All of these contribute to effective knowledge-
based academic writing, yet are not as crucial as the primary criteria. An outstanding
disciplinary essay might well possess all primary and secondary qualities, meriting a
score of 4. Such an essay would seem appropriate for an applicant of a graduate program.
(Note, however, that this is a suggested standard for a passing mark. Setting actual
passing scores properly requires a research study.) An essay that is strong in primary
objectives but less so in secondary ones will be marked a 3; this might be an appropriate passing mark for an applicant to an undergraduate program. If an essay is only marginal in primary criteria, earning a 2, it might still merit generally successful score of 3 by virtue of outstanding secondary objectives performance. However, if an essay is deficient in primary criteria, an above-2 score cannot be salvaged through secondary criteria, for these are only capable of slightly moderating the final score.

The CLEAR Essay Team, a subset of the entire committee, found improved agreement through using this scoring tool in practice sessions. The trials required raters to assess essays from the Fall 2002 and Spring 2003 pilots with essays where there had previously been consistent and divergent scores. Using the new scoring tool instead of the original rubric, raters now focused on the same qualities and better concurred on ranking the qualities and the essays. This agreement, moreover, came with efficiency. Whereas this agreement may arguably be a function of the Essay Team's shared vision, the present study found harmony between the Team scoring, that of the researcher, and that of the content area expert, as will be reported in Chapter 4. However, not addressed at present is the critical quality of decision accuracy: determining the accuracy of pass or fail decisions is for the future, after investigation into setting the passing score, perhaps even after a predictive study.
Materials

This section will detail test materials, the text and videotaped lecture that provide context for this knowledge-based dynamic test. Later in this section are discussed the research materials, the instruments employed to collect data from examinees, raters, and examinees' teachers.

Description of Text Used in the CLEAR

Recall that the CLEAR test is given in the context of an undergraduate psychology course for non-specialists. Although the actual stimulus, like the test itself, cannot be appended because of security concerns, a description may prove helpful. The text used for the CLEAR is a one-chapter excerpt from a textbook commonly used in the introductory course. The chapter is 19 pages long; it is visually pleasing, with charts, photographs, and graphics on many pages. The chapter features introductory questions and chapter-end summaries. Each examinee receives a photocopy of this text which is placed in the individual's designated manila folder. Examinees are free to mark, highlight, and annotate the photocopies as they wish; writing paper is also included in the study folder.

The psychology professor reported that he had found the readability level appropriate for undergraduates. Nuttall (1996, p. 175) states that indices of readability are generally calculated based on average numbers of syllables per word and sentence length of a 100-word extract. Using the Fry (1968) Readability technique, two CLEAR members
analyzed the text. Both calculations resulted in 11th Grade Readability in Fry analysis terms.

The reading was easy or manageable in the view of many pilot test participants. Their understanding of the task, however, may affect their judgment. For example, consider the changed perception of one examinee. Asked her opinion of the reading on the first day of a week-long CLEAR assessment, she reported with delight, “I can understand!” On the last day of the CLEAR, after taking the knowledge test and the essay assessment, this same participant said, “Oh, I wasn't reading the right way for the test,” recognizing she had not prepared well for information-based testing.

**Description of Videotaped Lecture Used in the CLEAR**

The psychology lecture on videotape related to one of the themes of the chapter, although the lecture expands on this topic in much more detail. The CLEAR Team's psychology consultant prepared the lecture following his usual procedures. The major difference for the CLEAR was in duration, for the videotaped lecture only lasted 25 minutes compared to the usual 50-minute class session. The lecture was videotaped by a professional organization in a lecture hall wired for audiovisual media creation. In attendance was a small group of the consultant's students, which improved the verisimilitude. The videotape used in the CLEAR was made under the second recording session. While the first session proceeded smoothly, the resulting lecture appeared far too polished in the perception of the CLEAR Team. The consultant was requested to re-
record his lecture, being free to speak in his natural style with all the pauses, hesitations, and back-channel cues that he would normally use. Also more naturalistic in the second session was that technical special effects were omitted, so viewers would have an experience more like that in a live lecture. The resulting lecture was considerably improved. The CLEAR team, comprised of applied linguists and ESL teachers, judged it satisfactory.

Examinee reactions to the video ranged somewhat. (Their feedback was collected on the first day and on the final day using tools described below.) One student considered the lecture to be interesting; another judged that he or she had taken good notes during the video lecture. Many of them complained about the television monitor or about the sound. While the audio-video quality was not perfect, it was acceptable in the judgment of this researcher and the research assistant, both NSEs. Most of the complaints were registered on the first day of testing; only one complaint was recorded on the final feedback session at the end of the CLEAR week. Thus, it seems possible that the complaints might be a compensatory reaction to cover a sense of frustration about their lecture listening. This might be evidenced in the data from one examinee who remarked, “The lecture focused on [i.e., addressed] native students. It's really different from EAPI classes,” proceeding to admit, “I'm not good at taking notes during the lecture.”
Research Instruments: Examinee Feedback

Examinee feedback in the latest pilot was solicited at the end of the first day and at the end of the last day of the entire CLEAR Test; the same feedback form was used at both sessions. The instrument, like all others, uses the examinee code number but not the person's name. The use of two data points was a change from earlier pilots; examinee reactions did indeed change over the test period, so the present research employs examinee feedback collection at the two time points. Examinees are asked to rate CLEAR components in difficulty, then to respond to forced-choice and open-ended items concerning preferences, verisimilitude with target situation, and awareness of own ability, among other items. A copy of the instrument is available in Appendix E.

Research Instruments: Rater Feedback

The next CLEAR instrument is the feedback form to be used with essay raters and/or administrators. (A copy of this form is located in Appendix G.) This addresses issues for raters only. (The administrator feedback section is not needed in the present study, as this researcher administered the CLEAR at the both sites.) The rater feedback form solicits data on the raters' perceptions of the CLEAR essay scoring tool in terms of usefulness, verisimilitude, and quality. Essays had previously been scored by the CLEAR Team. The present study diverged from this and consulted a subject area expert and a measurement expert. In addition to scoring duties, they were asked to reflect upon the
Essay prompt and testing conditions, as well as the scoring tool and the rater training session.

**Research Instruments: Teacher Ratings**

Teacher rankings are collected in the final instrument, the Teacher Report Form (a copy of which is attached in Appendix H). The judgment of the ESL teachers was sought to support criterion-related claims of validity concerning their students' ability; that the teachers did not participate in the CLEAR is not a concern. The participants' ESL teachers were explicitly informed to base their rankings not merely on English proficiency but on their judgment of students' overall abilities. The examinees' ESL teachers were asked first to rank order the students from strongest to weakest. They were also asked to circle one of four choices to represent their estimation of the student's likelihood of success in university work: “probably succeed,” “might succeed,” “might fail,” and “probably fail.”

**Procedures**

Before detailing the procedures for administering and analyzing the CLEAR, an overview may prove helpful. Table 4 provides the summary of test-related events.
Table 4:
CLEAR Testing Events

Day 1
1:00 p.m. - 4:00 p.m.
   Introduction to the study; informed consent (:45)
   Supervised reading of textbook chapter (1:00)
   Video lecture and notetaking (:45)
   Examinees complete feedback questionnaire (:30)
4:00 p.m. to 6:00 p.m.
   Supervised Study Hall: optional event for testees
   Distribute form collecting teacher data rating examinees

Day 2
12:00 p.m. - 1:00 p.m.; 3:00 p.m. - 5:00 p.m.
   Supervised Study Hall: optional event for examinees

Day 3
9:00 a.m. - 12:00 p.m.
   Knowledge Test, without reading folders (1:00)
   Essay Assessment, with reading folders (1:50)
   Examinees complete feedback questionnaire (:10)
   Focus group (:20)
   Collect teacher data form, after dismissing testees

Issues Concerning Test Duration and Timing

The CLEAR simulation occurs over a 48-hour period, from noon on a Wednesday through noon on a Friday. Six hours are required to participate in the CLEAR mandatory components, such as the video viewing, text reading, testing and reporting feedback. In early versions, the test battery events took place over a 15-day period, then over a week.
The test developers initially perceived the advantages of conducting the CLEAR over a long period of time, minimizing the effects of memorization, permitting deeper learning to take place better resembling real world conditions. Piloting revealed unexpected problems: fitting CLEAR events into the EAPI schedule, flagging enthusiasm as the test progressed, examinee non-compliance, difficulty in supervising test administration, and limited examinee motivation to study. The test developers agreed to sacrifice a degree of verisimilitude for better efficiency and restrict the CLEAR to a 48-hour period. The testing in the current study started the CLEAR one afternoon and completed it by noon on the third day. This schedule permitted the CLEAR developers to control test administration at remote sites: travel and transportation are more feasible with this schedule than over a longer duration. Controlling the test administration is also good measurement, since it minimizes a source or measurement error. Finally, the 48-hour period would seem to facilitate the logistics of the CLEAR: finding rooms and equipment over a few days is generally more feasible than over a longer period.

The CLEAR study was administered in late June, 2004 at Site #1 and at Site #2 a few weeks later. An overview of the administration schedules is as follows. Session 1 was held from noon to mid-afternoon on Day 1, a Wednesday. In this session, examinees underwent the informed consent process, received and studied the reading materials, next viewed the video lecture while taking notes, then completed the feedback questionnaire. Then examinees were allowed to study the reading material silently in a proctored
environment. Teacher data forms were be administered this first day, for collection at the end of the CLEAR. On Day 2, supervised study was allowed. The morning of Day 3 was the final session of the CLEAR assessment. The session, which lasted from 9 a.m. to 12 noon, comprised the knowledge subtest and the essay subtest. The final component with the test-takers was a second administration of the examinees' feedback data collection. The experimenter lastly collected teachers' data forms ranking the examinees.

**Role of Specialists in Validating the CLEAR**

Two specialists were employed to advise the researcher in this initial test validation process, an expert in the content area and an expert in testing. A content-area expert was selected on the basis of psychology faculty recommendations. This expert is a professor of psychology and co-director of a research center. In terms of instructional experience, she has taught graduate and undergraduate courses, and has won awards for teaching excellence. This psychologist focuses on cognitive and neural sciences and clinical psychology. She is known nationally for her research in child psychology and child second language development. The second expert is a psychometrician: her specialization is in testing and measurement. Following extensive work with major testing companies and a period in academia, she now is an independent consultant. She enjoys a national reputation as an expert in test development, innovative item types, and computer-based testing. The specialists were offered honoraria, paid by the researcher, for their valuable service.
The expert session occurred over the course of one day during July, 2004. The content area expert had reviewed the stimuli and knowledge test prior to the meeting, as a gracious consideration to the productivity of the session. The meeting commenced, after signing informed consent releases, with an overview of the CLEAR, which served as a review for the content judge and an orientation for the psychometrician. The experts were invited to evaluate and comment upon the stimulus materials. Attention then turned to the objective subtest, and the specialists provided their expert judgment of the knowledge test, although scoring that multiple-choice test was performed by the researcher.

Proceeding on to work with the constructed-response subtest, the experts were consulted on the essay prompt and on the essay scoring tool which uses the criteria developed by the CLEAR team. Discipline-based faculty, such as these psychology teachers, typically do not perceive themselves as writing instructors: they focus upon the academic content of student writing (see, for example, Schleppegrell, 2002 and Waters, 1996). Indeed, the psychology consultant who advised the CLEAR team at the early stages of test development said that he graded essays strictly on the presence or absence of particular factual criteria established for each essay. Thus, a discussion seems warranted concerning the relationship between the criteria in the CLEAR essay scoring tool and those criteria employed in the subject disciplines for scoring essays. Such will help establish the merit of the CLEAR scoring tool for its intended purpose. The psychology expert consulted during this validation study fully endorsed using the essay scoring tool developed by the
CLEAR team. She considered that the scoring criteria were well selected, with appropriate emphasis of primary and secondary factors. She also commented that the scoring tool appeared easy to use. She judged that the CLEAR scoring tool, though yielding insufficient detail for diagnostic purposes, was quite appropriate for making the holistic decisions for which it was developed. The measurement expert suggested revising the language of the prompt: constraining the examinee to focus on content aspects would likely help testees demonstrate their knowledge and also improve the scoring consistency. Implementing her recommendation will be explored in the final chapter of the present work.

The judges next prepared to rate compositions by "range-finding practice" that employed essays from earlier trials of the CLEAR. These seven essays had been scored previously by the CLEAR test builders with the present holistic scoring tool. Using the CLEAR team's mean rating score from the past (“Team”), essays were selected to represent the different score levels. Scoring these essays anew helped to illuminate how well the CLEAR team and the researcher ("RES") are in agreement with the content expert ("SME"). The quality range helped prompt discussion of the different characteristics important to successful university writing, as became evident in the discussions following the scoring of each practice essay.
The content expert and the researcher then individually scored 21 of the CLEAR essays, with discussion on reasoning for rating assignment. Because time pressure limited the expert's further participation, the researcher completed the remaining scoring alone. (As the expert's ratings and those of the researcher correlated rather well (r = .83), the one-rater scoring was not as great a concern as it might have been.) Essays were awarded the mean of the two rating scores. The raters never diverged by more than one point on the 0-4 point scale. Had that case arisen, a third judge would have been invited to rate the problem essay, with the disputed essay receiving the mean of all three raters' scores.

As well, the SMEs were invited to discuss whether examinees seem academically ready and should earn an overall “pass” on the CLEAR dynamic simulation. Obtaining experts' recommendations is valuable, particularly since discipline-based instructors may have a perspective different from that of ESL faculty (cf Horowitz, 1986a, 1986b; Johns, 1991; Leki, 1995; Leki & Carson, 1997). By soliciting the judgment of some psychology

<table>
<thead>
<tr>
<th>Testee/Rater</th>
<th>#204</th>
<th>#205</th>
<th>#206</th>
<th>#210</th>
<th>#312</th>
<th>#314</th>
<th>#315</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SME1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>RES</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5
*Ratings Assigned During Initial Practice Session*
instructors in the current study, the present study can contribute to future work on standard-setting.

All of these qualitative insights contributed by expert judges are scrutinized and have contributed to the findings of this study. The expert judgments also contribute in the recommendations for improving the CLEAR.

*Examinee Motivation in Taking the CLEAR.*

CLEAR examinees, as noted elsewhere, are presented incentives to perform well. Under pilot conditions, the volunteers, all students from one IEP, were awarded extra credit in their ESL classes regardless of the CLEAR scores they earned. The researcher did not seek to have host sites change their grading policies. Wishing to have both sites' participants similarly motivated, the researcher offered different incentives in the present validation study. The first incentive was the possibility of recommendation for admission to LSRU, exempt from taking the TOEFL. Second, prizes were awarded by lottery to three examinees among the 20 highest-scoring examinees of the 36 participants. These measures were intended to help promote extrinsic motivation to CLEAR volunteers. Additionally, pilot participants had reported motivation by gaining a sense of the university experience through the CLEAR Test.

**Analyses**

Test validation is principally concerned with assessing the qualities of reliability and validity, discussed in more detail in the following sections. Thereafter, the discussion
will consider the guiding research questions for this validation study and the analyses conducted to help answer these questions.

**Reliability**

Reliability indicates “the degree to which test scores are free from errors of measurement” (AERA *et al.*, 1985, p. 19). Variance not related to the construct under scrutiny is measurement error. Error can be traced to four sources: the test itself, the administration of the test, the scoring of the test, and the examinee. The test itself can contribute to error if it is too short, if the items are badly written, or if the domains are poorly sampled. The test administration can skew results, as when a teacher coaches examinees during the test or when testing conditions vary in comfort or distraction. Scoring contributes to measurement error if an item is not keyed correctly or if judges are inadequately trained. Finally, the test-taker is a source of variance: examinees are humans and come to the test with such causes of measurement error as lack of sleep, illnesses, and personal problems.

Reliability is quantified by correlating two scores. For example, "test-retest reliability" is a correlation of scores earned on the first test administration with that of the second. Two forms of a test can be correlated: such is termed "alternate forms reliability." "Split-half reliability" is the result of correlating scores earned on the first part of a test with the remaining component. Comparing performance on two different parts of a test is another form of reliability, called "internal consistency." This measure estimates the
relationship of response patterns among test items, for example, the association between response patterns to one test item and response to the remainder of the test. Internal consistency is appropriate for a test with dichotomously scored items, particularly useful for tests measuring a unidimensional trait.

Among these different possibilities, the internal consistency technique was selected, appropriate for a rather brief test administered only one time, not compared to another instrument. Several techniques exist for calculating internal consistency. The Kuder-Richardson 20 (KR20) formula is often recommended for estimating the consistency of a measure with dichotomously scored items. Statistically equivalent is Cronbach’s coefficient alpha. The latter technique is easily generated in SAS statistical software, and was the statistic herein employed. In these and all reliability correlations, the result is a coefficient, \( r \), that ranges from as low as 0, for perfectly unreliable test scores, to 1.0 for perfectly reliable test scores.

**Validity**

Validity, the second quality under examination, is “the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences” (AERA et al., 1985, p. 9). Validity is unitary, but three general categories of evidence may be distinguished. First is content-related evidence, wherein the actual content of a test is
scrutinized to determine its representativeness with the universe that the content should sample. This examination was conducted by a subject area expert hired to advise the researcher. Second is criterion-related evidence, a systematic connection of test scores with other outcome criteria. The concurrent measures here employed were scores earned on static tests of language proficiency and teachers' ratings of examinee likely success in university work. The third aspect of validity study was examined through construct-related evidence, previously raised through theoretical framework and argumentation, and further explored through inter-correlation of the two CLEAR dynamic simulation subtests.

**Research Questions and Analyses**

Recall that three research questions guide the present study. The questions are reprinted below for reader convenience, followed by the analyses selected to help answer the questions.

Prior to answering any specific research question, it is critical to establish whether the scores from the two testing sites can be appropriately pooled. In order to evaluate sample equivalency across testing sites, $t$-tests were conducted; analyses were also calculated to determine equivalent test functioning across subgroups (language, gender, proficiency level). Additionally, item analyses were conducted. Items on the knowledge test were described in terms of facility value and discrimination index, and distractor
analysis helped complete the statistical description of the knowledge subtest.

**Research Question One:** What is the consistency (i.e., reliability) of the CLEAR?

Consistency was examined in both subtests. The specific analyses to answer Question One were as follows. Cronbach's alpha was calculated to estimate the consistency of the knowledge subtest items. For the essay subtest, essay scorers' inter-rater consistency were analyzed by a $t$-test of mean rating score and by correlation. Raters' scores were compared with each other and, in a limited subset, across time with CLEAR team scoring. Inter-tester dependability was not necessary, as the test administrator delivered the test at both sites. The CLEAR consistency was also examined by calculating descriptive statistics and the standard error of measurement (SEM). This statistic is in inverse relationship with standard deviation (SD); the SEM is useful in estimate how far the test scores vary from examinees' true ability.

**Research Question Two:** What evidence exists for claims for validity of inferences based on CLEAR scores? To answer this, validity-related evidence was marshaled for the content, concurrent, and construct aspects of validity. Evidence for content validity was solicited from the content expert. Note that certification tests, such as the CLEAR dynamic simulation, are fundamentally grounded upon content validity evidence. The consultant was invited to judge test content (stimuli, knowledge test items, and essay prompt) for representativeness and for domain sampling. The expert also judged the essay prompt and rated examinee essay performance (a sample of which was
correlated with CLEAR team scores). These formed the content-related evidence sources collected. The next facet of validity is concurrent validity, wherein support for validity claims is found through scores on other measures collected at approximately the same time. Two broad categories of evidence were collected. The first source was in the form of teachers' rankings of examinees' university readiness; this was expected to provide convergent support for CLEAR validity claims. The second type of evidence came from placement and proficiency test scores; these were anticipated to provide discriminant evidence since these measures are so distinctly different from the CLEAR dynamic simulation. These scores were correlated with CLEAR dynamic simulation subtest scores. The third facet concerns evidence for validity claims in terms of the construct, AR. Certification tests are more centrally concerned with close connection to the target domain: underlying constructs are not of focal interest with certification tests as they are in tests of latent psychological traits. Nevertheless, this aspect of validity is not without interest. Construct-related evidence of validity claims was sought, qualitatively and quantitatively. Qualitative evidence for construct-related validity claims might be partly supported by this present study in terms of the literature review, the quality of argumentation, and the overall quality of the research. Quantitative construct-relevant evidence for validity were found in terms of inter-subtest correlations (Alderson, Clapham, & Wall, 1995, p. 184). The predictive aspect of validity-related evidence would help establish the utility of the CLEAR dynamic simulation assessment. Limitations of
time and funding, however, did not permit including this predictive aspect of validity claims into the the scope of the present work.

**Research Question Three: What is the examinee perception of the CLEAR?** This question of how well the CLEAR strikes a balance between ideals and efficiency was explored through feedback data collected from participants in the CLEAR dynamic simulation. These data were analyzed by qualitative analysis, following techniques recommended by Lincoln and Guba (1985, pp. 344-350). They detail a procedure of content analysis that consists of two major stages. The first is an analysis stage, finding basic data units that are heuristic or informational and that cannot be broken down into smaller units. The basic information unit should then be recorded on an index card and coded. The second stage is a synthesizing stage. Here, the investigator sorts and organizes data into different categories.

The essential tasks of categorizing are to bring together into provisional categories those cards that apparently relate to the same content; to devise rules that describe category properties and that can, ultimately, be used to justify the inclusion of each card that remains assigned to the category as well as to provide a basis for later tests of replicability; and to render the category set internally consistent. Note that the category set that emerges cannot be described as *the* set; all that can reasonably be required of the analyst is that he or she produce *a* set that provides a "reasonable" construction of the data. "Reasonable" is most easily defined as a judgment that might be made subsequently by an auditor reviewing the process. (p. 347) [emphases in the original]
Lincoln and Guba recommend a "constant comparison" method. In this technique, each unit card is scrutinized before considering where it best fits. The analyst will first consider whether the new unit card fits with an existing category. If not, a new category is created. This procedure is followed until all the data units have been studied and categorized. The categories are next scrutinized for consistency, by pondering what theme or propositions can account for all the data in this group. The theme is tested on data units within and outside the category. Upon completing this process with a category, the process is replicated with all other categories. The knowledge base as a whole is later examined, particularly to account for miscellaneous data units, to seek to clarify overlap between categories, and to investigate any relationship among categories. The last point is important, for such might clarify understanding to subsume small categories together under a general cover.

Having considered the methodology of the present research, recall that the CLEAR is intended to give ESL students an opportunity to demonstrate their abilities under conditions that resemble the real world but with attention to rigorous measurement quality. Evaluating the quality of the test in its present state, firstly, will permit the test developers to use the CLEAR for its intended purpose, with a sense of what the test scores mean. Moreover, the evaluation might contribute to the research and development agenda for ongoing improvement of the CLEAR.
CHAPTER 4

RESULTS

Overview of Research Questions

The present research is conducted in order to investigate the measurement quality of the CLEAR. As an initial validation study, this investigation must examine various fundamental issues prior to considering the three research questions that form the heart of the research. Numerous analyses have been conducted, for example, on the stimulus materials, the knowledge subtest, the essay subtest, the essay scoring tool, and the raters' scoring agreement. As an aid for the reader, the list below recapitulates the analyses comprised in this research. Thereafter, the data analyses will be examined in detail and results reported.
### Preliminary Analyses

Factors analyzed first are those matters important in establishing a foundation prior to examining the actual research questions. First of these fundamental factors is:

#### Table 6

**Recapitulation of CLEAR Analyses**

<table>
<thead>
<tr>
<th>Preliminary analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine sample equivalence across testing sites</td>
</tr>
<tr>
<td>Determine equivalent test functioning across subgroups (i.e., proficiency level, language, gender)</td>
</tr>
<tr>
<td>Conduct item analyses (facility value, discrimination index, and distractor analyses)</td>
</tr>
</tbody>
</table>

**Research Question 1: What is the consistency of the CLEAR Test?**
- Knowledge subtest: consistency of items per Cronbach’s alpha test of internal consistency
- Essay subtest: essay scorers’ inter-rater consistency
- CLEAR Test overall: descriptive statistics (mean, mode, median, range, and standard deviation), and the standard error of measurement (SEM)

**Research Question 2: What is the evidence for the validity of inferences based on CLEAR Test scores?**
- Content-related evidence, evaluated by a content expert
- Criterion-related evidence, by concurrent measures of ability
  - Correlations of CLEAR knowledge subtest scores with teacher ratings
  - Correlations of CLEAR essay subtest scores with teacher ratings
- Construct-related evidence
  - Correlations of CLEAR knowledge subtest and essay subtest scores with placement test scores, proficiency test scores for discriminant evidence
  - Intercorrelation of knowledge subtest and essay subtest scores

**Research Question 3: What is the examinee perception of the CLEAR?**
- Test quality according to examinee feedback collected at the end of the first day and at the conclusion of the simulation; examined by qualitative analysis.

---

---

---

---
determining sample equivalence across testing sites, to establish whether it is appropriate to collapse together those participants into a single group.

Table 7
Mean Student CLEAR Scores by Test Site

<table>
<thead>
<tr>
<th>Subtest Scores</th>
<th>M</th>
<th>SD</th>
<th>n</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>knowledge</td>
<td>14.80</td>
<td>12.63</td>
<td>3.90</td>
<td>4.06</td>
</tr>
<tr>
<td>essay</td>
<td>2.05</td>
<td>1.75</td>
<td>0.94</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Note: Maximum knowledge score = 32, maximum essay score = 4. Essays were scored only if they followed the test guidelines.

Test score means, for both subtests, are not statistically significantly different between the two testing sites. The mean score comparison indicates it is reasonable to collapse the data from the two locations, and consider such as one sample hereafter.

Regarding the score range, the examinee performance was poor in objective terms. Important points in interpreting the scores are that the examinees had very limited time to prepare for the test. Also, many examinees reported that they were not reading effectively. Perhaps they were preparing for an ESL test by learning vocabulary...
definitions and identifying the main ideas, rather than learning factual information reported in the text. ESL students in the real world will have more time to study before an examination than they did in the simulation. The students can also learn from failure and improve subsequent performance, although the CLEAR parameters did not permit this.

The next preliminary matter of investigation concerns equivalent test functioning across subgroups. Such analyses consider whether test performance differences exist at a group level, analyses important in establishing test fairness. An analysis was conducted to illuminate the appropriacy of opening the CLEAR to examinees of different ESL proficiency levels.

<table>
<thead>
<tr>
<th>CLEAR Subtest Scores</th>
<th>M</th>
<th>SD</th>
<th>n</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>knowledge</td>
<td>Level Y</td>
<td>13.20</td>
<td>14.20</td>
<td>3.42</td>
</tr>
<tr>
<td>essay</td>
<td>Level Y</td>
<td>1.71</td>
<td>2.21</td>
<td>0.82</td>
</tr>
</tbody>
</table>

*Note: Maximum knowledge score = 32, maximum essay score = 4
*Essays were scored only if they follow the test guidelines.

Scores were analyzed across proficiency groups, comprising the highest level (“Level Z”) and the penultimate level (“Level Y”) of English language mastery at the different language institutes. (Recall that the two testing sites categorize students into...
four or five proficiency levels, so "Level IV" students at one site are in the highest level but in the second-highest level at the other. To avoid confusion, the highest level is herein called "Level Z" and "Level Y" is the proficiency just below that.) For both the knowledge test and the essay assessment, score differences were not statistically significantly different between Level Z and Level Y at these different English proficiency levels. This suggests that it is reasonable to administer the CLEAR at these two levels.

Another preliminary issue involved equity across language groups. Analyzing test results among the different L1 groups was planned in order to evaluate the CLEAR for linguistic bias. The 36 study volunteers represented nine different languages (Spanish, Korean, Japanese, Thai, Turkish, Arabic, Chinese, Portuguese, and Yoruba). Groups were not sufficiently large enough to subject to statistical analysis. Nevertheless, frequency counts indicated that the sample was linguistically heterogeneous. This variability is a positive feature for an initial validation study. An examination of linguistic bias would, however, remain for future research.
The next fundamental test examined the performance of CLEAR examinees across gender groups. This test was to see evidence, whether of fairness or bias, in CLEAR scores by gender. On the essay, $t$-test results show that mean scores (females=1.71 SD=0.72, males=2.21 SD=1.05) were not statistically significantly different, $t(26) = -1.46, p = .15$. That is, score differences by gender could have occurred by chance.

<table>
<thead>
<tr>
<th>First Language</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>Korean</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>Japanese</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Thai</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Turkish</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Arabic</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Chinese</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Yoruba</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: Percent might exceed 100 because of rounding.
The knowledge subtest mean score (females=12.37 SD=4.64, males=15.00 SD=3.19) t-test results also found no statistically significant differences. Male examinees' essays scores were apparently higher compared to the essay scores of the female examinees, but the statistical test revealed no significant difference at the .05 level, t(34) =-2.01, p = .0477. Despite the small sample size, (14 females and 14 males), the results nevertheless attest that the score difference might well be chance, rather than a real difference in the multiple-choice test performance. These findings indicate that the visually better scores by male students could be anomalous. No relationship should therefore be assumed.

A quantitative item analysis was conducted to assess the quality and functioning of the knowledge test items and distractors; the results are presented in Table 11. The

<table>
<thead>
<tr>
<th>CLEAR Subtest Scores</th>
<th>M</th>
<th>SD</th>
<th>n</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>knowledge</td>
<td>Female</td>
<td>12.37</td>
<td>4.64</td>
<td>16</td>
<td>-2.05</td>
<td>34</td>
</tr>
<tr>
<td>essay</td>
<td>Male</td>
<td>15.00</td>
<td>3.19</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1.71</td>
<td>0.72</td>
<td>14*</td>
<td>-1.46</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>2.21</td>
<td>1.05</td>
<td>14*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Maximum knowledge score = 32, maximum essay score = 4
*Essays were scored only if they follow the test guidelines.
research environment employed SAS statistical software version 8.2 for Windows operating system using Cody's (1997, pp. 273-276) programming code. Note that, in item analysis, the unit under scrutiny is not the examinee, but the test item as evidenced by responses to the item. Basic statistical calculations yield a mean score for each item. This mean score is an important statistic: it is the \( p \)-value, or the probability that an examinee will answer the item correctly. This value is also variously called the item facility value, or item difficulty. This statistic is reported in the second column of Table 11, labeled "diff." How this statistic relates to the future development of the CLEAR will be explored in the final chapter of this study, but a few comments may be of interest here.

Following the consideration of item difficulty, the question presents itself: who are the examinees answering correctly? Are they the examinees who perform well on the test or those who perform poorly? The statistic expressing such value is item discrimination (the "discrim" column in Table 11), the point-biserial correlation \( r_{pbis} \) between response to an item and responses to the whole test. An item discriminates positively, for example, when high-scoring examinees answer correctly and low-scoring examinees do not; negative item discrimination occurs when strong test-takers miss an item but weak students answer correctly. Items that discriminate negatively might have been badly written or even scored incorrectly. The item discrimination analysis reveals that five items in the CLEAR (numbers 2, 6, 21, 23, and 28 in the published numbering system) correlate negatively with the overall test, although items 2 and 28 are only
slightly negatively correlated (-0.06 and -0.08, respectively). The items had all been scored correctly. The negative discrimination might be due to confusing wording. All have been removed from the live version of the test pending revision. In this test version, many items discriminated negatively. This underscores the need for field-testing items to assess their functionality, and the need to build a larger bank of quality items. The item quality will be further discussed in the final chapter of this study.

Item discrimination may furthermore be scrutinized through viewing the four quartile columns in Table 11. These columns report response behavior by group performance. One may thus compare how the poorest scoring test-takers of Quartile 1 answered an item versus the top examinees in Quartile 4. Consider, for example, item 11, which was answered correctly by 49% of all test-takers. Only 22% of the lowest quartile examinees answered this item correctly, while 67% of the top students answered it correctly. This item appears to function well. Other items draw correct responses more from middling examinees or from poor examinees. Such results seem to indicate the need for item inspection and revision, a matter discussed further in the final chapter of this report.

One more result is included in Table 11, the distractor analysis, displayed in the five columns at the far right side of the table. Here is reported how many examinees selected each of the multiple-choice options. The correct response is found beside the item number in the first column; below each column of the distractor analysis is
displayed the percentage of examinees selecting each option. This analysis indicates how attractive distractors are: in item 1, for example, very few testees chose option A while the large majority was attracted to option D. Several other test items have distractors that attract few examinees. Distractor analysis results can help shed light on why an item functions poorly, and can indicate item options in need of improvement.

Table 11

<table>
<thead>
<tr>
<th>#</th>
<th>Key</th>
<th>diff</th>
<th>p-value</th>
<th>discrim</th>
<th>Qrtl 1 % correct</th>
<th>Qrtl 2 % correct</th>
<th>Qrtl 3 % correct</th>
<th>Qrtl 4 % correct</th>
<th>a %</th>
<th>b %</th>
<th>c %</th>
<th>d %</th>
<th>f %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 d</td>
<td>.78</td>
<td>.31</td>
<td>55.6%</td>
<td>72.7%</td>
<td>88.9%</td>
<td>100%</td>
<td>3</td>
<td>3</td>
<td>17</td>
<td>*78</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 d</td>
<td>.47</td>
<td>-.06</td>
<td>44.4%</td>
<td>54.5%</td>
<td>55.6%</td>
<td>28.6%</td>
<td>19</td>
<td>6</td>
<td>28</td>
<td>*47</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 d</td>
<td>.69</td>
<td>.15</td>
<td>77.8%</td>
<td>36.4%</td>
<td>77.8%</td>
<td>100%</td>
<td>14</td>
<td>14</td>
<td>3</td>
<td>*69</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 a</td>
<td>.44</td>
<td>.25</td>
<td>44.4%</td>
<td>27.3%</td>
<td>44.4%</td>
<td>71.4%</td>
<td>*44</td>
<td>14</td>
<td>36</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 a</td>
<td>.61</td>
<td>.42</td>
<td>22.2%</td>
<td>72.7%</td>
<td>77.8%</td>
<td>71.4%</td>
<td>*61</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 b</td>
<td>.33</td>
<td>-.42</td>
<td>66.7%</td>
<td>36.4%</td>
<td>11.1%</td>
<td>14.3%</td>
<td>28</td>
<td>*33</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 b</td>
<td>.53</td>
<td>.34</td>
<td>33.3%</td>
<td>45.5%</td>
<td>66.70%</td>
<td>71.4%</td>
<td>17</td>
<td>*53</td>
<td>19</td>
<td>11</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 b</td>
<td>.83</td>
<td>.30</td>
<td>66.7%</td>
<td>81.8%</td>
<td>88.9%</td>
<td>100%</td>
<td>6</td>
<td>*83</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 a</td>
<td>.06</td>
<td>.26</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>11.1%</td>
<td>14.3%</td>
<td>*6</td>
<td>25</td>
<td>44</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10 c</td>
<td>.58</td>
<td>.28</td>
<td>22.2%</td>
<td>63.6%</td>
<td>100%</td>
<td>42.9%</td>
<td>14</td>
<td>19</td>
<td>*58</td>
<td>8</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 d</td>
<td>.49</td>
<td>.25</td>
<td>22.2%</td>
<td>54.5%</td>
<td>55.6%</td>
<td>66.7%</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>*49</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 b</td>
<td>.19</td>
<td>.32</td>
<td>0.00%</td>
<td>18.2%</td>
<td>33.3%</td>
<td>28.6%</td>
<td>19</td>
<td>*19</td>
<td>56</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 a</td>
<td>.81</td>
<td>.62</td>
<td>33.3%</td>
<td>90.9%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>*81</td>
<td>3</td>
<td>11</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>14 b</td>
<td>.22</td>
<td>.26</td>
<td>22.2%</td>
<td>0.00%</td>
<td>33.3%</td>
<td>42.9%</td>
<td>25</td>
<td>*22</td>
<td>36</td>
<td>17</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 b</td>
<td>.75</td>
<td>.54</td>
<td>44.4%</td>
<td>72.7%</td>
<td>88.9%</td>
<td>100%</td>
<td>3</td>
<td>*75</td>
<td>3</td>
<td>19</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Above item order does not represent actual item order, so that security of live test form may be protected. Quartile 1 consists of the lowest-scoring examinees, while the highest-scoring examinees are in Quartile 4. Distractor analysis cells filled with a "0" indicate no examinee selected that option. An asterisk (*) marks the correct option.

continued on the next page
The findings of the item analysis findings can be briefly considered here. The first major factor to examine is item difficulty. A generalization may be made that the $p$-values represent a range of difficult to considerably difficult items: 11 of relatively less difficulty ($p$-values below .34), 13 of middling difficulty ($p$-values between .35 and .66),
and 8 of considerable difficulty, \((p\)-values above \(.67\)). This type of distribution of item difficulty optimizes the spread of test scores and permits rank-ordering examinees. This spread is generally associated more with norm-referenced tests (NRTs) than criterion-referenced tests (CRTs).

The second major dimension of the item analysis is the discrimination index. Characterizing item discrimination may be simply deeming the value as acceptable or not. Kubiszyn and Borich (2000) note that experts disagree on answers to the question, "How high is a 'good' discrimination index?" (p. 139). Some judge that a positive value is sufficient; others prefer a discrimination index of \(.30\) at minimum. Adding to the complexity is that discrimination is a function of difficulty: discrimination is lower with very easy or very difficult items, higher with moderately difficult items. For example, when a test item is extremely easy and all the examinees answer it correctly, such a \(p\)-value reduces the possibility of discriminating between strong and weak test-takers. The same result occurs with extremely difficult items. These factors help shed light on the CLEAR item analysis. Considering such, therefore, the discrimination indices found in CLEAR items can be judged as quite good. A summary table has been prepared to display the discrimination values.
An example of scrutinizing an item through item and distractor analysis may contextualize the data analysis. Consider item 19 for instance. This item appears strikingly easy, for the $p$-value is .94. The item discrimination functions acceptably, at .28 falling between the generally accepted boundaries of .20 and .30. The performance by quartile illustrates the positive discrimination, for the response pattern increases reading rightward from Quartile 1, then higher in Quartile 2, still higher in Quartile 3, and highest in Quartile 4. This illustrates that fewer of the low-achieving examinees and more of the high-achieving test-takers answered this item correctly.

The response pattern shows that option A attracted 6% of the examinees, option B drew 94% of the responses, whereas no test-taker chose option C nor option D. These

<table>
<thead>
<tr>
<th>Range</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -.4</td>
<td>-0.42</td>
</tr>
<tr>
<td>&lt; -.1</td>
<td>-0.12 -0.14</td>
</tr>
<tr>
<td>&lt; 0</td>
<td>-0.06 -0.08</td>
</tr>
<tr>
<td>0 - .19</td>
<td>0.12 0.15 0.16</td>
</tr>
<tr>
<td>.20 - .29</td>
<td>0.22 0.25 0.25 0.25 0.26 0.26 0.28 0.28</td>
</tr>
<tr>
<td>.31 - .39</td>
<td>0.30 0.31 0.32 0.33 0.34 0.36</td>
</tr>
<tr>
<td>.41 - .49</td>
<td>0.42 0.42 0.43 0.44 0.45 0.47</td>
</tr>
<tr>
<td>.51 - .59</td>
<td>0.54</td>
</tr>
<tr>
<td>.61 - .69</td>
<td>0.62</td>
</tr>
</tbody>
</table>
results raise questions. Possibly this item is a good one, but one that most students knew, not an unusual situation in a CRT. Perhaps options C and D are artificially easy, answers that are quite improbable, thus causing the item to function as a de facto true-false item. Because of the response pattern, the item distractors invite inspection and consideration. Minor revisions might strengthen options C and D into appealing distractors.

Lest this sound deceptive or unfair, recall that attractive options help a selected-response test function effectively. Appealing distractors help ensure that an item is answered correctly by examinees who correctly understand the material, but the same item is missed by examinees who are guessing. The quality of response items is, therefore, crucial in a multiple-choice test to ensure that the test measures the construct under consideration but minimizes construct-relevant variance. A spread of varied responses across the different options is generally desirable. Note, however, a hypothetical pattern wherein each of the four options drew 25% of the examinees: this pattern might indicate that everyone is guessing. Contrast such with the responses found in the present distractor analysis. In general, the distribution of response proportions for most CLEAR knowledge subtest items would suggest that the response options function well. The response patterns generally show good spread across the choices, thus examinees are choosing various options. Such results, coupled with the good discrimination values, are positive indicators of quality in CLEAR knowledge subtest items.
Research Question 1: What is the consistency
of the scores derived from the CLEAR?

Proceeding with the first research question is appropriate having concluded the
fundamental analyses. Examining score consistency (i.e., reliability) is achieved through
basic statistical analysis of the shape and dispersion of the score distribution. Following
such, estimating the CLEAR internal consistency is appropriate. Cronbach's coefficient
alpha was employed to estimate the knowledge test internal consistency. The rationale for
this choice is as follows. Firstly, a statistical test was chosen that did not rely on
comparisons across time (as in a test-retest situation). Secondly, the knowledge subtest is
comprised of dichotomously scored items. Consequently, therefore, recommended tests
of internal consistency are the Kuder-Richardson 20 (KR20) formula or Cronbach's
alpha, its statistical equivalent. Cronbach's alpha was employed to analyze the
consistency of the items comprising the CLEAR knowledge subtest.

<p>| Table 13  |
| CLEAR Knowledge Subtest Scores Descriptive Statistics |</p>
<table>
<thead>
<tr>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
<th>α</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>15.11</td>
<td>3.64</td>
<td>15.00</td>
<td>14.00</td>
<td>7-21</td>
<td>-0.31</td>
<td>-0.40</td>
<td>.528</td>
<td>2.50</td>
</tr>
</tbody>
</table>

*Note: Maximum knowledge score = 32*
Descriptive statistics indicate that the mean knowledge test score for the 36 examinees is 15.11 and the standard deviation is 3.64. The score distribution forms a rather flat curve (-0.31 kurtosis) with a moderately negative skew (-0.40). Internal consistency is indicated by the alpha of .528. The standard error of measurement, a function of the standard deviation and the alpha, is 2.50. This statistic estimates the difference between an examinee's observed and true score.

Internal consistency, a concern raised by these findings, could be improved by removing problem items. Table 14 shows the effect of removing poor items one at a time. This table provides information useful in test revision. Item 6, for example, exhibits negative correlation with the rest of the test. Along with all the other negatively discriminating items, this item was scrutinized for causes of problems. One possible problem source considered was whether the item was derived from the text or lecture, and whether the source seemed a problem. Inspection revealed that neither item source was remarkable for problem items. The item wording was considered, as was the distractor quality. Another possibility was the item placement: an item with only three distractors might have confused examinees because it was placed at the end of a page, causing them to flip between pages searching for a fourth distractor. These possibilities were pursued with all of the negatively correlating items, and the test owners, the EAPI, were advised to remove the items pending revision. The final chapter will address this
quality of the knowledge subtest, an important constituent of the CLEAR simulation of the university experience.

Table 14
Knowledge Subtest Reliability Improvement by Removing Poor Items

<table>
<thead>
<tr>
<th>Deleted Test Item</th>
<th>Raw Variables</th>
<th>Correlation with Total</th>
<th>Alpha if Item is Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.00</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.20</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.12</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.51</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.21</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.21</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.20</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.14</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.11</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.21</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.55</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.15</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.44</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.11</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.02</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.30</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.22</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

Note: Above item order does not represent actual item order, as a means to protect the security of a live test.

continued on the next page
Table 14, *continued*

<table>
<thead>
<tr>
<th>Deleted Test Item</th>
<th>Raw Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation with Total</td>
<td>Alpha if Item is Deleted</td>
</tr>
<tr>
<td>1</td>
<td>.20</td>
<td>.51</td>
</tr>
<tr>
<td>2</td>
<td>-.20</td>
<td>.57</td>
</tr>
<tr>
<td>3</td>
<td>.03</td>
<td>.53</td>
</tr>
<tr>
<td>4</td>
<td>.12</td>
<td>.52</td>
</tr>
<tr>
<td>5</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>6</td>
<td>-.51</td>
<td>.60</td>
</tr>
<tr>
<td>7</td>
<td>.21</td>
<td>.51</td>
</tr>
<tr>
<td>8</td>
<td>.21</td>
<td>.51</td>
</tr>
<tr>
<td>9</td>
<td>.20</td>
<td>.52</td>
</tr>
<tr>
<td>10</td>
<td>.14</td>
<td>.52</td>
</tr>
<tr>
<td>11</td>
<td>.11</td>
<td>.52</td>
</tr>
<tr>
<td>12</td>
<td>.21</td>
<td>.51</td>
</tr>
<tr>
<td>13</td>
<td>.55</td>
<td>.47</td>
</tr>
<tr>
<td>14</td>
<td>.15</td>
<td>.52</td>
</tr>
<tr>
<td>15</td>
<td>.44</td>
<td>.48</td>
</tr>
<tr>
<td>16</td>
<td>.11</td>
<td>.52</td>
</tr>
<tr>
<td>17</td>
<td>.02</td>
<td>.54</td>
</tr>
<tr>
<td>18</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>19</td>
<td>.22</td>
<td>.52</td>
</tr>
</tbody>
</table>

*Note:* Above item order does not represent actual item order, as a means to protect the security of a live test.
Having examined the knowledge subtest, the question of consistency must be considered with the essay subtest. The content expert and the researcher began the scoring session with practice on essays from pilot versions of the CLEAR. Beginning with these exemplars allowed the two raters to discuss how to apply scoring criteria to seven real essays. Table 15 shows the Pearson product moment correlation for the subject area expert (SME), for the researcher (RES), and the mean ratings from the CLEAR team's earlier scoring (TEAM). The researcher scored only the last four of the seven practice essays because of the demands of leading the essay session. Increased harmony of perspective developed as the practice session progressed, as manifested in scoring consistency. After scoring each essay, the raters engaged in a thorough discussion of criterion values and their realization. Likely this extensive discussion contributed to shared perspective.

<table>
<thead>
<tr>
<th></th>
<th>SME</th>
<th>RES</th>
<th>TEAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SME</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RES</td>
<td>.59</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>TEAM</td>
<td>.50</td>
<td>.91</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Seven essays were employed in the range-finding exercise.
Having completed the initial practice, the scorers set to work on rating actual test products. Mean score comparison was the statistical test recommended for this analysis. In the actual CLEAR essays, mean scores were quite similar across the two raters, as shown in Table 16. The content expert's mean rating was 1.43 and the researcher's was 1.42.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
</tr>
</thead>
<tbody>
<tr>
<td>SME</td>
<td>21</td>
<td>1.43</td>
<td>1.32</td>
<td>1.00</td>
<td>1.00</td>
<td>0 - 4</td>
<td>-0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>RES</td>
<td>36</td>
<td>1.42</td>
<td>1.13</td>
<td>1.00</td>
<td>2.00</td>
<td>0 - 4</td>
<td>-0.24</td>
<td>0.46</td>
</tr>
</tbody>
</table>

More will be discussed concerning the mean scores, but first a description of the score distribution is in order. Considering other aspects of the ratings, note the similarity in SD and median scores for the two raters. Differences do exist: consider the mode score of 1.0 for the subject-area expert and 2.0 for the researcher. More distinctive are differences in the respective scores distributions and shapes. While both raters' scores formed a platykurtic, or flat-topped, curve, the SME score curve is somewhat flatter than that of RES. Both score curves are positively skewed, with somewhat more skew in the curve of SME than in RES.

A further t-test, displayed in Table 17 below, indicated no statistically significant difference in the mean scores of all essays by each rater. Thus, it is unlikely that chance
accounts for the strong score similarity. The agreement might be due to the CLEAR team appropriately using discipline-relevant criteria in the essay scoring tool; it might be because of the effectiveness of the essay scoring tool; it might be thanks to the effectiveness of the anchor session and discussion. Perhaps some combination of these factors accounts for the agreement. The results of these tests suggest that CLEAR score users can have some degree of confidence in essay scores by raters trained with the essay scoring tool, even when the raters are not content experts.

<table>
<thead>
<tr>
<th>N</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>8.83</td>
<td>50</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The apparent similarity of the raters' scores is further illustrated in the correlation of essay scores, found in Table 18. The test indicates a rather strong correlation between the scoring of the two raters. Thus the scores of the subject expert and the researcher bear actual agreement. This supports the scoring of content essays by the researcher (and indirectly by the CLEAR team) employing standards of the discipline. The results reflect well upon the foundational work performed by the CLEAR team in developing the essay scoring tool and selecting the criteria reflected therein, as well as the procedures used in the range-finding session. The good results suggest that score users may have some degree of confidence in the meaning of the CLEAR essay subtest scores.
Research Question 2: What is the Validity of Inferences Based on Scores Derived from the CLEAR?

At the heart of test quality lies the issue of validity. Validity has traditionally (Kubiszyn & Borich, 1999, p. 297) been considered a property of a test: whether a test measures what it is supposed to measure. Validity as a function of test score interpretation is the contemporary consensus perspective (Messick, 1989; see Boorsboom, Mellenbergh, & van Heerden, 2004, for a dissenting opinion). Whereas the quality of validity is unitary, validity evidence may be perceived through different facets. Three distinct approaches exist through which a researcher may discern evidence for claims of validity.

First is content-related evidence, wherein the actual content of a test is scrutinized to determine its representativeness with the universe that the content should sample. Second is criterion-related evidence, a systematic connection of test scores with other outcome

Table 18
*Essay Assessment Inter-rater Score Correlations*

<table>
<thead>
<tr>
<th></th>
<th>SME</th>
<th>RES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SME</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=21</td>
<td></td>
</tr>
<tr>
<td>RES</td>
<td>0.8529</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>n=36</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Eight essays were not scored due to evidence of plagiarism, an infraction of the guidelines.*
criteria. Third is construct-related evidence, which may be manifested in a theoretical framework or argumentation or correlation with relevant measures.

**Content-related Evidence for Claims of Validity in the Knowledge Subtest**

Estimating test content validity is a straightforward matter in tests that measure educational achievement. Such a test ought to bear a strong resemblance to the substance of instruction, as stipulated in the test blueprint or in instructional objectives. Kubiszyn and Borich (1998) state, "The content validity of a test is established by examination. Test questions are inspected to see whether they correspond to what . . . should be covered by the test" (p. 298). The subject area expert was consulted to judge the CLEAR test content as well as the stimulus materials. Ideally multiple judges would be engaged for this consultation. Although such was not possible during the dissertation, further consultations could take place in future sessions.

Concerning the textbook chapter and videotaped lecture stimuli *vis à vis* the subject domain, the psychologist attested that these materials sampled the domain appropriately. She furthermore approved the topic and the materials as representative of an introductory university course, particularly at the undergraduate level although not exclusively so.

The cognitive and academic level of presenting the subject matter is acceptable for an introductory course, stated the content expert. However, the language of the test items and directions could be simplified, she and the measurement expert agreed. The
consultants did not specifically recommend linguistic simplification, but did advise research trials of the CLEAR with "average" native English speaking university students who have a "C" grade point average. Indeed, Kubisyn and Borich (1999) point out a shortcoming of content validity analysis: such research "gives information about whether the test looks valid, but not whether the reading level of the test is too high or if the items are poorly constructed. A test can sometimes look valid but measure something entirely different than what is intended, such as guessing ability, reading level, or skills that may have been acquired before instruction" (p. 298, emphasis in the original).

Evaluating the knowledge test, the subject matter expert deemed the content coverage sufficiently broad and representative. The psychometrician questioned how closely the test specifications fit with the content of the textbook chapter and the video lecture: the content judge thought that the CLEAR items corresponded well with the stimuli. She also noted that asking a quantity of items devoted to particular topics is reasonable when both stimuli cover those topics. She proceeded to dismiss examinee complaints that items tested specific knowledge of terms, concepts, scholars. The CLEAR team goal of testing every level of Bloom's (1956) taxonomy is neither necessary or even recommended, said the psychometrician. She suggested collapsing the taxonomy into two simple levels: knowledge/comprehension level and above.

Discussing improving the technical quality of the knowledge test, the consultants recommended increasing the number of knowledge test items. The textbook publisher
might provide a bank of test items, which might prove helpful in inspiring CLEAR item writers. The subject-area expert did not endorse using actual items from this source, questioning the quality of these items. She advised employing more scenario-type items to improve the test. The measurement consultant remarked that several items could be based on a single scenario, creating a "testlet." The testlet approach, used by some psychology professors, would enhance the efficiency of scenario-based items. However, she noted that testlets cannot appropriately be analyzed by standard item analysis procedures since the statistical assumption of item independence is violated.

The psychometrician, moreover, attested that "excellence" is not necessarily "representative." The CLEAR should certainly be a good test, but more importantly, it should resemble tests given to university students. Such is crucial in order that the CLEAR can serve its intended purpose. Certainly the knowledge test would be improved in consistency (i.e., reliability) with more items; nevertheless, the concerns for content-related evidence for validity claims were allayed by the subject area consultant. This content consultant judged the knowledge test appropriate for its purpose in terms of content representativeness and domain sampling.

*Content-related Evidence for Claims of Validity in the Essay Subtest*

Examining the CLEAR dynamic simulation assessment for evidence supporting validity claims must, in the case of the essay subtest, comprise scrutiny of the essay prompt as well as the scoring.
Concerning the essay prompt, the psychology consultant evinced strong support. The prompt "captures the essence" of the reading and the lecture, she stated. The psychometrician commented on measurement, her area of expertise, but did not comment on the subject matter. She noted that different essays would result from a restricted-range prompt as opposed to an open-range one. She advised that the prompt be phrased to structure content response, guiding the examinee to address particular points of basic knowledge as well as crafting an evidence-based argument. The content expert remarked that the current prompt does not dictate responding from a particular position, but a college writer would certainly be expected to cover major elements of the domain in a position paper responding to this prompt. These results will be explored further in the final chapter.

*Concurrent Measures Yielding Evidence for Claims of Validity*

The present research marks the first systematic study of the measurement quality of the CLEAR dynamic simulation assessment. Central to this research is investigation into validity claims, a key constituent of which is the comparison of examinees' performance on the CLEAR with their performance on other measures. Convergent evidence for validity claims from concurrent measures was sought from the examinees' ESL teachers. The teachers' input was sought particularly because of their expertise and insight into the ESL students' effectiveness in learning, for AR is believed to comprise not merely CALP and topical knowledge, but also personal characteristics, academic
skills (reading and writing skills, mathematics and science skills, reasoning skills such as analysis, synthesis, and investigation), and academic auxiliaries (motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies). ESL instructors completed a form rating the examinees on their likely success in university work. On this data collection form, the teachers were directed to base their rating not simply on English proficiency but teacher judgment of students' overall abilities. Teachers were to classify each ESL student into one of four categories: probably succeed, might succeed, might fail, and probably fail. These categories were transformed into a four-point scale, the "teacher rating scale."

Data were collected from 13 different instructors on all of their students; the researcher logged only the scores for the CLEAR participants. One teacher had only three students who took the CLEAR, while other teachers numbered eight or more CLEAR test-takers. Thus, the teacher rankings are a "patchwork" of data. Results are first reported for the knowledge subtest, then for the essay.

| Table 19 |
| Correlations of CLEAR Knowledge Subtest and Teacher Ratings |

<table>
<thead>
<tr>
<th></th>
<th>Knowledge</th>
<th>Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=36</td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>.376</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>n=36</td>
<td>n=36</td>
</tr>
</tbody>
</table>
The validity coefficient for the relationship between the knowledge subtest scores and the teacher ratings of student university success, \( r (36) = 0.38 \), demonstrates a correlation despite the low reliability of the knowledge subtest scores, which would tend to attenuate the strength of an association. Thus, this coefficient quite probably reflects a relationship, not random resemblance. The teacher ratings, therefore, bear a real association with how students actually performed in demonstrating new content acquisition on the multiple-choice knowledge subtest. This association suggests that the ESL teachers effectively judged university demands and student abilities for this task. Such findings attest support for the meaning of the knowledge subtest scores.

In terms of the essay, correlations of essays and teacher ratings reveal no statistically significant association. An association had been anticipated in the present study, as agreement had been seen in pilot studies. The lack of relationship might be due firstly, to statistical limitations or secondly, to the basis of the teacher ratings. Firstly, statistical factors may have suppressed variability by means of the limited four-point scales used for the teacher ratings and for the essay scores. Secondly, the teacher ratings might have been based on factors more aligned with student performance on selected-response tests rather than performance in writing a content-responsible essay. Some implications of these findings are considered in the final chapter of this study.

Table 20

*Correlations of CLEAR Essay Subtest and Teacher Ratings*
Having examined associations between the knowledge subtest scores and the essay subtest scores with the concurrent measure of teacher ratings, this concludes the data analysis relevant to criterion-relevant evidence for validity claims. Construct-related evidence for claims of validity will be discussed in the ensuing section.

**Construct-related Evidence for Claims of Validity**

In responding to Research Question 2, concerning validity claims for the CLEAR, this report has addressed content and concurrent aspects of validity. Attention now turns to examining claims of validity in light of construct-related evidence. As noted elsewhere, foundational evidence for the construct of AR was laid in the literature review earlier in this report; the overall quality of this research might lend qualitative support, buttressing construct-related evidence of validity claims. Additional, quantitative, affirmation may be revealed in results presented in this section. First reported are results of correlating CLEAR subtest scores with placement test scores and proficiency test scores.
scores. These correlations sought discriminant evidence, as these tests seem to measure constructs different from AR, as measured by the CLEAR. Following such will be reported the second calculation to test evidence for construct-related validity claims, an inter-subtest correlation: this correlation estimates the overlap between the knowledge and the essay subtests.

The first test to provide construct-related evidence for validity claims consisted of correlations calculated between scores on the CLEAR subtests and scores on the TOEFL proficiency test and the MTELP placement test. Results were anticipated to show discriminant evidence, little or no correlation between the CLEAR and the external measures. Discriminant evidence would suggest support for the construct of AR as distinct from the qualities tapped in the external measures.

Results (reported in Table 21) demonstrate that the knowledge subtest scores are not associated with the MTELP placement test at a statistically significant level; the apparent relationship indicated in the validity coefficient might be due to chance. However, a strong correlation is seen between the knowledge subtest scores and the TOEFL proficiency test scores, r(28)=.58. Such a validity coefficient very likely represents an association, despite the weak reliability of the knowledge subtest. This analysis was conducted for discriminant validity evidence, expecting to find little association, and the outcome was not expected.
The correlation, however, may indicate that the selected-response testing mode employed in the knowledge test and the TOEFL draws upon similar abilities. That is, multiple-choice test scores might demonstrate knowledge of test-taking strategy, or knowledge of language features, or content knowledge learned through language. However, selected-response subtest scores do not represent the full scope of an examinee's AR, as will be demonstrated in the section under construct-related validity evidence. Moreover, the scores do not undercut the value of the dynamic simulation testing approach: the powerful effect of taking the CLEAR was reported by many test-takers, as will be recounted later under Research Question 3.

<table>
<thead>
<tr>
<th>Table 21</th>
<th>Correlations of CLEAR Knowledge Subtest and External Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Knowledge</td>
</tr>
<tr>
<td>Knowledge</td>
<td>1.00</td>
</tr>
<tr>
<td>MTELP</td>
<td>.16 n=19</td>
</tr>
<tr>
<td>TOEFL</td>
<td>.58 n=28</td>
</tr>
</tbody>
</table>

Proceeding to the correlations between the essay scores and external measures (displayed in Table 22), results reveal findings in accord with expectations: the essay scores were not strongly correlated with the other measures. That is, only weak to moderate correlations exist between the essay and the MTELP placement test, and between the essay and the TOEFL proficiency test. These findings were anticipated, for
these tests measure different constructs and do so from different approaches. This negative evidence therefore provides discriminant support for the claims to construct validity made on the basis of CLEAR essay scores.

Table 22
*Correlations of CLEAR Essay Subtest and External Measures*

<table>
<thead>
<tr>
<th></th>
<th>Essay</th>
<th>MTELP</th>
<th>TOEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essay</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTELP</td>
<td>.24</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>n=20</td>
<td>n=19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEFL</td>
<td>.37</td>
<td>.41</td>
<td>1.00</td>
</tr>
<tr>
<td>n=20</td>
<td>n=12</td>
<td>n=28</td>
<td></td>
</tr>
</tbody>
</table>

Having considered discriminant evidence for construct-related validity claims from external measures, attention now turns to examining the internal evidence within the CLEAR for this aspect of validity. Such was probed by an inter-subtest correlation, seeking evidence for a relationship between scores on the knowledge subtest with scores on the essay subtest. The results are summarized in Table 23.

Table 23
*Intercorrelation Between CLEAR Subtests*

<table>
<thead>
<tr>
<th></th>
<th>Knowledge</th>
<th>Essay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>n=36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>n=28</td>
<td></td>
<td>n=28</td>
</tr>
</tbody>
</table>
The correlation results demonstrate a distinct relationship between the scores of the subtests, \( r(28)=.50 \). The strong correlation coefficient indicates that the subtest scores likely bear a real association to one another. The level of association suggests some shared variance, but not complete redundancy, between the two subtests of the dynamic simulation. These findings evince construct-relevant support for validity claims, a central concern in gauging the quality of a test.

In order to understand better AR as measured in the CLEAR, an additional correlational test was conducted. Examined herein were the relationships of the CLEAR simulation subtests with static tests of English proficiency (the TOEFL university admissions test and the MTELP test for ESL placement), along with the teacher ratings, and the examinee gender. The numerical data are reported in Table 24.

<table>
<thead>
<tr>
<th></th>
<th>Knowledge</th>
<th>Essay</th>
<th>TOEFL</th>
<th>MTELP</th>
<th>Teacher</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay</td>
<td>.50</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEFL</td>
<td>.58</td>
<td>.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTELP</td>
<td>.16</td>
<td>.24</td>
<td>.41</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=19 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=19 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=12 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=19 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>.38</td>
<td>.22</td>
<td>.41</td>
<td>.19</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=19 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>.33</td>
<td>.28</td>
<td>.23</td>
<td>.58</td>
<td>.27</td>
<td>1.00</td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=28 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=19 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n=36 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

179
The network of relationships may be analyzed as follows to understand better the meaning of CLEAR scores. Thus, the CLEAR knowledge subtest and essay subtest overlap, though not fully. The CLEAR knowledge subtest and the TOEFL share variance, and both share a portion of teacher ratings. Examinee gender is associated with knowledge scores, although not to a great degree. Examinee gender seems to be the only factor which links MTELP placement test scores with the rest of the scores. MTELP is not associated with any other factor. Thus the knowledge subtest appears to act as a nexus: with this subtest, all the other measures are directly or indirectly associated. These matters will be further considered in Chapter 5.

Research Question 3: What is the Examinee Perception of the CLEAR?

The final research question concerns the examinees' perception of the CLEAR. Examinees were asked their impressions on various matters related to test content, difficulty, verisimilitude, and administration. A tension exists between the ideals of perfect measurement, on the one hand, and respecting the time constraints of examinees and institutes on the other. The examinees are key stakeholders, so their input is important; moreover, they are the only persons who can shed light on the experience of taking the CLEAR. For these reasons, the perceptions of the examinees are useful in assessing the quality of the test. These data were collected at the end of the first day of
the CLEAR, and at the very end of the dynamic simulation. Responses were grouped thematically and counted; representative responses appear below the charts. Results are presented generally in the same order as items are presented on the feedback form. Two exceptions are made: appearing at the end of this section are the open-ended items, number 2 ("Comment on any section of the test") and number 10 ("If you have any other comments about the CLEAR or the test process that you would like to share, please comment below").

In Feedback Item 1, examinees were asked to rate the difficulty of the video lecture, the readings, the knowledge test, and the essay test. Statistical analyses in Table 25 represent the distribution of ratings of video difficulty on the first and final days of the simulation. (Note that examinee difficulty ratings for the knowledge test and the essay test were collected at only one point, after they had actually taken these tests.)

<table>
<thead>
<tr>
<th>Data point</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR start</td>
<td>36</td>
<td>2.53</td>
<td>0.99</td>
<td>2.50</td>
<td>2.00</td>
<td>1-4</td>
<td>-0.99</td>
<td>0.01</td>
<td>15.17</td>
<td>35</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CLEAR end</td>
<td>35</td>
<td>2.68</td>
<td>0.93</td>
<td>3.00</td>
<td>2.00</td>
<td>1-4</td>
<td>-0.91</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Four-point scale ranges from 1=easy to 4=difficult.

Responses to Item 1a form score distribution curves that diverge from normal firstly in the difference between median and mean scores, and secondly in kurtosis, sketching platykurtic, or flat-topped, curves. The statistically significant difference
(p<.0001) in mean scores indicates that very likely a real difference exists: that at the CLEAR simulation end, examinees considered the lecture more difficult than they had on the first day. This changed perception may result from heightened awareness after taking the two content-responsible subtests. Perhaps, upon taking the two subtests, the examinees better appreciated that careful listening and notetaking in university lectures are crucial for success in content-based tests.

Feedback Item 1b responses change shape at the two data points, as shown in Table 26. Examinees' first impressions of reading difficulty describe fairly normal, if somewhat platykurtic, score distribution curves. The second response curve is positively skewed and more platykurtic than initially. The mean score at the second data point is .13 lower than at first, showing examinees evaluated the reading as easier after the CLEAR was ended. T-test results show this difference is statistically significant (p <.0001). The results thus show that at the end of the simulation, these examinees retrospectively found the reading text easier than it first appeared. Possibly they were initially discouraged at the length and substance of the text, but study time made the content more understandable.

While examinees' judgment of difficulty changed to a statistically significant degree for both the video lecture and the reading, direction of change differs. That is, by the end of the dynamic simulation, the lecture seemed more difficult whereas the reading seemed less so. Perhaps this direction is amenable to change, if examinees are focused
and prepared for the CLEAR. Indeed, such data may be useful to relate to examinees in preparing them for the CLEAR simulation, and could be of interest to ESL teachers as well.

In Table 27 are reported the findings from Feedback Item 1c, which queried examinees on their estimation of the difficulty of the knowledge subtest. As previously noted, only one data point is relevant for this item. Most examinees rated the knowledge subtest as rather difficult: mean, median, and mode are at or near the 3 rating. That examinees considered the knowledge subtest so difficult might be attributed to their lack of practice in English-mediated content learning that is followed by content-based testing. This supposition is supported by feedback reported later in this section.

Table 26
*Feedback Item 1b: Examinee Ratings of Difficulty of Readings*

<table>
<thead>
<tr>
<th>Data point</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR start</td>
<td>36</td>
<td>2.55</td>
<td>0.81</td>
<td>3.00</td>
<td>3.00</td>
<td>1-4</td>
<td>-0.34</td>
<td>-0.01</td>
<td>18.96</td>
<td>35</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CLEAR end</td>
<td>35</td>
<td>2.42</td>
<td>0.92</td>
<td>2.00</td>
<td>2.00</td>
<td>1-4</td>
<td>-0.65</td>
<td>0.22</td>
<td>18.09</td>
<td>35</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note:* Four-point scale ranges from 1=easy to 4=difficult.

Table 27
*Feedback Item 1c: Examinee Ratings of Difficulty of Knowledge Test*

<table>
<thead>
<tr>
<th>Data point</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR end</td>
<td>34</td>
<td>2.94</td>
<td>0.78</td>
<td>3.00</td>
<td>3.00</td>
<td>1-4</td>
<td>-1.30</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Note:* Four-point scale ranges from 1=easy to 4=difficult.
Next are the examinee ratings of the essay difficulty, displayed in Table 28. Examinees found the essay quite challenging. The strong negative skew describes a curve with scores heaped towards the upper end. Such is borne out by the high mean score, 3.23. The impression of difficulty may be due to a number of different factors: essay writing is complex; writing under a time limit in class is a challenge; composing a content-responsible essay is demanding. Performing all these in a second language is indeed a difficult task.

<table>
<thead>
<tr>
<th>Data point</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR end</td>
<td>35</td>
<td>3.23</td>
<td>0.87</td>
<td>3.00</td>
<td>3.00</td>
<td>1-4</td>
<td>1.55</td>
<td>-1.31</td>
</tr>
</tbody>
</table>

Note: Four-point scale ranges from 1=easy to 4=difficult.

The final section of Feedback Item 1 concerns examinee ratings of overall difficulty of the dynamic simulation. The mean, median, and mode scores cluster near 3; the moderate negative skew and kurtosis indicate a somewhat narrow peaked score distribution towards the upper end of the scale. All these data support results from other elements of this feedback item: test-takers judged the CLEAR as a challenging assessment.
Recalling that Feedback Item 2 is considered at the end of this section with Item 10, attention turns to considering Feedback Item 3. In this and subsequent items, examinees' remarks will be reproduced to illustrate the themes identified. Examinee statements are presented in order of the randomly assigned code number; following each statement is the code number for the examinee being quoted. The quotations are corrected for spelling and punctuation only where necessary for clarity. Any verbiage inserted by the researcher is set off by square brackets. Examinees sometimes covered several themes in one response, yielding more total responses than actual examinees.

Item 3 invites test-takers to complete the sentence starting, "From taking the CLEAR, I learned . . . ." Twelve examinees commented on the closeness of the CLEAR dynamic simulation with a real university experience. A few quotations will illustrate this theme. Example statements include, "[I learned] how the academic exams will be," (914); "Study hard is a powerful tool to pass a test, because I would like to turn back time and review the author's theories" (906); and "[I learned] how regular USA college classes will be to us." (933). Another strong theme, indicated by 11 comments, was the need for language skill improvement made evident after the CLEAR dynamic test. "I

<table>
<thead>
<tr>
<th>Data point</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>median</th>
<th>mode</th>
<th>range</th>
<th>kr</th>
<th>sk</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR end</td>
<td>29</td>
<td>3.07</td>
<td>0.84</td>
<td>3.00</td>
<td>3.00</td>
<td>1-4</td>
<td>-0.39</td>
<td>-0.52</td>
</tr>
</tbody>
</table>

Table 29 Feedback Item 1e: Examinee Ratings of Difficulty Overall
have to improve my writing skill, because I present a lot of difficulties when I have a
time limit" (908); "Reading skill is much more important than I thought to take a lecture
at university" (912). The third theme, noted by eight comments, focused on the value of
study and understanding: "Reading before class is not enough. Student should read,
lecture and back to reread what in lecture and in book to know concept" (924). Another
said, "I can understand a little more than I thought" (989). Some remarked on the need for
study skill improvement: "I need to go to the study hall!!" (903) and "I need to improve
my learning and writing" (931). Another response set discussed learning psychology, the
academic topic of this CLEAR dynamic simulation. Three examinees did not respond to
the item.

Figure 6
Results from Feedback Item 3, "From Taking the CLEAR, I Learned . . . ."

Note; Examinees may have given more than one response.

Asked wherein examinees best performed on the CLEAR simulation in Feedback
Item 3, 18 remarks addressed subtests. Representative remarks commenting on the
knowledge test follow: "...knowledge, because it's easier to answer questions than to write an essay" (902); "...knowledge test, because it's not difficult. Just I was confused" (913); "...knowledge test, because I understand and learned what I read and listened" (922). Those who judged the essay performance best wrote such comments as these: "...essay, wrote my best, used the others speech" (904) and "...essay, because I could response what I understand from the lecture and think about what the chapter says" (907).

Figure 7
Results from Feedback Item 4, "I Did Best on the ___ Section, Because . . . . "

Note: Examinees may have given more than one response.

A second theme, represented by 14 comments, was the focus on language skill as the best CLEAR performance. Some feedback thus categorized includes: "...reading, because I can see all words of conversations." (927); "... reading, because many hours
and a interesting subject." (933); ". . . video lecture, because I'm good in listening." (928); ". . . video lecture, because Professor had interesting styles and not so hard, not so easy." (909); and ". . . video lecture, because I took good notes. " (919) The final category, labeled "Global", includes these two statements: ". . . overall, because I want to be a student in USA." (915) and ". . . Nothing is best for me." (935).

Feedback Item 5, summarized in Figure 8, invites examinees to state which university admissions test they would prefer and explain such preference by completing the following statement: "If I could choose between taking the CLEAR and taking the TOEFL for admission to this university, I would choose ____ because ______." Preference for the CLEAR based on test quality was the largest group, represented by 10 comments. Typical remarks are: "CLEAR, because I think it more useful than TOEFL to students who want to go to a American college." (907); "CLEAR, because it concern about how a student working." (921); "CLEAR, because CLEAR test, tests how much students understand, not how much students memorize." (922); "CLEAR because it's more real." (925); and "CLEAR, because it could judge my academic ability. "(933). Other comments favoring the CLEAR were based on the CLEAR being easier, as in "CLEAR because TOEFL is more difficult and have many sections. I think CLEAR test is more helpful for me." (920). A few examinees expressed a preference for the CLEAR but were silent on the reasons.
Preference for the TOEFL was widespread and for an assortment of reasons. Test quality did not appear as a reason. Seven persons favored the TOEFL, considering it easier than the CLEAR: "TOEFL, because I think it is easy than CLEAR." (935). Four test-takers preferred the TOEFL because of the test conditions: "TOEFL because I have more time to study at home by myself." (901); and "TOEFL, because it [. . . ] takes one day to finish it while the CLEAR takes 3 days." (919). The widespread acceptance of the TOEFL was noted by four examinees: "TOEFL because its the popular admission to the universities." (931). Familiarity was the reason four other examinees listed: "TOEFL, because I know TOEFL styles very well but CLEAR is good but the test maker has to make sure that he or she explain about the test well enough." (909).
Examinee impressions were solicited concerning the duration of the dynamic simulation by completing the prompt "Taking this test over a 48-hour period was . . ." The largest response group, 20 comments, addressed the quality of the CLEAR experience. Eight of these said it was interesting or exciting; Representative statements are: "Kind of exciting and tiring." (907) ; "An interesting experience, you learn a little bit more of colleges methods." (908); "Good. I found that I have to pace myself all the time." (909); "So far so good at first then a little tired, finally interesting." (929). Also in this theme are the seven calling the duration tiring: "A long run with more experiences and stresses." (910) and "Make me very tired because I don't have enough time to practice in lecture. I have more time to practice I might do it better. But I feel very happy in CLEAR test." (924). This category includes five remarks on the challenge: "Hard work for me." (905) and "Difficult for me." (911). The next largest theme, with 11 remarks, focused on the sufficiency of time. Six persons said the time was too short, while five said it was too long. Some examples from this category are: "Too short for me to study all details." (915) "Too long." (918, 936). Responses classed as "Acceptable" or "No Preference" were variously expressed as "Fine. (922) "Okay." (930) and "I think it has any influence." (902).
In Feedback Item 7, examinees completed the statement, "If I could change one thing about the test or the test administration, I would change _____." Two important themes were represented with 12 comments each on test components or test conditions. Example remarks on test components follow: "The questions of the knowledge test should be more general." (906); "Maybe, not to be so specific in details and focus more on the main ideas in the multiple choice section." (908); and "Essay section." (914). Comments related to test conditions include: "Bring the book home and study. But cannot bring the book [out of] the test room." (901); "If sample questions about knowledge test could be given before reading section it would be perfect for us to figure out what we need to focus while reading." (922): "I want to use my note during the knowledge test."
Nine responses suggested changes in the stimulus materials. "Knowledge test." "Video lecture into a real one. A more difficult topic." (929); "Video lecture. I will invite a professor. I will change real lecture." (932); "No essay." (935). Three comments expressed that the CLEAR should not be changed at all. Examples are "Nothing I think it is perfect." (924) and "Not change, but try to be more common." (926). Three feedback forms had no response.

Figure 10
Results from Feedback Item 7, "If I Could Change One Thing . . . ."

![Bar chart showing responses]

Note: Examinees may have given more than one response.

The verisimilitude of the CLEAR dynamic simulation was the focus of Feedback Item 8, depicted in Figure 11. This item asked examinees to respond "Yes" or "No" to the
following prompt, then explain why: "The tests I took gave me a feel for what I will have
to do in the university." Affirmative responses totaled 28, sharply greater than the 7
negative responses. Some examinees attesting the CLEAR-university resemblance
explained their reasoning on the basis of external features such as task elements, as seen
in the following statements: "Yes. I understood I will need to read a book, not article, it
means so long reading." (912); "Yes. It's close to university approach of class and
research." (918); "Yes. It like a research. Read what other researcher did in this area and
present your question or opinion.' (921); "Yes. Test is almost similar to university studies.
CLEAR is like a demonstration of university studies." (922); and "Yes, because is the
real world. The university is like this." (925). Other affirmations of the CLEAR
verisimilitude with the university experience are based on the internal qualities of the
respondent, such as these examples: "Yes. I need to work hard with writing and listening
to lecture." (915); "Yes. You see the way to study." (917); and "Yes. I need to study more
and prepare myself to confront with many many information that I will get from class.
"(935).

Seven examinees asserted no resemblance between the CLEAR dynamic
simulation and the university. The support offered for this assertion is not founded on
objective evidence. Three examples may illustrate this theme: "No. Because I didn't like
the main topic." (903); "No. Because I couldn't remember about whole lecture. So, I
didn't chose clear answer." (920); and "No. My major is completely different. I think college topics are more difficult." (929).

Responses were markedly positive to Feedback Item 9, asking examinees whether they would recommend a friend take the CLEAR. As shown in Figure 12, "Yes" responses totaled 31, "No" answers totaled 3, "I don't know" came to 2. Of examinees who would recommend the CLEAR to a friend, 14 based their reasons on the good experience or verisimilitude. Some example comments are: "Yes. It can be a good experience for him to understand the university lectures and exams." (904); "Yes. CLEAR displays the reality of university." (922); "Yes. It is an excellent challenge and it is a good option to engage the capacity." (925); "Yes. It is good opportunity for ESL
students to know what's going to happen in college or university." (930); "Yes. Because it will be good experience." (934). Another 8 comments related to the personal growth or self-awareness resulting from the CLEAR dynamic simulation, as these statements indicate: "Yes. I realized I need more works to enter and adjust to a college in USA." (915); "Yes. Because it is useful to improve skills and knowledge." (916); "Yes. It's good to test yourself." (929). Five remarks focused on comparing the CLEAR and the TOEFL: "Yes. I have many friend that prefer to study at LSRU, but they were struggled with a restrict in TOEFL." (910); "Yes. Good idea and easier than the others." (926); "Yes because better than TOEFL." (927). Four affirmative responses did not specify their reasons why.

As noted above, three examinees said they would not recommend the CLEAR to a friend. One person gave no reason; the other two remarks are: "No. TOEFL has strategies, easier to study for the test." (912) and "No, because [the CLEAR] is too long." (928). Two examinees did not know if they would recommend the CLEAR and offered no explanation.
Two feedback items remain in the data analysis responding to the research question probing for examinee perceptions of the CLEAR dynamic simulation. Feedback Items 2 and 10 are now analyzed: both items gave the examinee the choice of topic. Item 2 read, "Comment on any of the sections of the test." Item 10 stated, "If you have any other comments about the CLEAR or the test process that you would like to share, please comment below."

Item 2, displayed in Figure 13, drew considerable attention across a broad variety of themes. All but five examinees addressed this item. The largest response theme (15 remarks) concerned the university-type learning of the CLEAR, different from ESL learning. Example statements are: "The knowledge test was not exactly what I was
expecting. Most of the questions were specific questions about author's names and not about what I could understand from the text." (906). "The words in the question of the essay should be easy to understand because if a professor wants me to write an essay I would use lots of dictionaries, internet, library. If the goal of this test is test students as in real life, the language should be understandable." (922); "In essay, it's very difficult to remember everything and make in one essay even though I can open text book" (935).

The next three themes comprise examinee attention to elements of the CLEAR dynamic simulation. The second-largest theme, termed "text/video/notetaking," consisted of 10 examinee comments on interacting with CLEAR simulation stimuli. Representative quotations are: "I think is difficult take good notes while we are watching the video." (903); "If I were in real lecture situation, I would understand easier than the video lecture." (907); "The text book was not very difficult, but very long." (912). The third theme was represented by nine remarks focused on the essay component of the dynamic assessment. Examinee feedback included these statements: "I'm not strong at writing, so the CLEAR was difficult for me." (905) and "Essay test - I think that if we didn't use anything (dictionary, reading book, and source) during the test, it is better for me." (913). Nine other remarks, in the fourth theme, focused upon the knowledge test. Examinees made such comments as: "There were lots of details on the knowledge test. I should know that before I took it." (909). The challenges of the CLEAR are the central points in the fifth and sixth theme, regarding "time pressure" and "test difficulty" respectively.
Sample data from the "time pressure" theme are: "It's like cram before taking. It should take more time to review book. If you want to evaluate the ability to study in university." (901); "Essay need more time especially if we have to make citation that we have to look for in the chapter." (918). A representative comment of the "test difficulty" theme is: "Most sections were difficult to me." (933). A few examinees responded to this item with complimentary evaluation of the CLEAR. Four responses form the last theme, compliments. One such remark was: "Test in multiple choice is very good they are all concept in lecture" (925).

Figure 13
Results from Feedback Item 2, "Comment on Any Section of the Test."

Note: Examinees may have given more than one response.
Feedback Item 10, inviting additional comments, drew less attention than other items, as may be seen in Figure 14. This might be because of respondent fatigue or urgency to complete the feedback form. The largest theme was comprised of six comments, which were classed together under the label "test conditions." Sample remarks are: "I was very confused during the essay test. Because I'm not accustomed to using the source during the test." (913); "Study these study text and try to memorize those text within one day these included many person and their report. It's not very easy. But in Master or university, when you do research, you can check any information but you must know where you can find the information what you need. You needn't memorize but you need to know where you can find the information." (921); "If sample questions about knowledge test could be given before reading section it would be perfect for us to figure out what we need to focus while reading." (922); and "The examiner should better what we will do in the test. Cause in the beginning all of us have no idea what we should until test." (931). The next theme responding to Item 10 was termed "compliments." Three examinees wrote such remarks as: "I feel very happy in CLEAR test." (923) and "Thank you!" (930). Two comments addressed the relevance of the CLEAR dynamic simulation assessment. these statements are: "All the test it's focus in a topic that could not be relationated with what the student want to study." (906) and "Taking the CLEAR was
very good experience for me. I could know about the lecture and test at university in the U.S." (913).

Figure 14
Results from Feedback Item 10, "Other comments"

Note: Examinees may have given more than one response.

Above has been reported a rich source of examinees perceptions concerning the CLEAR. The examinees are ambivalent about the test duration: time for study is desirable but the simulation is an demanding experience. Test quality was a CLEAR feature, and so was the verisimilitude of the simulation. Examinees stated that they valued the dynamic simulation experience and would recommend a friend take the CLEAR. The CLEAR was, for many, an event which affected their attitude towards studying, bettered their self-assessment, and increased their motivation. The results present generally strong support for the CLEAR, and seem to indicate that the assessment strikes a reasonable balance between ideal quality and efficiency.
In conclusion, this chapter has presented results from analyses into the measurement quality of the CLEAR dynamic simulation. After having found sample equivalence across testing sites, research attention turned to determining equivalent test functioning across subgroups. Results of item analyses were reported, specifically the \( p \)-value, discrimination index, and distractor analyses for the knowledge subtest. Attention next turned to answering Research Question 1, What is the consistency of the CLEAR Test? Here the knowledge subtest consistency was analyzed in terms of Cronbach's alpha, whereas the essay subtest consistency was measured by essay scorers' inter-rater consistency. Descriptive statistics for the overall knowledge test were reported, as was the standard error of measurement (SEM). Next was examined data to answer Research Question 2: What is the evidence for the validity of inferences based on CLEAR Test scores? Content-related evidence for claims of validity was provided by a content expert evaluation. Criterion-related evidence for claims of validity was reported by correlating concurrent measures of ability (placement test scores, proficiency test scores, and teacher ratings) with the two CLEAR subtests. The third aspect of validity to be investigated was construct validity, analyzed by inter-correlation of knowledge subtest and essay subtest scores. Completing this chapter was the report of examinee feedback to answer Research Question 3: What is the examinee perception of the CLEAR? These data help illuminate the relationship between test effectiveness and efficiency. The implications of these findings will be explored in the last chapter of this report.
CHAPTER 5

DISCUSSION

Introduction

This investigation was conducted to estimate the measurement quality of the CLEAR. Years of developmental research have been invested into this test for university admissions purposes. The present instrument seeks to elicit evidence of AR by means of dynamic assessment (Grigorenko & Sternberg, 1998). CLEAR examinees experience a simulation of the university experience, where learning opportunities are followed by testing how well students learned and understood the academic content.

Despite advances made in the test development, only now has it been possible to commence a validation study, a key step in the test development cycle. This initial validation is crucial to shed light on the test strengths and weaknesses, thereby helping point the direction to future research and development for the CLEAR test.

This chapter will summarize research findings following each section with a brief discussion. The ensuing section will discuss in more detail results of particular interest.
Discussion of Findings

Preliminary findings indicated the appropriacy of combining data from testing sites. Although a larger participant group would be preferable, the present sample (N=36) was acceptable for many analyses in this initial validation study. Other preliminary findings supported decisions to offer the CLEAR to ESL students at the highest- and second-highest proficiency levels. The participants' diverse L1 backgrounds yielded a linguistically heterogeneous sample, a positive feature for an initial validation study, yet the low numbers from each L2 did not permit linguistic bias investigation.

The preliminary question of similar performance by gender led to additional analyses. Differential performance by gender group was seen in the knowledge subtest wherein mean scores were 2.68 points higher for males than females. However, no statistically significant difference was found in the t-test, $t(34)=-2.05$, $p=0.0477$. In the essay, males earned scores 0.5 points higher than did females, a score difference that was not statistically significant in the t-test, $(26)=-1.46$, $p=0.1550$. These findings indicate that the visually better scores by male students could be a chance anomaly, and no relationship should be assumed.

Nevertheless, interest compelled examination of these participants' performance on other measures. In terms of TOEFL scores, the pattern of lower scores by these female
examinees seemed to continue. Females' mean score (491.27) was lower than males' mean score (509.76), a difference of 18.49 points. The TOEFL score difference by gender was not statistically significant, \( t(26) = -1.23, p = 0.2291 \). A similar pattern was seen in the teachers' mean ratings of examinee likely success. The mean ratings for females (2.5568) was lower than mean rating assigned to the male participants (3.0786), by .893 points on the four-point scale. This apparent difference was not statistically significant, \( t(34) = -1.63, p = 0.1120 \). The pattern continued in scores these participants earned on the MTELP placement test, where females' mean score (33.778) was 11.52 points lower than that earned by males (45.3). In the MTELP scores by gender, therefore, was found the first statistically significant difference in group performance, \( t(17) = -2.91, p = 0.0098 \).

Taken together, these findings do not show a pattern of real differences between the performance of the males and females in the sample. Indeed, superficial score differences exist, but these are borne out as statistically significantly different only in the MTELP measure. Nevertheless, a sample of only 36 participants is might be susceptible to sampling bias, particularly research where the sampling was without random assignment. Further investigation, with a larger sample, seems warranted to ensure the fairness of the CLEAR to both males and females. Scores derived from the CLEAR simulation, and indeed from every test, should distinguish differences in the construct under examination: score differences should not be due to construct-irrelevant variance. The matter of gender-related performance differences will be considered again in the
discussion related to Question 3, and evidence for validity claims.

Proceeding with the next item of preliminary study, the knowledge subtest item analyses provided data useful for the test developers. Item statistics showed a very broad range of item difficulty values or \( p \)-values. Items with \( p \)-values around .50, for instance, are answered correctly by half of the examinees. In norm-referenced tests such as language proficiency tests or IQ tests, item difficulty ideally hovers about .50; such difficulty values will permit the greatest spread of scores among examinees. This might seem problematic for the CLEAR, with item difficulty values from .06 to .94, a range wider than would be desirable in a norm-referenced test. As the CLEAR is a criterion-referenced test (CRT), this feature is not necessarily a flaw when the knowledge subtest items closely conform to the blueprint for the dynamic simulation.

With CRTs, the item difficulty standards are different. Since these tests are intended "to assess performance on a set of tasks representative of a well-defined domain" (Crocker & Algina, 1986, p. 329), they do not need to maximize the dispersion of scores and show differences among examinees. Thus, the 50% difficulty standard is not necessarily requisite in a CRT such as an educational achievement test. Crocker and Algina (1986, p. 330) advise that \( p \)-values be reported as the mean difficulty of all items measuring one objective. This suggests a procedure for future analyses of the CLEAR, to categorize items into objective-based clusters and calculate the \( p \)-value of the cluster mean difficulty. Such a technique might better illustrate the functioning of the CLEAR,
Continuing with quantitative item analysis, a point-biserial correlation was conducted to express the relationship between response to an item and responses to the whole knowledge subtest. The resulting coefficient, called "item discrimination," or "discrim," revealed how well items differentiated high-scoring and low-scoring students. That is, an item discriminates positively when, for example, strong examinees answer correctly and poor examinees do not. Negative discrim occurs when strong test-takers miss an item but weak students answer correctly. The item discrimination analysis revealed that many items discriminate positively, while five items correlated negatively with the overall knowledge subtest. (These items are numbers 2, 6, 21, 23, and 28 in the published numbering system.) Items that discriminate negatively might have been badly written or even scored incorrectly. While these items had all been scored correctly, the negative discrimination might be due to confusing wording. Removing or revising items that discriminate negatively is requisite, and easily accomplished. All five negatively discriminating items were recommended for omission from the live version of the test pending revision. In general, however, improving item discrimination will be problematical because of the high $p$-values found in this test. The matter of item discrimination arises again in the discussion on consistency, in Question 1.

The final component of the quantitative item analysis was the distractor analysis, showing response patterns by strong, middling, and weak students in terms of their
overall knowledge subtest scores. Thus, which distractors were and were not attractive to unsure examinees was evidenced through the distractor analysis. These data are useful to the test developers, who can invest effort into improving the plausibility of unattractive distractors. Rewriting these unattractive distractors will quickly improve the measurement quality. Distractor analysis may also prove helpful when making a pass decision about a borderline examinee. The distractor analysis, like the other preliminary analyses, indicated areas in need of improvement while also illuminating where the CLEAR dynamic simulation is functioning acceptably. Such information is useful to the test owners in planning item revision, the next step in progressing through the stages normally followed in the test development cycle.

Attention may now focus upon Research Question 1, concerning the consistency of the scores earned in the CLEAR simulation. The knowledge test scores, according to basic statistical analysis, describe a relatively normal curve, if somewhat negatively skewed. A normal curve distribution is not unexpected in CRT scores; the restricted range of the sample likely produced the score curve piled toward the high end of the scale (cf. Kubiszyn & Borich, 2000, p. 319). Quantified by the coefficient alpha of .528, internal consistency is low. Such may be improved with eliminating negatively correlating items, as will be discussed below. The standard error of measurement, a function of the standard deviation and the alpha, is 2.50. This statistic estimates the difference between an examinee's observed and true score. The SEM statistic is useful to
the CLEAR test committee in making pass decisions. Scores within the SEM, a range of +2.5 to -2.5 points, should be considered essentially the same. Of course, as a function of the test consistency, the SEM will narrow when the test consistency is improved.

Improving the internal consistency is a concern raised by these findings. The alpha is not too poor considering how few items are in the test; yet, improved internal consistency ought to be a goal for the next stage of test development. The recommended internal consistency level is not a stipulated level, but a matter of principles for the test designer to consider, according to several reference works consulted (AERA et al., 1985; Alderson, Clapham, & Wall, 1995; Bachman, 1990; Bachman & Palmer, 1996; Crocker & Algina, 1986; Hughes, 1989; Kubisyn & Borich, 2000). Bachman and Palmer (1996) note three factors in the test designer’s decision to set a particular reliability level. The first factor is test purpose: a high-stakes test ought to have higher reliability than would be tolerable in a lower-stakes test (p. 135). Two other factors affect the limit of reliability, first, how narrowly or broadly the construct is defined and second, the nature of the test tasks (p. 135). In the CLEAR, the broad construct of AR will likely not be measured as reliably as would a narrower, more tightly construed trait. The CLEAR selected-response measure is a fairly consistent, uniform testing task, from which could be expected higher levels of reliability. The CLEAR very likely will never attain the high reliability of the TOEFL. Since reliability and validity exist in tension with one another, such a goal may be unreasonable for a test constructed upon concern for validity (cf. Alderson, Clapham,
& Wall, 1995, p. 187). Taking the above into account, a recommendation for the test developers is to target the CLEAR knowledge subtest for a .75 to .85 reliability coefficient.

Steps to improve internal consistency might begin by removing items that discriminate negatively, with degrees of improvement according to which item is deleted. However, not all negatively correlating items can be removed. Further analyses showed that removing certain test items affects how the remaining items correlate to the total. Such results suggest that certain test items function as sets or clusters (a question which might be later investigated by means of logical or factor analysis, as will be discussed later). This functioning is not necessarily a problem, in a criterion-referenced test wherein items are matched to assorted blueprint criteria. Indeed, item clustering behavior might illuminate one reason for the poor alpha: a test comprised of multiple clusters of items will not be as internally consistent as a test measuring one single trait. If the CLEAR dynamic simulation is indeed heterogeneous, alpha perhaps ought not to be based on the overall knowledge subtest, but more reasonably based on item clusters tied to the test blueprint, as Crocker and Algina (1986, p. 330) recommend for CRTs. Certainly, crucial and immediate means of test improvement are to increase item quantity and to better the quality of distractors.

Having now considered the issue of knowledge subtest consistency, the essay subtest may be examined. In this constructed response measure, consistency is a matter of
comparing the scores assigned by the essay raters. The content expert approved the CLEAR essay scoring tool, so this instrument was employed without any modification. Judging began with a range-finding session to practice applying the scoring tool criteria with essays from CLEAR pilot data. Following this, the judges worked with papers from the present study. The expert consultant graded 21 essays and the researcher, 36. Ratings showed good agreement. Mean scores were 1.42 and 1.43, respectively; no statistically significant difference ($p<.0001$) was found between the two mean ratings. Additionally, correlation showed scores of the content expert and the researcher were closely associated, $r=.85$. The strong correlation indicates a real relationship between the two raters' scores. The findings moreover contribute to the depiction of the quality of the CLEAR dynamic simulation assessment, since the essay scores form an important constituent of the CLEAR simulation of the university experience. These results complete the findings answering Question 1.

Research Question 2 asks, What is the validity of inferences based on scores derived from the CLEAR? This question is key, for at the heart of test quality lies the issue of validity. This research question was considered from three perspectives, first that of content-related evidence for validity claims, then criterion-related evidence, and lastly construct-related evidence.

For content-related evidence, the CLEAR dynamic simulation assessment was closely scrutinized by experts for evidence of domain sampling and representativeness.
The test stimuli were subjected to analysis in addition to the test instruments, as deemed fitting in this first systematic investigation of the CLEAR quality. The expert judge approved the textbook and the lecture materials. The knowledge subtest items, the essay prompt, and the essay scoring tool were commended. Also endorsed was the level of cognitive and academic demand in the CLEAR. The expert evaluations were therefore consistent and positive, very solid support for content-related validity claims for the CLEAR. These results help illuminate an important strength of this dynamic simulation assessment.

Examining grounds for validity claims next proceeded to considering criterion-related evidence. Such analysis normally seeks to find support for a new test because of its relationship with an established measure. The case of the CLEAR is different. The testing approach is different: assessing through dynamic simulation, as found in the CLEAR, might not bear a close relationship with measuring crystallized knowledge via static tests. The construct measured is different: AR as measured by the CLEAR may show little in common with tests of English proficiency. Finding shared variance would seem more likely if the tests were similar, or indeed shared similar measurement approaches. Thus, expecting to find disagreement, the research correlated examinees' CLEAR scores with their scores on static ESL tests, the TOEFL and the MTELP. As well, CLEAR scores were correlated with ESL teachers ratings of examinees' likely success, as a concurrent measure. The teacher ratings were expected to bear more in
common with the CLEAR than the static test scores.

In relation to the CLEAR knowledge subtest, results showed that scores on this subtest and on the MTELP placement test did not reveal a strong correlation, \( r(19) = .16 \), as might have been expected. However, between the CLEAR knowledge subtest and the TOEFL was found a strong association, \( r(28) = .58 \). The CLEAR knowledge subtest also correlated with the ESL teacher ratings to moderate level, \( r(36) = .38 \). The association found between the teachers' ratings and the CLEAR was anticipated, as was the lack of relationship between the CLEAR knowledge and the MTELP.

Also found was an association between the CLEAR knowledge and the TOEFL. A close connection was not anticipated, because the two measures are largely different. Nevertheless, finding some shared variance is not unreasonable, for in both measures examinees must display knowledge (whether about language or content) in a selected-response format. Perhaps this format draws upon similar abilities in learning, analysis, or test-wiseness. The overlap between the two tests is not great. Approximately 33% of the variance is shared between the TOEFL and CLEAR knowledge. Thus, most of the CLEAR knowledge cannot be explained by performance on the TOEFL: about 67% of the CLEAR knowledge variance is unique.

The pattern of relationships found for the knowledge subtest differ from the associations found for the essay subtest. CLEAR essay scores did not correlate to a statistically significant degree with the selected-response measures, nor with the teacher data. Unsurprising, perhaps, is the lack of association between the CLEAR essay and
the TOEFL, and the CLEAR essay and the MTELP. Quite likely, composing an essay engages skills that are beyond those needed in a multiple-choice test. However, a relationship was expected between essay scores and teacher ratings. Several possibilities are presented to explore this lack of correlation. Statistical limitations perhaps are the cause. A four-point scale is the extent of the range for both the essay and the teacher rating, which might mask an association by limiting variance. The basis of the teacher ratings might also be questioned. Perhaps the ratings were not based strictly on suggested elements such as student motivation, work ethics, and study habits. Student performance on multiple-choice tests could indeed have influenced the teachers' ratings. Student performance on standardized tests could, similarly, have been known to the teachers and affected how they rated examinees.

The nature of the CLEAR essay suggests a final possibility for understanding the teachers' ratings. Recall that ESL classes, like English-subject classes, typically employ a reading stimulus as a springboard for compositions, to inspire the student or to frame the prompt for the composition (Carson, 2001; Horowitz, 1986a, 1986b; Johns, 1991; Leki, 1995; Leki & Carson, 1997; Spack, 1988; Weigle, 2003; Weigle & Nelson, 2003). In that sort of composition, the reading passage need never be referenced. However, knowledge must be displayed in student writing for the disciplines. Such is the fundamental characteristic distinguishing discipline-based writing from subject-English compositions. Therefore the CLEAR essay requires examinees to demonstrate knowledge in their writing. Perhaps some of the ESL
teachers had never assigned such writing to their students, thus explaining some of the rating discrepancy.

A cautionary point should be raised, that not all essay data were subjected to analysis. Eight essays showed evidence of plagiarism, an infraction of the guidelines, and so were not scored. What might have been the outcome if these essays had been rated and included in the data set? The possible contribution by these essays can be inferred by examining the papers. Providing direct samples is not possible, since this is a live test, but generalizations may be made. The illicit quotations evidencing plagiarism were notable for misunderstanding or lack of judgment; these essays centered upon unimportant scholars or sideline issues. The types of text plagiarized and the examinee writings did not indicate nefarious intent, but demonstrated weak command of English and academic norms. A reasonable extrapolation, therefore, is that these essays would have scored poorly; lacking their scores has truncated the essay score range.

The third and final aspect of validity investigation addressed construct-related evidence for validity claims. This correlation analyzed the level of association between the two CLEAR subtests, finding a strong validity coefficient $r(28) = .50$. That is, this correlation very probably expresses a real relationship. The overlap between the knowledge and essay subtests indicates some variance in common. That not all variance is shared indicates that the two subtests are not fully redundant, but each test component
reveals information not provided by the other. Keeping both subtests seems a reasonable conclusion based on these results.

In order to understand better AR as measured in the CLEAR, an additional correlational test was conducted. Here were examined the relationships of the CLEAR simulation subtests with static tests of English proficiency, the TOEFL and MTELP, with the teacher ratings, and the examinee gender. The web of relationships is offered in graphical representation (in Figure 15) as a possible expression of the correlational findings. Thus, the CLEAR knowledge subtest and essay subtest overlap, though not fully. The CLEAR knowledge subtest and the TOEFL share variance, and both share a portion of teacher ratings. Examinee gender is apparently associated with knowledge scores: although no statistically significant correlation exists, gender is the only factor under examination which can connect MTELP placement test scores with other scores.

Within the web of relationships, the knowledge subtest scores are central. It is with this subtest that all the other measures are directly or indirectly associated. Perhaps this relationship suggests that selected-response tests similarly tap analytical abilities and teachers make a connection between such performance and university achievement. Possibly examinees can successfully employ test-taking strategies in selected-response tests, and teachers rate their likely successful taking into account this test-wiseness. Such ability, however, is of little assistance in constructing a text-responsible essay. Evidently, the essay draws on abilities different from those used in multiple-choice tests. Thus,
despite the knowledge subtest correlations with all other measures, no claim is made that the knowledge is the central element. Rather, this pattern would suggest that the essay is quite valuable, providing information not available from any other source. The inter-relationships should certainly be supported by research replicating the present study for greater confidence in these associations. These findings complete the results from test data, and conclude reporting the results responding to Research Questions 1 and 2.

The next area of findings responds to Research Question 3, "What is the examinee perception of the CLEAR?" Across the various items of the feedback form, examinee remarks may be grouped into seven different categories: language, CLEAR stimuli,
CLEAR subtests, task engagement, time, study, verisimilitude, and CLEAR impact.

Language skill limitations were noted in many comments. Examinees saw the need to be able to read long texts, to take notes while listening to a lecture, to write showing understanding. CLEAR stimuli perceptions changed over time, after study and after taking subtests. In hindsight, the video lecture difficulty appeared greater, yet reading was rated not so difficult after days of study.

The CLEAR subtests were of interest to examinees. They considered the CLEAR overall, the knowledge subtest and especially the essay subtest all difficult. Some examinees remarked that the knowledge subtest was too specific, though others judged it fair. Some examinees perceived that the CLEAR simulation measures learning and understanding. Most suggestions for improving the test were unfeasible: examples include employing a live lecturer for every test and allowing examinees to take test materials home. Practical suggestions included considering simplification of language level in the test; another idea was giving students example questions at the beginning of the simulation, so they could envision taking a content-based test in English.

Task engagement emerged as an important theme. Many examinees found the lecture and text interesting. While a few individuals commented that the topic was out of their discipline or was too easy, far more indicated positive comments about the topic and stimuli. The duration of the CLEAR dynamic assessment allowed interest in the
simulation to develop. The CLEAR was exciting to some examinees because of the demanding challenge and bringing the target environment to life.

Exciting, challenging, and tiring are how many participants viewed the CLEAR, in part because of the time demands. Responses were mixed: more time for study was desirable, but the experience was also tiring. Pacing oneself was noted to be helpful in the CLEAR simulation. Timing was especially a challenge in the essay subtest. Timed essays seemed to be unfamiliar to some CLEAR participants.

The importance of studying and learning was noted by numerous respondents. Academic learning, some saw, is beyond language skill development. Remarks noted the utility of reading for understanding, and of studying before taking a test. Others addressed the need to devote more time and effort to study. After the CLEAR dynamic simulation, many examinees perceived study as more important than before.

Verisimilitude was a key quality of the CLEAR. Examinees said that they experienced the university world through the CLEAR dynamic simulation assessment. Participants stated that the simulation events resembled those in the target environment; they also noted the resemblance in task demands such as the lengthy readings and data-based writing. Some aspects were demanding in their unfamiliarity: the open-book, timed conditions made the essay notably challenging. Succeeding in the CLEAR is not a matter of a few strategies, noted some test-takers. Numerous participants characterized the CLEAR dynamic assessment as "real." A large majority of examinees would recommend
that a friend take the CLEAR to experience this quality.

Many participants experienced better understanding of themselves, of their strengths and weaknesses, after the CLEAR dynamic simulation. Thus, the test impact became the final theme emerging from feedback. Examinees noted that simply taking the CLEAR test helps these ESL students improve language and learning skills. Moreover, measuring themselves against real-world demands was valuable in aiding self-assessment. Many said they knew themselves better after the CLEAR simulation.

Implications for Theory

The CLEAR dynamic simulation assessment was constructed on evidence for the university demands upon students. Recall that a close connection with the target domain is crucial in a criterion-referenced measure, and in a simulation. For a criterion-referenced simulation, the practice-to-theory approach is appropriate.

Additionally, the CLEAR was informed by theory and by observation in building a test to measure AR. Recall that AR is defined as the direct, present evidence of ability to learn academic content via the L2 as demonstrated during the dynamic simulation, learning new subject-area material to a level appropriate for native-speaking students, and having learned under conditions similar to the same. AR indicates that ESL examinees' second-language (L2) proficiency is above the "threshold" level (Adamson, 1993; Alderson, 1984; Clapham, 1996; Clarke, 1980/1988; Cummins, 1979; Elder, Erlam, & von
Randow, 2002; Laufer, 1997; Laufer & Sim, 1985; Lee & Schallert, 1997; Ridgway, 1997; Taillefer, 1996). and therefore sufficiently advanced to allow the examinees to allocate resources to learning new subject matter information transmitted in the L2, which has been argued to be the central activity of university students (Elliot & DiPerna, 2002, p. 10). The construct of AR is hypothesized to include CALP (i.e., above-threshold academic language proficiency) as well as academic skills developed in the L1, that is, reading and writing skills, mathematics and science skills, reasoning skills such as analysis, synthesis, and investigation (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002), the importance of particular skills varying according to the instructional event. Academic auxiliaries also contribute to AR and, in turn, enable academic success. The auxiliaries are posited to include the dimensions of motivation, study skills, engagement, work drive, emotional stability, affective schemata, and metacognitive strategies.

The construct of AR has brought together knowledge from research into academic competence and L2 learners (Adamson, 1993; Bachman, 2002; Bachman & Palmer, 1996; Garner & Borg, 2005; Huong, 2001) and investigations into academic achievement among native-English-speaking university students (Elliot & DiPerna, 2002; Mroch, 2002; Mroch, Lang, Elliott, & DiPerna, 2002; Ridgell & Lounsbury, 2004). Topical knowledge is another individual resource the above-threshold L2 user can deploy, demonstrating AR. No longer "short-circuited" (Clarke, 1980/1988) by inadequate L2
ability, the individual's topical knowledge can benefit learning content in the L2 (Clapham, 1996; Ridgway, 1997).

This construct of AR was assessed through the CLEAR. Despite the need for test revision, the present study may make some contribution to understanding EAP students and AR. The present study seems to suggest support for the construct of AR, particularly the CALP and academic auxiliary dimensions. The CLEAR dynamic simulation appears to draw upon these elements, by the linguistically intensive discipline of psychology, by the stimulus materials, by the essay requirements, and by the duration of the dynamic simulation.

Dynamic simulation offers a unique approach to testing EAP students' learning potential. The use of dynamic simulation may offer rich insight into ESP examinee ability. Of particular interest is the considerable shared variance between the knowledge and essay subtests. This suggests that a more complete depiction of learner ability is provided by the two modes of testing. Particularly interesting is the essay-related evidence suggesting that student abilities are not frequently directed to content-responsible writing. Such is regrettable, for knowledge-based writing will very likely be assigned in their university work (Carson, 2001; Horowitz, 1986a, 1986b; Weigle, 2002, 2003; Weigle & Nelson, 2003). Also intriguing were the indications that, having experience the CLEAR dynamic simulation, examinees improved in self-assessment and motivation. If further investigation supports these early indications, such might contribute
to a more detailed depiction of EAP students, promoting a better understanding of their needs. Such might be of interest to the SLA and language testing communities, who might also find interest in measuring learning potential through dynamic simulation assessment. Conversely, the present research might offer some contribution to the dynamic testing literature through implementing the simulation model of testing AR.

Implications and Recommendations for the CLEAR Test

The CLEAR dynamic simulation assessment has been investigated in this validation research study, illuminating the present strengths and weaknesses of the measure.

Strengths of the CLEAR might begin with global factors, the academic discipline and topic. As was the intention of the test developers, these choices helped make the test effective. Examinees were not familiar with the topic or discipline, so they truly experienced learning in the simulation-based assessment. Examinees also experienced a test of factual learning and understanding, another CLEAR feature. The close connection to the target domain and university-type learning was attested by a subject-matter expert, buttressing content-validity-related claims. These factors work in synergy to support the effectiveness of the dynamic simulation assessment. These outcomes indicate that the investment of time and attention into the CLEAR development was not without effect.
The test developers might advantageously follow much of the early CLEAR procedures when constructing alternate test forms.

The two subtest components function smoothly in the present version of the test although the CLEAR is not without need for improvement. The knowledge subtest, for example, requires revisions of problem items and distractors, and adding new test items. Moreover, the test items need to be mapped to objective clusters (whether such is performed through statistical or logical means) so that new consistency estimates may be calculated upon the mean cluster scores. These data will perhaps better depict the consistency of the CLEAR, and one more appropriate to the multidimensional nature of the knowledge test. With this information, the test owners will be better able to evaluate examinee performance.

Regarding the essay, the subject area expert consultant evinced strong support. The prompt was approved by this judge, although the measurement consultant advised narrowing the prompt to guide responses. The prompt, for example, can direct writers to construct an argument which must include particular points of knowledge. Such would likely improve measuring construct-relevant variance, agreement among raters, and enhance measurement quality. The scoring criteria, their relative importance, and the scoring tool were commended. Using the scoring tool with the content consultant yielded good scoring agreement. The scoring pattern agreement indicated harmony of scoring among the CLEAR team, the consultant, and the researcher. These favorable outcomes
help defend content-related validity claims, and may help guide the test-builders in future development activity.

A concern, however, is that the correlations for the CLEAR essay scores do not necessarily form a clear and convincing pattern at present. That is, support for construct-related validity claims is suggested in the inter-subtest correlations. Although essay subtest scores share some variance with the knowledge subtest scores, the essay subtest scores are not related with teacher ratings. These results were not expected, based on early pilot studies where were seen associations between CLEAR scores and teacher ratings. The present findings may speak to the uneven "patchwork" of data from 13 teachers. Perhaps the findings relate to these teachers' unfamiliarity with the construct of AR. Another possibility is that these ESL teachers ignored the directions and rated students on their linguistic proficiency rather than their general work drive, motivation, study skills, and other AR-related dimensions. Other studies (Leki, 1995; cf. Santos, 1988; Schleppegrell, 2002; Waters, 1996) report occasions wherein ESL teachers' judgment is at variance with that of non-ESL teachers, and the present findings might be another instance of such. An important distinction is that the CLEAR presents a timed, text-responsible essay, markedly different from the personal, narrative writings often assigned by ESL teachers (Horowitz, 1986a, 1986b; Leki, 1995; cf. Santos, 1988; Schleppegrell, 2002; Waters, 1996; cf. Carson, 2001; Johns, 1991; Leki & Carson, 1997; Spack, 1988; Weigle, 2003; Weigle & Nelson, 2003).
Better data collection might result if the teachers were surveyed, given more information about what is meant by AR and "university success". Teacher interviews are another data collection option. Teachers should also be queried about the types of instructional tasks they assign the examinees. Narration and opinion, typical ESL communication tasks, might not be good indicators for the content-responsible writing of the CLEAR essay. Better data collection may help clarify the meaning of teacher ratings, how these are related to CLEAR subtest scores, and to the construct called AR.

Academic ability is measured by the CLEAR, in the view of some examinees. Test quality, others noted, was an outstanding feature of the CLEAR. The CLEAR relevance and utility, examinees reported, were reasons they would recommend a friend take this test. Thus, many of the examinees' perceptions were quite encouraging, particularly since the CLEAR was built for verisimilitude.

The verisimilitude can present initial surprises to examinees: most had learned academic content only via their native languages, and the dynamic simulation concept was novel to all participants. A few test-takers suggested that the CLEAR administrator give a better overview of the test at the outset of the dynamic simulation, to help them better prepare for learning academic content not ESL. This can be accomplished by creating a test user guide with examples from a different discipline. Examinees can be better prepared for the CLEAR by considering how meteorology, for example, is taught differently by an ESL teacher and by a science instructor. The ESL teacher might test
knowledge of meteorology by requiring students know the definition of "barometric pressure" and how to use such in a sentence. The science instructor for native speakers might do the same, but might also ask, "If your barometer reads 24 inches of mercury, is your barometer working or is it broken? What weather condition are you experiencing?"

Presenting such an illustration at the beginning of the CLEAR would perhaps examinees help examinees direct their energies appropriately, and better engage in the micro-world of the simulation.

The verisimilitude comes at a cost. The CLEAR demands resources. Test participants--administrators, examinees, and rater--all invest more effort into the CLEAR than into a static test. Although a commercial standardized test would have spared these costs, the EAPI did not seek convenience in constructing a customized assessment. The Institute sought to develop a measure closely congruent with the university ecology. The validity-related evidence suggests that the verisimilitude has yielded good data; the participant comments demonstrate that the CLEAR is a powerful experience.

Recommendations for Future Research

Bearing a close connection to the target environment, however, means work for test developers and researchers. The present form of the CLEAR exists in one test form, which marks a restriction on the test administration. A test could be given again with little difference if it is a static test. However, the same CLEAR test cannot simply be re-
administered on a short-term basis. Learning must occur in a dynamic simulation. Thus, the test developers will certainly construct new forms.

With those forms, and this first version, further research will be needed to investigate the CLEAR association across academic disciplines, and across university levels of study. Across academic disciplines, discourse communities interact differently. For example, people in engineering frequently employ a problem-solution format. This presents a problem for the CLEAR, for it is not yet evident whether the CLEAR effectively reveals AR for those engineers and for examinees from other disciplines. One possibility is to construct several essay prompts, and require examinees write two or more essays. This would allow inspecting different papers from the same examinee, permitting inter-rater and intra-rater comparisons. Inserting the new task into a live test permits such comparison between an established prompt and a new one, with the examinee held constant. Such would also yield data on the utility of the scoring tool employed with a new prompt. Another recommendation for follow-up research is to contact present research participants, across their various academic fields, in order to explore their present studies and degrees of success.

Another question that cannot yet be answered is how well the CLEAR might reveal AR by university level. The present study was not able to address this question, since the small sample size limited data. Future data collection should ask examinees the level of attained education as well as the intended level of study. The topic is an
important one; results would inform the test owners on the appropriacy of the CLEAR for different examinee populations.

A final question presents a different possibility for investigation, exploring the effect of the CLEAR dynamic simulation on the examinees. Many examinees report, upon experiencing the CLEAR, new level of improved self-assessment, a new level of motivation. How extensive are these changes, and do they persist over time? A longitudinal study tracking examinees after their CLEAR simulation experience could prove informative. The consequences of taking the CLEAR dynamic simulation might potentially be of interest to curriculum designers, instructors, scholars, and test developers.
REFERENCES


229


Interactive Approaches to Second Language Reading (pp 168-182). Cambridge:
Cambridge University Press.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In M. H. Long
Language Vocabulary Acquisition: Cambridge Applied Linguistics Series (pp. 225-


Collier, V. P. (1989). How long? A synthesis of research on academic achievement in
a second language. TESOL Quarterly, 23(3) 509-528.


Cummins, J. (1979). Linguistic interdependence and the educational development of

Cummins, J. (1980). The cross-lingual dimensions of language proficiency:

Implications for bilingual education and the optimal age question. TESOL
Quarterly, 14(1) 175-187.

Cummins, J. (1982). Language proficiency and academic achievement. In J.W. Oller,


English Language Institute of the University of Michigan (n.d.). Michigan Test of English Language Proficiency [MTEL]. Ann Arbor, MI: Author.


UCIEP pre-conference Workshop conducted prior to the annual meeting of NAFSA: Association of International Educators conference, Philadelphia, Pennsylvania.


246


APPENDIX A, Comparison of Site #1 and #2
<table>
<thead>
<tr>
<th>City Comparison</th>
<th>Site#1 City</th>
<th>Site#2 City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>People</td>
<td>46966</td>
<td>270027</td>
</tr>
<tr>
<td>Pop. Density</td>
<td>10043.9</td>
<td>25450</td>
</tr>
<tr>
<td>Median age</td>
<td>35.5</td>
<td>24.08</td>
</tr>
<tr>
<td>Household size</td>
<td>2.55</td>
<td>2.27</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor</td>
<td>12.60%</td>
<td>12.00%</td>
</tr>
<tr>
<td>Professional</td>
<td>14.70%</td>
<td>12.00%</td>
</tr>
<tr>
<td>Technical</td>
<td>16.60%</td>
<td>9.00%</td>
</tr>
<tr>
<td>Sales</td>
<td>12.50%</td>
<td>12.30%</td>
</tr>
<tr>
<td>Clerical</td>
<td>15.50%</td>
<td>10.00%</td>
</tr>
<tr>
<td>Blue-collar</td>
<td>40.70%</td>
<td>40.00%</td>
</tr>
<tr>
<td>Environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school grad.</td>
<td>79.50%</td>
<td>75.00%</td>
</tr>
<tr>
<td>4 yr. college grad.</td>
<td>17.50%</td>
<td>12.50%</td>
</tr>
<tr>
<td>PhD graduates</td>
<td>1.50%</td>
<td>7.40%</td>
</tr>
<tr>
<td>Population ratio</td>
<td>19.5</td>
<td>17.4</td>
</tr>
<tr>
<td>Crime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Violent crime</td>
<td>466.1</td>
<td>225.4</td>
</tr>
<tr>
<td>Property crime</td>
<td>4162.2</td>
<td>1075.1</td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual snow</td>
<td>54.6</td>
<td>49</td>
</tr>
<tr>
<td>Frost days</td>
<td>103</td>
<td>220</td>
</tr>
<tr>
<td>Precip. Days</td>
<td>83</td>
<td>4</td>
</tr>
<tr>
<td>Avg. Annual Snow</td>
<td>26.8</td>
<td>20.1</td>
</tr>
<tr>
<td>Elevation (ft)</td>
<td>750</td>
<td>15</td>
</tr>
</tbody>
</table>

Cost of Living Indexes
- **Overall 100 = national average**
- **Lower = better**
- **Housing comprises 31% of overall cost of living**
- **Food and groceries comprises 16% of overall cost of living**
- **Transportation comprises 10% of overall cost of living**
- **Utilities comprises 8% of overall cost of living**
- **Health comprises 5% of overall cost of living**
- **Miscellaneous comprises 30% of overall cost of living**

<table>
<thead>
<tr>
<th>Housing</th>
<th>Site#1 City</th>
<th>Site#2 City</th>
</tr>
</thead>
<tbody>
<tr>
<td>House purchase cost median home value - Q3 2000</td>
<td>$128,300.00</td>
<td>$94,300.00</td>
</tr>
<tr>
<td>Home appreciation</td>
<td>7.20%</td>
<td>10.30%</td>
</tr>
</tbody>
</table>

**Economy**
- **Unemployment Rate**
- October, 2000: 4.00% 2.70% 2.80%
- **Recent Job Growth**
- Most recent 12 months: 1.30% 3.40% 3.70%

---

251
APPENDIX B, CLEAR Essay Prompt/Guidelines
CLEAR: Essay Assessment

[Institute name]
[University affiliation]

code ____________________________
date ____________________________

Parents certainly influence their children, but not everyone agrees upon the extent of this influence. Indeed, some authors assert that the parental role is primary, while others consider the parents' influence minimal.

Imagine that the lecturer, [name omitted], is your teacher and has assigned you this essay for a midterm. For this professor, write an essay of approximately two pages (typed). You will have 90 minutes (1.5 hours) in which you answer the following question:

[text omitted to maintain security]

Writing Guidelines

Follow this plan for your writing:

- First, choose how you will answer the question.
- Remember to show that you learned content information. An unsupported opinion is not an effective essay.
- Next decide how you will organize your essay. Use the back of this sheet for brainstorming or outlining.
- Make sure that you state a position, and develop it. Show that you learned from what you heard and read.
- Then, where appropriate, include select quotations from the lecture and the reading.
- You must include citations from both the lecture and the reading in your essay. These may support your position or contrast with it. They may be numbered or language, directly or indirectly reproduced.
- Make sure you cite the original source. Report the name and date when you quote any author or speaker. This applies to direct and indirect references.
- Above all, be sure that the quotations or excerpts are woven into your own arguments, a string of quotations does not make an effective essay.

Resources Guidelines

You may use:
- your own dictionary
- your reading reserve folder, including the reading photocopy, and your notes written in your study hall
- a manual or provided by the study skills center
- a computer-aided dictionary or thesaurus
- Microsoft Word spellcheck and thesaurus (select a word and press “Shift+F7”)

You may not use:
- your own dictionary
- an electronic or a bilingual dictionary
- the Internet or email
- any software other than Microsoft Word
- any other sources or materials.

Formatting Guidelines

If your response is typed:
- your own dictionary
- contain your code number, not your name, in the upper right hand corner of every page
- doublespaced
- one-inch margins
- Times New Roman font
- 12-point font size

If your response is handwritten:
- your own dictionary
- contain your code number, not your name, in the upper right hand corner of every page
- doublespaced
- one-inch margins
- legible writing or printing in ink, not pencil

Submission Guidelines

Special instructions for people using computers:
- Stop writing at the warning announcement 15 minutes left.
- Correct the spelling and formatting.
- Save the document to disk.
- For the document name, use your code number.
- Write your code number on the disk.
- Print one copy of the essay.

All test takers:
- Remain silent until you leave the test room.
- Make sure your code number is in the top right corner of every page.
- Wait by the seat for the exam supervisor or assistant to come to you for the essay.
- Turn in the essay printout, the disk, and this essay guideline.
- Do not leave until permitted by the exam supervisor.

253
APPENDIX C, Original CLEAR Essay Rubric
### CLEAR: Rubric for Essay

<table>
<thead>
<tr>
<th>Competency</th>
<th>Skills</th>
<th>Superior</th>
<th>Advanced</th>
<th>Satisfactory</th>
<th>Unsatisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>Task</td>
<td>- Fully addresses the writing task.</td>
<td>Adequately addresses the writing task.</td>
<td>Addresses the writing task, indicating some areas.</td>
<td>Addresses only a few aspects of the writing task.</td>
</tr>
<tr>
<td>Information</td>
<td></td>
<td>- Demonstrates excellent ability in analyzing and integrating source materials.</td>
<td>Demonstrates strong ability in analyzing and integrating source materials.</td>
<td>Demonstrates adequate ability in analyzing and integrating source materials.</td>
<td>Demonstrates limited ability in analyzing and integrating source materials.</td>
</tr>
<tr>
<td>Details</td>
<td></td>
<td>- Develops main ideas with appropriate supporting ideas.</td>
<td>Adequately develops main ideas with appropriate supporting ideas.</td>
<td>Has an adequate intro and conclusion.</td>
<td>Has an ineffective intro and conclusion.</td>
</tr>
</tbody>
</table>

---

**CLEAR Rubric for Essay**

<table>
<thead>
<tr>
<th>Competency</th>
<th>Skills</th>
<th>Superior</th>
<th>Advanced</th>
<th>Satisfactory</th>
<th>Unsatisfactory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Expression</td>
<td>- Shows original sentence variety. Consistently uses expressions that are clear and comprehensible.</td>
<td>Shows adequate sentence variety. Generally uses expressions that are clear and comprehensible.</td>
<td>Shows some sentence variety. Contains some unclear expressions.</td>
<td>Shows little sentence variety. Contains many unclear/comprehensible expressions.</td>
</tr>
<tr>
<td>Conventions</td>
<td>- Consistently conforms to the conventions of standard written English.</td>
<td>Generally conforms to the conventions of standard written English.</td>
<td>Contains some minor deviations from the conventions of standard written English.</td>
<td>Contains major deviations from the conventions of standard written English.</td>
<td></td>
</tr>
<tr>
<td>In-Text Citations</td>
<td>- Consistently uses appropriate attribution to source materials.</td>
<td>Consistently makes appropriate attribution to source materials.</td>
<td>Makes minimal use of attribution to source materials with some mechanical errors.</td>
<td>Makes inappropriate attribution to source materials.</td>
<td></td>
</tr>
</tbody>
</table>

---

255
APPENDIX E, Feedback from Test-Takers
CLEAR: Feedback from Test-Takers  

Rate each part of the exam on a scale of 1 to 4 (1 is easy; 4 is difficult).

1. video lecture______  c. knowledge test______  e. overall______
2. readings______  d. essay test______

2. Comment on any of the sections of the test

________________________________________________________________________

3. From taking this test, I learned that:

________________________________________________________________________

4. I did the best on the __________ section of the test, because

________________________________________________________________________

5. If I could choose between taking the CLEAR and taking the TOEFL for admission to this university, I would choose __________ because

________________________________________________________________________

6. Taking this test over a 48-hour period was

________________________________________________________________________

7. If I could change one thing about the test or the test administration, I would change

________________________________________________________________________

8. The test I took gave me a feel for what I will have to do in the university. ___Yes ___No

Why?

________________________________________________________________________

9. I would recommend taking the CLEAR to a friend. ___Yes ___No

Why?

________________________________________________________________________

10. If you have any other comments about the CLEAR or the test process that you would like to share, please comment below.

________________________________________________________________________
APPENDIX F, Teachers' Report Form
Content Learning Experience: Academic Readiness (CLEAR) Test

Teachers' Report Form

Rank order your students from strongest (first) to weakest (last), and circle the description that best matches your estimate of the student's likelihood of success in university work. Base this NOT merely on English proficiency but on your judgment of student overall abilities.

<table>
<thead>
<tr>
<th>1st name</th>
<th>first name</th>
<th>circle one description per student</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>2nd</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>3rd</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>5th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>6th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>7th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>8th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>9th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>10th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>11th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>12th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>13th</td>
<td></td>
<td>probably succeed</td>
</tr>
<tr>
<td>14th</td>
<td></td>
<td>probably succeed</td>
</tr>
</tbody>
</table>

name:__________________________
APPENDIX G, Feedback from Raters and Administrators
1. Administering this test was _____________________________.

2. The Essay Scoring Tool was _____________________________.

3a. For test takers
Did the test simulate a university requirement?  
____Yes   ____No
Please explain__________________________

3b. For test administrators
Did the test simulate a university requirement?  
____Yes   ____No
Please explain__________________________

4. Students seemed __________________________ when taking the test.

5. Comment on the topic of the CLEAR.

6. Comment on timing of the test and test components.

7. Remark on the CLEAR in terms of quality, usefulness, and resemblance to the real world.

8. Suggestions for change:

9. Other remarks: ________________________________
About the Author

Sha Balizet taught ESL/EFL in four continents before earning her M.A. in Applied Linguistics and Ph.D. in Second Language Acquisition and Instructional Technology at the University of South Florida.

As a doctoral student, Miss Balizet taught graduate and undergraduate courses in linguistics and research methods. She developed innovations in technology coaching for ESL teachers and online instruction in educational linguistics, and also managed advising duties for the Linguistics graduate program.

Her test construction experiences span the development cycle, and her research work includes investigations into the principled use of audio in testing, individual differences in language achievement, developing a spiral model of distance education instruction, and a comparability study of computerized and traditional modes listening tests. Sha Balizet has delivered invited workshops and conference presentations. She has published, individually and jointly, in peer-reviewed journals. Sha Balizet is the Research Director of English Language Learner testing for Harcourt Assessment.