DOCUMENT RESUME

ED 423 675 FL 025 476

AUTHOR Cardoso, Rosana M. F.
TITLE Authentic Foreign Language Testing in a Brazilian University Entrance Exam.
PUB DATE 1998-00-00
NOTE 21p.; For complete volume of working papers, see FL 025 473.
PUB TYPE Journal Articles (080) -- Reports - Research (143) -- Tests/Questionnaires (160)
JOURNAL CIT Texas Papers in Foreign Language Education; v3 n2 p51-70 Spr 1998
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *English (Second Language); Foreign Countries; Higher Education; *Language Proficiency; *Language Tests; Psychometrics; *Reading Comprehension; Test Construction; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Brazil

ABSTRACT

This study analyzed English language tests administered in Brazil as part of a university entrance examination, focusing on the authenticity of its tests of second language reading comprehension, the concept of reading as an interactive process between reader and text, a proficiency-based view of language instruction, and the psychometric properties of a good test. Two reading tests that explicitly favor authenticity and proficiency are analyzed in two sections. The first concentrates on the types of questions in both tests and the language skill levels that are tapped by each question, according to American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines. The second part is a reliability study, in which the consistency of scores is statistically analyzed to determine the overall quality of the tests. The reliability test is also performed to verify the hypothesis that a shorter test is less reliable than a longer one. (Contains 25 references.) (Author/MSE)

Authentic Foreign Language Testing in a Brazilian University Entrance Exam

ROSANA M. F. CARDOSO

This study analyzes English tests administered in Brazil as part of a University Entrance Exam. It attempts to encompass the considerations about authentic tests of L2 Reading Comprehension, the concept of reading as an interactive process between the reader and the text, a proficiency-based view of language instruction and the psychometric properties of a good test. Two reading tests that explicitly favor authenticity and proficiency are analyzed in two sections. The first one concentrates on the types of questions in both tests and the language skill levels that are tapped by each question according to the ACTFL Proficiency Guidelines. The second part is a reliability study, where the consistency of scores is statistically analyzed in order to determine the overall quality of the tests. The reliability test is also performed to verify the hypothesis that a shorter test is less reliable than a longer one.

INTRODUCTION

Within the last two decades, research on language assessment has increasingly advocated the use of authentic materials (Shohamy & Reves, 1985; Stevenson, 1985; Wiggins, 1994; Young 1993). According to Clark (1975), direct tests (nowadays referred to as "authentic tests") are an attempt to duplicate as closely as possible the circumstances and efficacy of real-life situations. A growing commitment to a proficiency-based view of language learning and instruction brings the necessity of authenticity in all areas of language assessment. Accordingly, tests of L2 Reading Comprehension have been following the same trend, presenting real-life situations in which language proficiency is ordinarily demonstrated (Davis, 1994; Valette, 1994).

Refuting the traditional "indirect tests," which did not tap real-life language but had the advantage of being easily analyzed psychometrically (Shohamy & Reves, 1985), authentic assessment directly examines student performance on intellectual tasks (Wiggins, 1994). The use of authentic materials in reading tests, defined as texts written and read by native speakers in ordinary real-life situations, has been advocated by ESL and FL researchers as a replacement for edited texts (Byrnes, 1987; Swaffar, 1981; Young, 1993). Even though such texts were presumed to ease the students' reading processes, studies reveal that the linguistic simplicity of edited texts can make reading more difficult (Young, 1993).

Reading is considered here an active process where the readers do not merely decode what the text encodes, but they construct text meaning by synthesizing their prior knowledge, which may be linguistic, cognitive or experiential, with textual data. Reading cannot be seen anymore as simple "passive process of extracting meaning from the printed page, but rather as an active and interactive

process in which the reader uses knowledge of the language to predict and create meaning based on the text" (McLeod & McLaughlin, 1986). In all language tests, one of the main concerns should be the relationship and interference among skills. A reading test must tap only the students' reading ability, not writing or listening comprehension, for instance. Donin and Silva (1993) show that, at least at intermediate levels of L2 proficiency, the use of L2 production (writing) tends to underestimate and distort L2 reading comprehension. They also found that the lack of inferencing that has been attributed to L2 comprehension may be a result of the assessment techniques, for example whether the students have to produce responses in L1 or L2. Therefore, this is an issue that deserves great consideration on the part of researchers and test developers.

Another constant preoccupation is the quality of the test, from classroom tests to standardized tests. According to Friedenberg (1995), a good test should be carefully designed and empirically evaluated to ensure that it generates accurate and useful information. The design phase should contain clear definitions about the test's purpose, cover a specific content or domain, and define a set of administration and scoring procedures. The evaluation phase should include collection and analysis of data, which are then used to identify the psychometric properties of the test. Those properties, which are the measurement characteristics of a test, are determined by analyzing responses to test items. A good test is reliable (it provides a consistent measure of current knowledge, skills or characteristics), valid (it indicates whether the test measures what it was designed to measure), and comprises items with good item statistics (the pattern of responses to individual test items is analyzed with the purpose of identifying items in need of revision).

As an attempt to encompass the considerations about authentic tests of L2 Reading Comprehension, the concept of reading as an interactive process between the reader and the text, and the psychometric properties of a good test, this study analyzes an English test administered in Brazil as part of a University Entrance Exam. The State University of Campinas (Unicamp) completely restructured its entrance exam in 1987 and moved away from traditional multiple-choice items towards proficiency-oriented essay questions. The changes affected not only the FL exam, but all subjects. The specific change of the FL Test (Bastos et al., 1993) focuses on measuring the students' reading competence instead of measuring their explicit knowledge of grammar or their writing skills. Aiming for language proficiency instead of knowledge "about" the language, Unicamp's FL Test evaluates explicitly and exclusively the students' reading competence.

A recent change in the exam format (the length of the test was changed from 16 essay questions to 12) raises several issues concerning the overall quality of the test and its psychometric properties. The last 16-question test and the first 12-question test are analyzed in two sections. The first one concentrates on the types of questions in both tests and the language skill levels that are tapped by each question according to the Ameri-

can Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for Reading. The second part is a reliability study, where the consistency of scores is statistically analyzed in order to determine the overall quality of the tests. The reliability test is also performed to verify the hypothesis that a shorter test is less reliable than a longer one.

SIGNIFICANCE OF THE STUDY

São Paulo State's Unicamp is one of the most distinguished Universities in Brazil. In 1987, "Vestibular Unicamp" (The State University of Campinas' Entrance Exam) was dramatically modified to better fit the new profile of the students the University was hoping to attract. That was a pioneer initiative that would later influence not only many other Brazilian universities to take that same path, but also secondary schools to reshape their curriculum—the "washback effect" discussed by Alderson & Wall (1993) and Peirce (1992).

Unicamp wanted to select applicants who had better analytical abilities. It was the University's belief that students would better reveal such capacity if they had a chance to show their line of reasoning while answering the entrance exams' questions. This would not be possible if Unicamp kept employing the traditional multiple choice tests used until then to select applicants. The solution then was to abandon the traditional multiple choice tests in favor of a more complex and thorough form of evaluation using essay questions on tests of all subject matters of Secondary Level (high school) education.

The objective was then to select students who were able to express themselves clearly, organize ideas, establish relationships, interpret facts and data, and develop explanatory hypotheses in all areas of knowledge. That objective is explicitly declared in the "Manual do Candidato," the Applicant's Manual brochure (Unicamp 1994) published every year by the Special Committee for Entrance Exams (Comissão Permanente para os Vestibulares da Unicamp) to better guide students who will be taking its examinations. This publication also describes the topics of all subject matters that the test will cover.

By giving applicants the opportunity to better demonstrate their knowledge and thus prove their reasoning capabilities, Unicamp makes it clear what it does not expect from its future students: the mere reproduction of information as the result of a passive relationship with knowledge, where they do not need a point of view about what they have learned. Instead, Unicamp wants students who can show their active relationship with knowledge through their writing.

The new entrance exam has two phases (Unicamp, 1994). The first one, mandatory for all applicants, is a four-hour test divided into two parts: one Essay and twelve General Questions (based on the content described in the nationwide Official Secondary Level School Syllabus). The second phase, taken by those applicants who scored in the fiftieth percentile on the first phase, is administered in four consecutive days. Each day is composed of one four-hour test on two

subject matters. Each subject matter test has a total of sixteen questions.

The FL test is administered with the Mathematics test on the last of the four consecutive testing days. The applicants are allowed to choose between English and French. Even though there is a choice between the two languages, both English and French exams are developed and scored according to the same general principles.

The fundamental change consisted of abandoning the traditional way of evaluating a FL in University Entrance Exams in general in Brazil, where the explicit knowledge of grammar was measured, as well as the students' writing skills in the FL. Considering that reading ability in a FL is indispensable in any academic field, Unicamp decided to evaluate the students' reading competence. The new assessment procedure in this area therefore became a test of Reading Comprehension only.

The new test is consistent with the trend of authentic FL tests, aiming for language proficiency instead of knowledge "about" the language (Omaggio-Hadley, 1993). This tendency appeared in the 1970s in response to the artificial circumstances of traditional "indirect" language tests (Shohamy & Reeves, 1985). It finds its place in FL tests in Brazil more than a decade later, and Unicamp was still for many years the only Brazilian university to develop such tests.

The new test utilizes authentic materials, such as newspaper articles, comic strips, ads, and fiction, to name a few. The need of authentic text use in FL reading tests is defended by Shohamy & Reves (1985) and Stevenson (1985) and reinforced by Young (1993). Being an attempt to reproduce circumstances of real-life situations, authentic tests suit well the frame of proficiency-oriented tests.

The students have access to detailed instructions printed in the front cover of their question brochure. The instructions refer to, among other things, the point value of each question and the duration of the test. A total of sixteen questions are asked in Portuguese, and the students are explicitly and clearly instructed to answer them also in Portuguese, the students' native language. Language of assessment in this case does not affect students' ability to demonstrate comprehension, as Wolf (1993a, 1993b) and Donin & Silva (1993) postulate.

Another change in the exam format took place in the 1995 exam. The reasons for the alteration were not made public by the University. The first phase remained unchanged. The second phase tests, however, had their size reduced from 16 essay questions to 12 (Unicamp, 1995). The students still had four hours of total test time and the content of the test was still the same. The applicants were made aware of the change approximately one semester prior to the test date when the Applicant's Manual was published (Unicamp 1995).

As stated earlier, such reduction raises issues concerning the overall quality of the tests, as well as their psychometric properties. Has the shorter test maintained the standards of the previous eight 16-question tests (from 1987 to 1994)? The last 16-question test (administered in 1994, representing all the previous tests) and the first 12-question test (1995) are examined. The analysis concentrates

on the items of both tests and the language skill levels that are tapped by each question according to the ACTFL Reading Proficiency Guidelines. These guidelines are considered an attempt to define and describe levels of functional competence on a FL (Omaggio-Hadley, 1993), and are taken here as the most appropriate tool to gauge FL proficiency.

Furthermore, a reliability study will ultimately determine the quality and effectiveness of the tests. When a test is shortened, its reliability mathematically decreases (Friedenberg, 1995; Anastasi, 1982). Is the shorter test really less reliable than the longer test? If so, how significant is the difference? The effects of the modification of test length on the reliability coefficient are analyzed, as well as the amount of error that can be found in every set of test scores.

ANALYSIS OF TEST ITEMS

The English tests contain a series of authentic texts, taken from various sources, such as newspapers, fiction, poetry, and advertisements. The general instructions and the questions are in Portuguese. The candidates are explicitly instructed to answer all questions in Portuguese. Even though there is no mention regarding the length of answers in the test's general instructions, answers are limited only by the space allotted to each one in the answer brochure.

On both tests, each question is scored on a scale from 0 to 5. However, the total point value of each test is determined by the number of questions: the maximum total score for the 16-question test (1994) is 80 points, and the 1995 exam with 12 questions is worth a total of 60 points.

The test construction and the grading system are of great importance. However, they will not be addressed here. The target of the present study is the test itself and its psychometric properties, not the test construction and grading system.

The translations of the questions into English are appended. Appendix A contains the translation of the 1994 test, and the translation of the 1995 test can be found in Appendix B. The complete original tests in Portuguese are available from the author upon request.

One of the merits of the tests is the fact that both the questions and the answers are in the applicants' native language. The students are not asked to write in the L2, which guarantees there is no skill interference. In fact, it is possible to state that none of the other language skills are required for students to complete the tasks except for their reading competence.

Both tests present a series of authentic texts about which the questions are asked. The 1994 test has two texts about popular science, one excerpt of a short story, one introduction to a novel and one comic strip. The 1995 test contains one small poem, two popular science texts, an excerpt of a novel, one summary of a book, and one advertisement. The 1994 test has fewer, longer texts. The 1995 test has one more text than the previous exam, but they are shorter.

The exams in general seem to be geared towards students' abilities found in ACTFL's Advanced to Superior language skill levels as defined by

the Reading Proficiency Guidelines (ACTFL 1986). The descriptions of those levels given by the guidelines are consistent with the type of student profile the University wants to attract. However, a more detailed inspection of the test items is necessary to verify such a claim.

Question 17 of the 1994 exam ("What is the way found by Calvin to get ten dollars from his father? Explain.") may not be as easy as it seems. Based on a comic strip, the question asks for the explanation of one character's intentions. The drawing is certainly a good support, but it does not give the students all the indications needed to get to a correct answer. In fact, the question itself could be considered a better clue. The students would also have to grasp the humor in the strip and make the character's intentions explicit. A simple translation would not arrive at the correct answer, for it neglects the ironic tone of the episode narrated on strip. The students would have to go beyond the words and interact with the text to get the "correct" interpretation.

Students' sample answers are cited by Victor & Senatore (in press). Example 1 illustrates one possible correct, yet simple answer (sample answers were translated from Portuguese by the author):

(1) "Calvin makes up a story of aliens who ask for ten dollars and he volunteers to take the money to them."

Sample answer 2 shows how one student addressed the humor (the essence of the story) by adding the word "cleverly," without which the answer would be an inadequate translation:

(2) "Calvin says that aliens landed and that they want ten dollars. His father said that he will give them the money, and Calvin cleverly says that his father is busy and volunteers to do him the favor of delivering [the money]."

Another sample answer shows an attempt at translation that considers the father's activity as a hypothesis, not a fact. That interpretation jeopardizes the logic of the story:

(3) "Calvin tells his father that if he is busy washing the car Calvin himself can take the money to the ETs."

Still on the 1994 exam, questions 18 and 19 seem to be simpler. The two questions are related, and their answers can be found in the same paragraph of the text. Question 18 ("Who is 'Grandfather'?") asks for a simple identification of a character. On the other hand, Question 19 ("'No one in Vietnam has a clock as tall as a man.' (quote in English) How does the narrator justify this statement?") is more difficult because it relies on a comparative, as well as on the reference for the pronoun "that" which appears in the next sentence in the text. Based on the same text, Question 20 ("Why does the narrator refer to things from his country to describe Mr. Cohen?") asks for the reason why the narrator uses comparisons to describe a character. This seems to be a more complicated task because getting to an acceptable answer implies understanding the metaphors.

The third text of the 1994 exam is the longest and it has more questions (21 to 26) based on it then on any other text in that exam. The length of the text required more attention on the students' part. The topic of the text

(a new material possibly harder than diamond) may even have overwhelmed them. If the students paid close attention to the content of the text, as pointed out by Alves et al. (in press), they would be able to recognize already known information (for instance, the Mohs scale, which is part of the High School Chemistry curriculum). That way they would be able to resort to a more comfortable and maybe more appropriate bottom-up processing to approach the text. The students could be misled into using inadequate translations as answers, as sample answers 4 and 5 to Question 21 ("In Dr. Lieber's opinion, what makes his work out of the ordinary?") cited by Alves et al. (in press) indicate:

(4) "Generally the experiment precedes the theory; in this case the contrary happens, that is, the experiment falls into the theory."

(5) "In the majority of the cases the theory comes after the practice and now it falls into practice."

In both answers, the students recognized the similarities (phonological and spelling) between "follows" and "falls" and performed an inadequate translation. This is not sufficient as an acceptable answer. Instead of attempting simple translations, the students should be interacting with the text at a deeper level.

The introduction to a novel is the basis for the next set of four questions (27 to 30). The text basically presents descriptions and narration. All four questions are about facts that are present in the text or opinions expressed by the author. The text is very clearly structured and the questions can be easily located by the students, which seems to make this set of questions the least problematic of the test. Unfortunately, no sample answers were available. However, Questions 29 and 30 were the third and fourth highest mean scores (respectively, 3.140 and 3.079 points) observed in the entire test, which seems to be meaningful.

The last two questions (31 and 32) were based on the text titled "Astronomy." Question 31 ("Based on your reading of the text, explain its title.") requires an explanation for the title. To be able to do that, the students needed to show their ability of synthesis. Question 32 ("Attribute a meaning to: (quotes in English) a) 'meander' [paragraph 1, line 7]; b) 'hurling' [paragraph 4, line 9]; c) 'dim' [paragraph 5, line 3]") asked students to infer the meaning of three words taken from the text. The choice of words for the inferences seems to reveal an intention of making the students also summarize the interpretation of the text as a whole. Any interpretation the readers might assign to the text would demand that they attribute some form of meaning to these words.

The first question of the 1995 exam (Question 13) asked the students to explain what the two lines of the extremely short text said ("The following segment was taken from "A Martian Sends a Postcard Home," by Craig Raine. Read it and answer: what is the similarity between the rain and the television, according to the Martian? Rain is when the earth is television. It has the property of making colours darker."). Students had few words to take into consideration in order to get to the answer. They could

also rely on the title of the text where the two lines were taken from (given in English as part of the question) to help them recover a great deal of information in order to achieve an acceptable answer. Sample answers of the 1995 exam were not made available for consideration.

The text for Questions 14 and 15 is a passage that presents certain discoveries in the field of science. Question 14 ("Based on the images obtained by the spacecraft 'Clementine,' what were the discoveries related to: a) moon's valleys and peaks; b) 'The South Pole Aitken Basin'?") is clearly structured and divided into two parts. It is organized to presumably guide the students' reading. This can probably be considered the reason why this was the question with the second highest observed mean score in the whole exam, a mean of 4.004 points. Question 15 ("How useful is it to calculate the depth of the moon's craters?") asks about the utilization of the discoveries referred to in the text. Probably the most complicated task of this question was the recovery of references, for example, the reference of "such craters" in the last paragraph and "them" in "collisions that created them."

Still on the general topic of science, the third text is a technical report that discusses certain modern technologies. Question 16 ("What is needed in order to be possible to use biofuels for cooking or lighting houses?") is about the interpretation of a grammatical structure. Question 17 ("Cite two advantages of those fuels.") asked for two advantages of the use of biofuels, while the text presents more than only two. This question presented the highest observed mean score of the 1995 exam: 4.132 points Question 18 ("Explain the process of obtaining biofuels.") required the students to describe a process, where the challenge was to separate the description itself from the example.

Questions 19 to 21 are about a set of texts, namely, the introduction of a book, an excerpt from one of its chapters, and the epigraph. The first two questions seem to be among the easiest questions in the whole exam, for the answers could be taken from a direct (and relatively simple) examination of the texts. Indeed, the mean scores observed for Questions 19 ("What are the stories of T. W. Burgess about?") and 20 ("What does having a big mouth bring to Grandfather Frog?") were 3.850 and 3.031, respectively the third and the fifth highest mean scores. However, Question 21 ("What made Grandfather Frog think about the word if?") is slightly more complicated. The answer was directly related to the previous question. The title of the chapter and the epigraph in particular can aid the students in finding the answer, as it could be retrieved in the first sentence of the chapter.

A book summary about the anti-drug policy of the Dutch government is the basis for questions 22 and 23. Question 22 ("What induced the Dutch government to adopt its current position related to the fight against drugs?") is rather direct, asking for the identification of the Dutch government's position on drugs. However, Question 23 ("How do the ideas presented in the book conflict with other international institutions' policy of war on drugs?") demands a more complex type of reasoning as it asks about a justification of the disa-

greement between the ideas about fight against drugs presented in the book and other international institutions' policies. In fact, Question 23 shows the lowest observed mean score in the whole test, only 0.935 points. Moreover, it is the lowest observed mean score of the two tests analyzed here.

Finally, Question 24 ("Explain the title of the advertisement.") is based on an advertisement. The students are supposed to explain the title of the advertisement based on its short text. The two pictures and the logo that accompany the text can also be very helpful. Even though this is a short and apparently easy question, the students must be able to derive meanings from extralinguistic knowledge, combined with their own knowledge of the language. The difficulty of this question is substantiated by the students' performance. Question 24 had a mean score of 1.782 points, the third lowest score on the test.

TEST ITEMS AND THE ACTFL GUIDELINES

In order to further assess Unicamp's reading tests, they were compared to ACTFL Proficiency Guidelines. After a careful review of the texts on which the questions are based and the test items themselves, it is possible to verify that test questions can be basically placed in three of ACTFL's language skill levels as defined by the Reading Proficiency Guidelines, namely, Advanced, Advanced-Plus and Superior levels. It is imperative to note that the guidelines for reading proficiency assume that all reading texts are authentic and legible (ACTFL, 1986).

The Advanced level characterizes readers who are able to read rather lengthy texts of several paragraphs that are presented with a clearly defined underlying structure. Texts at this level include basically descriptions and narrations, such as short stories, news, bibliographies, correspondence, among others. Comprehension is determined by increasing control of the language, as well as situations and subject matter knowledge.

The texts of the Unicamp tests examined here that have the characteristics of those described for the Advanced level are the texts for questions 18 to 20 and 27 to 30 of the 1994 test, as well as the texts for questions 13 to 15 and 19 to 21 of the 1995 exam. Even though question 13 can be considered as geared towards Advanced level readers, some more attentive Intermediate-High level readers may be able to comprehend the text. While Intermediate-High level texts do not significantly differ from those at the Advanced level, comprehension is less consistent and several readings may be required to achieve comprehension. From a total of 28 questions, 13 (almost half of the test items) can be classified as belonging to the Advanced level of reading proficiency.

Readers placed at the Advanced-Plus level are able to understand parts of texts that are conceptually abstract and linguistically complex, as well as texts that deal with unfamiliar themes and those that involve certain traits of the target culture. Readers at this level can also make pertinent inferences and are able to comprehend a variety of texts, including literary pieces, even though misunderstandings may occur. They

may be even able to follow written discourse at the Superior level in areas of special interest or knowledge with a certain degree of difficulty.

Text for question 17 of the 1994 exam, as well as the one for questions 31 and 32, can be considered at the Advanced-Plus level. In the 1995 test, the texts for questions 16 to 18, 22 and 23 can be regarded as belonging to this level of reading proficiency. A total of 8 questions out of 28 can be placed at this level. The above mentioned texts show some degree of linguistic complexity and/or call for inferences on the readers' part, which are basic characteristics of the Advanced-Plus proficiency level.

Readers at the Superior level are capable of attaining almost full comprehension of texts about unfamiliar subjects. Reading ability is not dependent on subject matter knowledge. Texts at Superior level feature academic and professional texts, characterized by hypotheses, argumentation, and supported opinions. At this level, readers are able to match meanings derived from extralinguistic knowledge with meanings derived from knowledge of the language. Text types at this level include basically a variety of literary texts, editorials, reports, and technical material in professional fields.

The texts found to pertain to the Superior level are the ones for questions 21 to 26 in the 1994 test and the one for question 24 in the 1995 exam. A total of 7 questions (out of 28) are related to texts that require reading strategies and involve some degree of argumentation and hypothesizing.

The hypothesis that the tests are geared towards students whose abilities match ACTFL's Advanced to Superior levels is corroborated by examination of both the test items and the descriptions of the proficiency guidelines for each level. It does not mean that the tests were specially constructed for readers at those levels, nor that students whose abilities are below the Advanced level do not achieve some degree of success on those tests. It means that the type of student profile Unicamp wants to attract is consistent with Advanced, Advanced-Plus and Superior levels established by ACTFL.

METHODS: RELIABILITY STUDY

The scores from Unicamp's 1994 and 1995 English exams were obtained directly from the University's Special Committee for Entrance Exams. The data collection took place in the State University of Campinas in December of 1996 and January of 1997. The data consisted of individual item scores and total test scores of all students taking the English tests, which amounts to 16,813 students in 1994 and 11,378 in 1995. The test takers were not identified in any way; neither were their majors, areas of interest or FL skills.

Scores of each test item of both exams, as well as the total test scores of all students taking the tests, were analyzed. The statistical analysis consisted of a reliability study, which is mathematically defined as the ratio of the true score variance (actual differences in test takers' knowledge) to the observed score variance (actual score earned by each test taker). In a reliable test, a greater proportion of the actual observed test score variance can be attributed to differences in true score, not to some kind of random error,

such as anxiety, illness, or poor testing environment.

The internal consistency or homogeneity approach was utilized, which is considered the most appropriate approach to estimating the reliability of tests that are scored with a varying number of points, such as essay questions. If a test is internally consistent, all test items are tapping the same area of knowledge. So, the items can be also thought of as homogeneous, i. e., taken from the same domain.

RESULTS

Mean, standard deviation, item variance, Cronbach's coefficient alpha, and standard error of measurement were calculated for both 1994 and 1995 Unicamp English exams. The results of the 1994 test are shown in Table 1, while Table 2 shows the 1995 results.

On both tests, each question is scored on a scale from 0 to 5. Therefore, the maximum total score for 1994 is 80 points, and for 1995 is 60 points.

According to the coefficient alpha, applicants who took the 1994 exam performed consistently across the different test items ($r = 0.912$, $df = 16{,}811$, $p<<0.01$). Therefore, the test shows good internal consistency. About 91.2% of the test score variance reflects true score differences, and the remaining 8.8% reflects random measurement error.

The 1995 coefficient alpha shows that people who took this test also performed consistently across the different test items ($r = 0.83$, $df = 11{,}376$, $p<<0.01$). Around 83% of the test score variance reflects true score differences. The remaining 17% shows random measurement error.

The 1994 exam has a SEM of 6.069. That means that 6.069 is the average number of test score points that can be attributed to random error in the 1994 exam. The SEM found for the 1995 exam is 5.454. For that exam, the average true score/test score difference is estimated as 5.454 points.

Table 3 presents the 95% and 99% confidence intervals for the mean scores of each test. In each case, the interval represents the range within which the true mean score is expected to fall 95% and 99% of the time considering the SEM (and therefore the reliability) of each test. In the 1994 exam, the true mean score is likely to fall between scores of 24.479 and 48.269 for 95% of the time. In other words, the likelihood of the true mean score falling outside the above mentioned scores is only 5%. The 99% confidence interval (between 20.716 and 52.032) is the range within which it is 99% certain that the true mean score is probable to fall. The 1995 exam has a 95% confidence interval between scores of 19.789 and 41.167, and a 99% confidence interval between 16.407 and 44.549.

The standard error of measurement difference is utilized to evaluate the differences between the mean scores of the two tests. Table 4 shows the SEM_{dif} and the confidence interval cutoffs. The mean score difference between 1994 and 1995 test is 5.896 points. Such difference is less than the cutoffs; therefore, it is not a statistically significant difference.

DISCUSSION

In this reliability study, the consistency of scores was statistically analyzed to determine the overall quality of the test. A comparison of

Table 1
Results of the Statistical Analysis Performed on the 1994 Exam

Questions (Total Number k = 16)	Number of answers scored	Item Variance (σ_i^2)	Standard Deviation (σ_i)	Mean ($\overline{X}$)
1	16,512	1.859	1.363	3.433
2	15,670	5.622	2.371	3.059
3	15,154	5.492	2.343	2.884
4	14,274	5.118	2.262	1.948
5	15,468	3.821	1.954	2.263
6	15,822	4.639	2.645	2.664
7	14,527	3.534	1.880	3.376
8	14,725	2.768	1.663	1.390
9	15,825	3.419	1.849	2.448
10	14,716	5.432	2.330	2.634
11	13,323	1.800	1.341	1.153
12	14,816	5.258	2.293	2.255
13	14,074	2.799	1.673	3.140
14	14,380	3.720	1.928	3.079
15	15,598	2.637	1.624	2.229
16	13,877	2.895	1.701	2.822
Total Scores	16,813(a)	420.496(b)	20.506(c)	36.374(d)
$\sum(\sigma_i^2)$		60.821		
$\sum(\sigma_i)$			31.227	
Coefficient α	0.912			
SEM	6.069			

(a) This total represents the total number of applicants (N) taking the test.
(b) This total represents the σ_i^2 for all test questions together rather than singly.
(c) This total represents the σ_i for all test questions together rather than singly.
(d) This total represents the $(\overline{X})$ for all test questions together rather than singly.

Table 2
Results of the Statistical Analysis Performed on the 1995 Exam.

Questions (Total Number k = 12)	Number of answers scored	Item Variance (σ_i^2)	Standard Deviation (σ_i)	Mean ($\overline{X}$)
1	10,513	4.618	2.148	2.908
2	11,288	1.753	1.324	4.004
3	11,205	2.375	1.541	2.426
4	10,720	4.896	2.212	1.444
5	11,234	2.782	1.668	3.692
6	11,101	2.624	1.620	4.132
7	11,133	2.799	1.673	3.850
8	10,120	4.418	2.101	3.031
9	9,850	3.567	1.888	1.882
10	10,428	5.289	2.299	2.250
11	9,530	2.295	1.515	0.935
12	9,254	4.526	2.127	1.782
Total Scores	11,378(a)	176.050(b)	13.268(c)	30.478(d)
$\Sigma(\sigma_i^2)$		41.947		
$\Sigma(\sigma_i)$			22.121	
Coefficient α	0.830			
SEM	5.454			

(a) This total represents the total number of applicants (N) taking the test.
(b) This total represents the σ_i^2 for all test questions together rather than singly.
(c) This total represents the σ_i for all test questions together rather than singly.
(d) This total represents the $(\overline{X})$ for all test questions together rather than singly.

Table 3
95% and 99% Confidence Intervals of 1994 and 1995 Exams.

Exam	SEM	$\bar{X}$	95% CI		99% CI	
			Lower Limit	Upper Limit	Lower Limit	Upper Limit
1994	6.069	36.374	24.479	48.269	20.716	16.407
1995	5.454	30.478	19.789	41.167	52.032	44.549

Table 4
Standard Error of Measurement Difference and Its 95% and 99% Confidence Intervals.

SEM_{dif}	8.160
Cutoff for 95% CI	±15.993
Cutoff for 99% CI	±21.052

the psychometric properties of the two tests showed that they were similar.

Contrasting the two alpha coefficients, it is possible to see that even though the 12-question test shows a smaller reliability when compared to the 16-question test (0.83 and 0.912, respectively), it still has a considerably high reliability coefficient.

The 1994 reliability of 0.912 indicates that 91.2% of the test score variance reflects true score differences. Only about 8.8% of the test score variance is due to chance factors, not to differences in the actual knowledge being measured. This is a high level of internal consistency, which means that the students performed consistently across the 16 different test items. With this high internal consistency level, all test items are thought of as homogeneous; that is, they were all drawn from the same domain.

The reliability of 0.83 of the 1995 exam is slightly smaller than that of the 1994 exam. It shows that only about 17% of the test score variance reflects random measurement error, which can still be considered a small percentage of variance occurring due to chance factors.

The fact that the reliability coefficients obtained for both tests are considerably high is of great importance. The smaller reliability coefficient found for the 1995 exam does not seem to imply that the shorter test is substantially less reliable. The 1995 reliability being smaller than the 1994 coefficient only substantiates that the longer test has a higher reliability, as it should have. In order to confirm that, a subsequent analysis was performed so that the findings could be corroborated.

The next stage, then, was to verify what the SEM of both tests meant. The SEM found for the 1994 exam is 6.069, whereas its 1995 counterpart has a value of 5.454 points. Such values signify the average amount of points by which true scores differ from observed test scores. In

other words, they are the average amount of error in each test score. Again, it is important to mention that SEM does not refer to the specific amount of error in any one test-taker's test score, it is only an average.

With a mean score of 36.374, the 1994 exam has little over 6 of its points attributable to random error. The 1995 exam has over 5 points out of a mean of 30.478 credited to random error. Both SEMs are approximately one-sixth of their respective exam's mean. That can be considered a rather stable characteristic of the tests.

Considering the large sample size of both exams ($N_{1994} = 16{,}813$ and $N_{1995} = 11{,}378$), it is possible to say that this study produces potentially stable reliability estimates of population reliability. The reliability depends more on the subjects' variability than on the number of subjects. However, the sample size enters into consideration when calculating reliability. Reliabilities obtained on large samples (such as those in this study) are considered more consistent estimates and more closely approximate the population parameters. Moreover, the estimation of confidence intervals of reliability estimates from large sample size studies provides a range of values with a specific probability of including the real reliability (Morrow, 1993).

The 95% confidence interval of the 1994 exam falls between the scores of 24.479 and 48.269 (a range of 23.79 points in a test with a possible total score of 80 points). The 99% confidence interval of the same exam occurs between 20.716 and 52.032 (a range of 31.316 points). The 1995 exam, with a possible total score of 60 points, has a 95% confidence interval between the scores of 19.789 and 41.167 (a range of 21.378 points) and a 99% confidence interval between 16.407 and 44.549 (a range of 28.142 points).

Taking into consideration the total point value of each exam, it is possible to see that the 1994 exam presents narrower confidence interval ranges when compared with its 1995 counterpart. Narrower confidence intervals, added to the higher reliability of the 1994 exam, allows for a greater certainty that differences among scores reflect true score differences.

The observed standard error of measurement difference of 8.16 points is used to perform a type of significance test on the difference between two scores. The purpose is to see whether or not the difference is likely to occur on the basis of chance. The observed mean score difference between 1994 and 1995 exams is 5.896 points. This value is less than the cutoffs of 15.993 points for the 95% confidence interval and 21.052 for 99%. Therefore, it is possible to say that the difference between the mean scores of the two exams is not statistically significant. There is more than a 95% or even 99% likelihood that this difference is the result of random factors.

CONCLUSIONS

This study found that the authentic English tests of Reading Comprehension analyzed here are consistent with the upper levels of the ACTFL Reading Proficiency Guidelines.

Both tests succeed on their attempt to be a "pure" reading test with authentic input. By eliminating the possibility of L2 production by demanding only L1 use, the tests better

measure the students' actual reading ability.

The reliability study performed on both tests corroborates the good quality of the tests. It confirmed the hypothesis that the reliability coefficient should be higher for longer tests, even though the difference was quite small. The tests have stable measures of internal consistency, and the average amount of points attributable to random error is small in both cases.

This study sheds light on the issues concerning authentic tests and their psychometric properties. The old belief that only objective tests lend themselves to psychometric analysis is not held. Psychometric analysis is a valuable tool which should be increasingly utilized by researchers of language testing.

Not only researchers, but also L2 and FL teachers are urged to make good use of reliability studies such as this one. They need to be reassured that the tests they construct and adopt for their classroom practice are precise measures of what they intend to be assessing.

REFERENCES

Alderson, J., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14* (2), 115-129.

Alves, I., Testa, M., & Pizolatto, C. (in press). Dr. Lieber "falou e disse" na prova de inglês do Vestibular Unicamp/94? [Did Dr. Lieber "tell it and say it" in the English test of the Unicamp Entrance Exam 94?]. *Trabalhos em Lingüística Aplicada*.. Campinas: IEL Unicamp.

American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Bastos, L., Rodrigues, S., Cherem, L., & Nery, R. (1993). *Vestibular inicamp: Inglês/Francês*. São Paulo, SP: Editora Globo.

Byrnes, H. (1987). Getting a better reading: Initiatives in foreign language reading instruction. In S. Savignon & M. Berns (Eds.), *Initiatives in communicative language teaching II: A book of readings* (pp. 171-203). Reading, MA: Addison-Wesley.

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In R. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp.). Arlington, VA: Center for Applied Linguistics.

Davis, J. (1994). Authentic assessment: reading and writing. In C. Hancock (Ed.), *Teaching, testing, and assessment: Making the connection* (pp. 139-162). Lincolnwood, IL: National Textbook Company.

Donin, J., & Silva, M. (1993). The relationship between first and second language reading comprehension of occupation-specific texts. *Language Learning, 43* (3), 373-401.

Friedenberg, L. (1995). *Psychological testing: Design, analysis and use.* Needham Heights, MA: Simon & Schuster.

McLeod, B., & McLaughlin, B. (1986). Restructuring or automaticity? Restructuring in a second language. *Language Learning, 36* (2), 109-123.

Morrow J., Jr. (1993). How significant" is your reliability? *Research Quar-*

terly for Exercise and Sport, 64 (3), 352-354.

Omaggio-Hadley, A. (1993). *Teaching language in context* (2nd ed.). Boston: Heinle & Heinle.

Peirce, B. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly, 26* (4), 665-689.

Shohamy, E., & Reves, T. (1985). Authentic language tests: Where from and where to? *Language Testing, 1* (2), 48-59.

Stevenson, D. (1985). Authenticity, validity and tea party. *Language Testing, 2* (1), 41-47.

Swaffar, J. (1981). Reading in the foreign language classroom: Focus on process. *Die Unterrichtspraxis, 14*, 176-194.

Unicamp. (1994). *Manual do candidato.* Campinas, SP: Coordenadoria Executiva dos Vestibulares da Unicamp.

Unicamp. (1995). *Manual do candidato.* Campinas, SP: Coordenadoria Executiva dos Vestibulares da Unicamp.

Valette, R. (1994). Teaching, testing and assessment: Conceptualizing the relationship. In C. Hancock (Ed.), *Teaching, testing, and assessment: Making the connection* (pp. 1-42). Lincolnwood, IL: National Textbook Company.

Victor, F., & Senatore, P. (in press). Quando vale o não dito pelo dito. [When what is said works as what is unsaid].*Trabalhos em Lingüística Aplicada*.. Campinas: IEL Unicamp.

Wiggins, G. (1994). Toward more authentic assessment of language performances. In C. Hancock (Ed.), *Teaching, testing, and assessment: Making the connection* (pp. 69-85). Lincolnwood, IL: National Textbook Company.

Wolf, D. (1993a). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal, 77* (4), 473-489.

Wolf, D. (1993b). Issues in reading comprehension assessment: Implications for the development of research instruments and classroom tests. *Foreign Language Annals, 26* (3), 322-331.

Young, D. (1993). Processing strategies of foreign language readers: Authentic and edited input. *Foreign Language Annals, 26* (4), 451-468.

APPENDIX A: 1994 ENGLISH TEST - TRANSLATION

Answer all questions IN PORTUGUESE.

17 - What is the way found by Calvin to get ten dollars from his father? Explain.

Read the following text and answer questions 18 to 20.

18 - Who is "Grandfather"?
19 - "No one in Vietnam has a clock as tall as a man." (quote in English) How does the narrator justify this statement?
20 - Why does the narrator refer to things from his country to describe Mr. Cohen?

Read the following text and answer questions 21 to 26.

21 - In Dr. Lieber's opinion, what makes his work out of the ordinary?
22 - Why was it not possible yet to test the hardness of the new material synthesized at Harvard University?
23 - To Malcolm Browne, in which aspects can the speed of light be compared to the hardness of diamond?
24 - The author of the above article states, in the fifth paragraph, that Harvard University has applied for a patent for the process of making carbon nitride. Justify the use of the expression "in any case" (quote in English) with which he introduces this statement.
25 - What is the relationship between the silicon nitride and the diamond, mentioned in the last paragraph?
26 - What should be altered if the "hardness" expected for the carbon nitride is confirmed?

Read the introduction of Frankenstein for one of the editions of the story and answer questions 27 to 30.

27 - What made the author accept the request from the editors to speak about the origins of Frankenstein?
28 - Why does Mary Shelley find natural the fact that she has been interested in writing stories since she was a child?
29 - Point to two differences mentioned by Mary Shelley between her writings and her childhood dreams.
30 - During her childhood, Mary Shelley lived on the coast of Scotland. What are her childhood feelings related to that place?

The following text refers to questions 31 and 32

31 - Based on your reading of the text, explain its title.
32 - Attribute a meaning to: (quotes in English)

a. "meander" (paragraph 1, line 7)
b. "hurling" (paragraph 4, line 9)
c. "dim" (paragraph 5, line 3)

APPENDIX B: 1995 ENGLISH TEST - TRANSLATION

Answer all questions IN PORTUGUESE.

13 - The following segment was taken from "A Martian Sends a Postcard Home," by Craig Raine. Read it and answer: What is the similarity between the rain and the television, according to the Martian?

Rain is when the earth is television.
It has the property of making colours darker.

The following text refers to questions 14 and 15.

14 - Based on the images obtained by the spacecraft "Clementine," what were the discoveries related to:
a) moon's valleys and peaks;
b) "The South Pole Aitken Basin"?

15 - How useful is it to calculate the depth of the moon's craters?

Read the following text and answer questions 16 to 18.

16 - What is needed in order to be possible to use biofuels for cooking or lighting houses?
17 - Cite two advantages of those fuels.
18 - Explain the process of obtaining biofuels.

In order to answer questions 19 to 21, read the following:

I. The introduction of the book "The Adventures of Grandfather Frog," by Thornton W. Burgess, and
II. A segment from one of its chapters.

19 - What are the stories of T. W. Burgess about?
20 - What does having a big mouth bring to Grandfather Frog?
21 - What made Grandfather Frog think about the word if??

Read the following pamphlet and answer questions 22 and 23.

22 - What induced the Dutch government to adopt its current position related to the fight against drugs?

23 - How do the ideas presented in the book conflict with other international institutions' policy of war on drugs?
24 - Explain the title of the advertisement.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

FL025473 - 79

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").