Beyond Statistics: Systematic Development of a High Stakes Reading Comprehension Exam[1]

KAREN ENGLANDER, UNIVERSIDAD DE BAJA CALIFORNIA

Since multiple choice tests continue to be used for assessing student performance, their manner of construction is critical to both test creators and, especially, test takers (Thorndike 1971). Proper interpretation of test results is dependent upon the test being reliable and valid. Statistical measures of reliability and validity assure that suitable conclusions regarding the test-taker's ability can be drawn. When the test has important consequences (high stakes) for the test-taker, such as admission, graduation and/or certification, the validity of the test is crucial (Messick 1995). Several practitioners (Alderson 1990a, Cronbach 1984, Baxter & Glaser 1993, Paxton 2000) have stated that verbal reports and protocol analysis might help evaluators understand why test takers choose the answers they do. In this study, verbal reports were used to verify that items on a standardized exam were valid according to the testing objective of each item. Results showed that, in fact, some items that were statistically acceptable became invalid when subjected to a verbal report.

This paper will discuss the systematic development of the *Universidad Autónoma de Baja California*'s English Exit Exam (EXEDII). Criterion-referenced testing will be contrasted with norm-referenced testing and the role of construct validity will also be examined. The method of using verbal reports is presented and the two specific methods of validating the construct of this particular exam--statistics and verbal reports--are outlined. Results and discussion which demonstrate that statistics alone cannot provide sufficient information to determine the validity of an exam item follow.

Description of the English Exit Exam

At the *Universidad Autónoma de Baja California* we believe that all graduates must have at least an intermediate command of English to be successful in further study or career. Consequently, educational testing experts and English as a foreign language teachers collaborated in creating a high stakes exit exam of English language proficiency called EXEDII. All undergraduates who have not taken three semesters of English language instruction (approximately 480 hours) must pass the EXEDII before they can receive their graduation diploma. The exam is comprised of 100 multiple choice items. Reading comprehension is one

[1] This is a refereed article.

of three areas or *domains* (listening comprehension and grammar are the other two) in which students must demonstrate competency. The reading comprehension section of the EXEDII was created by a team of instructors (Adriana Usabiaga, Leonora Velasco, Bill Richter and the author) at the *Escuela de Idiomas*.

Criterion-referenced testing of reading comprehension

The exam was built as a criterion-referenced test. Following from the work of Glaser (1963, 1973) and Popham and Husek (1969) these tests are different from norm-referenced tests. The TOEFL and TOEIC are well-known norm-referenced tests; that is, the tests provide a measure of performance interpretable in terms of an individual's standing in relation to all the other people who took the test at that time. In contrast, criterion-referenced tests are interpretable in terms of clearly defined, specified tasks or "behavioral objectives." The test administrator can judge student performance in relation to specific objectives (Hambleton 1994)--in this case whether or not the student possesses intermediate-level competence in English--and not be concerned with how the student performs in relation to other students.

For our reading comprehension exam, the skills and sub-skills which characterize reading comprehension had to be identified. Munby's (1978) classic taxonomy of 260 language skills served well. Fifty-seven of these skills are relevant to reading (see Appendix 1). To obtain appropriate representation of the L2 reading comprehension skill, we combined the specific skills outlined by Munby, the interactive strategies of making meaning (Carrell, Devine & Eskey 1988), and the top-down/bottom up processing of text (Grabe 1988).

For our purposes, the exam had to be representative of the domain of reading, and it had to determine whether students had an intermediate level of proficiency. To establish an intermediate level, we adopted the skills and strategies listed in the scope and sequence overview provided in the textbook[4] used in the university's intermediate-level EFL classes as our starting point. Each skill or strategy related to reading was translated into a behavioral objective (see Appendix 2). Each objective was measured by particular test items. It was assumed that if the item functioned successfully, and the student possessed the targeted proficiency level, then the student would arrive at the correct answer. If the student chose an incorrect answer, it would be the result of not possessing the targeted proficiency level.

This concept of assuring that the exam tests what we want it to test raises the issue of validity. Validity has traditionally been construed as having three aspects: content, construct and criterion. (Standards,

[4] *On Target One* (1990) Scott Foresman, Longman

1974, 1984, APA, AERA & NCME). These three aspects of validity determine what the exam is testing, whether that exam is representative of that ability, and, if the student passes that exam, whether he or she will be able to perform the ability the exam is intending to test (see Moss 1992 for a synthesis). In other words, we want to be sure the exam is testing reading comprehension (and not spelling); that it is testing understanding of sentence connectors (because that is a reading sub-skill); and that it is testing at the intermediate level (and not higher, nor lower). For the purposes of this study, we were particularly interested in verifying construct validity: Whether the score really reflected the test-taker's ability in that skill (in this case, reading comprehension at the intermediate level).

Talk-aloud protocols – Possibilities and limitations

Verbal reports, variously called think-aloud, talk-aloud, self-reports, self-observations and self-revelations, have been used to identify the mental processes readers use to understand the printed word (Cohen 1987, Alderson 1990b, Anderson 1991, Presley & Afflerbach 1995). To perform a verbal report, a person is asked to say aloud everything that he or she is thinking while performing a task. Ericsson & Simon (1984, revised 1993) claim that these verbal reports can be systematically analyzed through a procedure called protocol analysis to understand cognition (Newell & Simon 1972). Verbal reports allow researchers a "window...to peer into the workings of the mind" (Smagorinsky 1998).

Protocol analysis was first proposed as a way of revealing actual cognitive processes (Ericsson & Simon, 1984). However critics of the method, based on Vytgosky's (1987) conception of thought and words, argue that in the process of forming words, thought (cognition) itself is altered. Thought is much like a storm cloud, whereas speech is a shower of words (Smagorinsky 1998). Words are differentiated and then put in sequence to form speech, which then forms verbal reports. In other words, speech is not a "window" to cognition because cognition is altered--some would say created--through speech.

Critics argue that Ericcson & Simon's premise is wrong, and verbal reports and protocol analysis cannot let us know *how* people think. In our study, we were interested in verbal reports not as windows to the cognitive processes of the mind itself, but rather as a report of activity. In other words, by asking students to talk aloud as they answered the test items, we were not looking for evidence of how their mind worked, we simply wanted to know what information, which was available in their short-term memory, they were using to answer the questions. Our aim is well suited to the methodology of verbal reports and avoids the criticism of researchers such as Smagorinsky (1998).

Statistical evidence

Educationel statisticians from the *Instituto de Investigación y Desarrollo Educativo* analyzed the performance of 710 students who took the exam. The measures of difficulty, discrimination, and discrimination coefficient produced numbers all in the acceptable range. Table 1 shows the ten items that were later subjected to verbal reports. These particular ten items were chosen simply because they represent the items corresponding to the first long text of the exam (items 64-69), and a selection of four items corresponding to four of the seven short texts (items 94, 96, 97, and 99) of the exam.

Item	p	D	r^{pbs}	Item #	p	D	r^{pbs}
#64	.62	.50	.40	#69	.38	.49	.40
#65	.54	.71	.57	#94	.45	.76	.59
#66	.46	.30	.21	#96	.24	.48	.41
#67	.65	.34	.31	#97	.35	.55	.45
#68	.52	.43	.32	#99	.37	.64	.50

TABLE 1. MEASURES OF THE PILOTED ITEMS BASED ON 710 TEST TAKERS.

To interpret these statistics, the following definitions were used. The symbol p equals difficulty. The closer p is to 1, the easier the item is, and conversely, the closer it is to 0, the more difficult. According to Thomson & Levitov (1985) the ideal difficulty of an item on a 100-item test with four multiple choice options for each is p = .63.

The symbol D indicates the discriminatory ability of the items and is calculated using the top 27 percent and bottom 27 percent of the scores. The higher the D number, the more likely that the high scorers got this item right and the low scorers did not. If all test takers get it right, or all get it wrong, the item is not discriminating well and its value is: D = 0. The figures can be interpreted as follows (Ebel & Frisbie 1986):

If D = .40 or higher, the item is very good;
.30 to .39, the item is reasonably good but subject to improvement,
.20 to .29, the item is a marginal item and needs revision;
less than .19, the item is poor.

The term r^{pbs} is similar to D but calculates the discriminatory power of an item based on the scores of all the test-takers, not just the 54 percent taken into account for determining D. The statistics reveal that item 66 is weak based on its discriminatory ability although at the same time it is not an especially easy question. Items 69, 96, 97 and 99 are quite difficult but function well in discriminating high-ability scorers from low-ability scorers. Consequently, based on these statistics, only item

66 might deserve a closer look. The other nine items functioned adequately statistically.

We, however, were primarily interested in the language interaction of the texts, items and responses. A pilot study of verbal reports was initiated.

Method

Ten items (see Table 1) were selected from the 34-item ding Comprehension Section of the EXEDII exam. They represented a 120-word text *(Chocolate: A World Favorite)* and its accompanying six questions, and four questions each relating to its own two- to three-sentence text.

Four students from different classes volunteered to participate. Each student had recently completed the Intermediate level at the Language School and so was presumed to possess the targeted proficiency level for this exam as determined by its criterion-referenced construction. Each student sat in a classroom with one member of the research team who had never had contact with that student. The students were instructed to "keep talking" while reading the texts and completing the ten test questions. They were permitted to use their L1 (Spanish) or L2 (English) as they wished. To determine what information in the text itself was being used, students were also prompted with questions such as, "What in the text helped you choose that answer?", "How did you arrive at that answer?" or "How do you know?" The four verbal reports were tape-recorded and later transcribed.

Results

Strikingly, two of the ten items we piloted were found to violate the exam's specifications; in other words, the item was not testing what it was intended to test. All four students chose an incorrect answer for item 66 using a similar rationale. The item was intended to determine the student's ability to differentiate fact from opinion. It reads:

> In this text, "chocolate is more wonderful when candy makers combine it with other ingredients" is:
> A. an opinion of somebody other than the author
> B. a generally accepted fact
> C. an opinion of the author
> D. a verified fact

The prompt in the text reads, "Some people think that..." Note three of the students' verbal reports:

> *Tomas:* "Some people think' is an opinion but it cannot be letter A. 'A' say an opinion of somebody and people no is somebody, is many people, so I say, hm, I say the answer is letter B, a generally accepted fact."

Both Maribeth and Eduardo hint at the same problem.

Maribeth: They mention *people* (her emphasis) like it. It mentions people so it be generally accept.

Eduardo: It's a generally accepted fact because it's talking about the people.

The fourth student, too, chose B – a generally accepted fact.

Item 69, in which the students were asked to identify and correctly infer the author's opinion, was also problematic. The specifications require that the author's opinion be demonstrated through qualification of an adjective or adjectival phrase.

The item reads:

Complete the following sentence with the best option. The author thinks that...

A. chocolate with fruits, nuts or coconuts can be heavy.
B. chocolate with shapes, such as flowers, are nice.
C. sugar and fat are bad for health.
D. chocolate with ingredients is sold all over the world.

While the correct answer is B, Maribeth tells us, "The author thinks that chocolate with ingredients is sold all over the world." Eduardo, who struggles with this one, and is tempted by D, finally decides, "Okay, in the title it says 'Chocolate: A World favorite' I think it's gonna be letter D." The use of the word "world" in the title and nowhere else in the text has an overwhelming influence which was not considered in preparing that item. This distractor is clearly misleading.

Discussion

The high stakes nature of the EXEDII, i.e. students must pass it in order to receive their undergraduate degrees, requires that the validity of the exam be closely evaluated.

The care which was taken in developing the Reading Comprehension Section of the EXEDII seems to meet virtually all the relevant criteria in order to provide a heuristic framework for test interpreters. This process is outlined below.

The EXEDII test developers specified the boundaries of the domain to be assessed by looking at the psycholinguistic process of reading and then referencing Munby's (1978) taxonomy of skills. The content of the exam is relevant to the domain and is representative of it. Cronbach (in Moss 1992) recommends "inspecting items" as a means of gathering evidence, and this was also done.

To comply with construct validity, we set an intermediate level of proficiency as the desired level within the domain of reading comprehension, and adopted the syllabus definition of which skills constituted that content domain on this particular exam. Further statistical analysis of item difficulty, discrimination and discrimination co-efficient generated evidence that the interpretation of the scores was valid. Messick (1995) suggests another method of gathering evidence of an exam's

construct validity: the substantive aspect. This refers to the theoretical rationales for consistency in test responses. We constructed the Reading Comprehension Section of the EXEDII based on a psycholinguistic understanding of the interactive nature of reading, incorporating both bottom-up and top-down processing which is consistent with the theoretical underpinnings of the domain. Careful attention to the specifications for items served to ensure that skills and sub-skills were balanced and complete. The substantive aspect of validity requires empirical evidence that the theoretical processes are actually engaged in by the respondents in the assessment task. To obtain this evidence, Cronbach (cited in Moss 1992) suggests administering the test to individuals who are asked to think aloud.

Think-aloud procedures and the interpretation of the data collected are not transparent operations. Smagorinsky (1998) undermines Ericsson & Simon's (1984, 1993) assertion that cognitive processes can be revealed through verbal reports and protocol analysis. Bearing in mind these limitations of verbal reports, we set the following goal: To elicit what students said about how they chose a particular answer from among the four multiple choice options. We adopted Ericsson & Simon's assertion that short-term memory is accessible. We did not presume that we could uncover cognitive processes. The verbal reports were elicited to reveal what specific data test-takers were using to solve the task. Test items which seem well constructed both by observation and statistical analysis still need to be subjected to this evidentiary procedure (Moss 1992).

As we demonstrated, directed verbal reports confirmed test items that operated as intended, and revealed problems with several items that otherwise had gone undetected. At this time we are revising the problematic items. Then we will subject these items to both the statistical procedures of difficulty and discrimination and to verbal reports. We also intend to undertake research using verbal reports with the two other sections of the EXEDII: Grammar and Listening Comprehension. As our work with verbal reports for the Reading Comprehension Section indicates, this investigative technique is indispensable in assuring the validity of the substantive aspect of an exam. As test creators we are compelled to use all the available evidentiary procedure – and not rely solely on statistics – especially when the consequences of examination scores carry high stakes.

REFERENCES

Alderson, J.C. (1990a). Testing reading comprehension skills (Part one). *Reading in a Foreign Language, 6*, 425-438.

Alderson, J.C. (1990b). Testing reading comprehension skills (Part two): Getting students to talk about taking a reading test (A pilot study). *Reading in a Foreign Language, 7*, 465-503

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974, 1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, N.J. (1991) Individual differences in strategy use in second language reading and testing. *Modern Language Journal, 75*, 460–472.

Baxter, G.P. & Glaser, R. (1992). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues & Practice, 17*, 37-45.

Carrell, P., Devine, J. & Eskey, D. (Eds.). (1988). *Interactive approaches to second language reading*. New York: CUP.

Cohen, A.D. (1987). Using verbal reports in research on language learning. In Faerch, C. & Kasper, G. (Eds.) *Introspection in Second Language Research*. Clevedon: Multilingual Matters.

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

Cronbach, L.J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper and Row.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Ericsson, K.A. & Simon, H.A. (1984, revised 1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist 18*, 519-521.

Glaser, R. (1971) Educational psychology and education. *American Psychologist, 27*, 557-566.

Grabe, W. (1988). Reassessing the term "interactive." In Carrell, P., Devine, J. & Eskey, D. (Eds.). *Interactive approaches to second language reading*. New York: CUP.

Hambleton, R.K. (1994) The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice, 13*(4) 21-26.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4) 5-8.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*. 229-258.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: CUP

Newell, A. & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, New Jersey: Prentice Hall.

Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education, 25*, 109-119

Popham, W.J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.

Popham, W.J. & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Smagorinsky, P. (1998). Thinking and Speech and Protocol Analysis. *Mind, Culture and Activity, 5*, 157-177

Thomasson, B. & Levitov, J.E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer 3*, 163-168.

Thorndike, R.L., (Ed.). (1971). *Educational measurement*, (2nd ed.). Washington DC: American Council on Education.

Vygotsky, L.S. (1987). Thinking and speech. In R. Rieber & A Carton (Eds.), L.S. Vygotsky: collected works (Vol. 1, 39-285). New York: Plenum

APPENDIX 1
Munby's (1978) taxonomy of skills relevant to reading.

19 Deducing the meaning and use of unfamiliar lexical items, through understanding of word formation:
- stems/roots
- affixation
- derivation
- compounding

19.2 contextual clues

22 Understanding information in the text, not explicitly stated, through
- making inferences
- figurative language

24 Understanding conceptual meaning, especially
- quantity and amount
- definiteness and indefiniteness
- comparison; degree
- time (esp. tense and aspect)
- location; direction
- means; instrument
- cause; result; purpose; reason; condition; contrast

26 Understanding relations within the sentence, especially
- elements of sentence structure
- modification structure
- negation
- modal auxiliaries
- inter-sentential connectors
- complex embedding
- focus and theme;
- thematic fronting; and inversion
- postponement

31 Understanding relations between parts of a text through grammatical cohesion devices of
- reference
- comparison
- substitution
- ellipsis
- time and place relaters
- logical connectors

37 Identifying the main point or important information in a piece of discourse, through
- topic sentence, in paragraphs of
- inductive organization
- deductive organization

39 Distinguishing the main idea from supporting details, by differentiating primary from secondary significance
- the whole from its parts
- a process from its stages

a category from its exponent
statement from example
fact from opinion
a proposition from its argument

45 Skimming to obtain
the gist of a text
a general impression of a text

46 Scanning to locate specifically required information on
a single point, involving a simple search
a single point, involving a complex search
more than one point, involving a simple search
more than one point, involving a complex search
a whole topic.

APPENDIX 2

Relation between Reading Comprehension Skill, EXEDII Exam Behavioral Objective for the Item, and Munby's Taxonomy of Skills Relevant to Reading

Skills or strategy from textbook	Behavioral objective (Translated from Spanish)		Munby's code
Recognize main idea	#1	Identify and understand the main idea, directly stated in text	37.4
	#7	Identify and understand the main idea, not directly stated in text	37.4
Recognize main idea distinct from secondary ideas	#15	Identify and understand the main idea, not directly stated in text but paraphrased from secondary ideas	39
Understand information in sequence	#27	Identify and understand specific information: sequence	32.6
Recognize an appropriate title	#24	Identify and understand the main idea: select a title	45.1
Scan for specific information: numerical or lexical	#2	Identify and understand numerical information	46.1, 46.2, 46.3, 46.4
	#23	Contrast information in different parts of the text	
	#26	Infer logically information that requires simple calculation	24.1
Identify and use examples and/or definitions	#8	Identify and understand specific information located in amplification of an idea through explanation or definition	28
	#14	Identify and understand specific information in an example	28
Identify what the example exemplify	#10	Identify and understand specific information which follows a discourse marker of location	32
Recognize the difference between comparison and superlative	#33	Identify and understand the difference between the comparative and superlative	24.3
Identify correct and incorrect inferences	#12	Infer logically objective information	24.7
Differentiate fact from opinion	#9	Identify and differentiate fact from opinion: a fact	32.1
	#19	Identify and differentiate fact from opinion: identify different options	32.1
Identify opinions and/or beliefs of the author or others	#3	Identify and differentiate fact from opinion: opinion of the author or of others	32.4, 32.5
	#6	Infer logically the author's attitude	

Recognize markers of sequence	#32	Understand discourse marker: chronology	24.4, 28.5
	#34	Identify and understand the sequence expressed by adverbs or prepositions	24.5
Recognize markers of cause or reason	#30	Understand a discourse marker, result or consequence	24.7
Recognize pronoun reference, including relative clauses	#21	Identify and understand a reference of a personal pronoun or pronominal phrase	32.3
	#25	Identify and understand reference information in a relative clause	32.1
	#28	Identify and understand pronoun reference of 'that' or 'which'	32.1
	#29	Identify and understand pronoun reference of 'that' or 'who'	32.1
Identify words in context by explanations	#4	Understand vocabulary in context; clue: rhetorical structure of example	19.2
	#20	Understand vocabulary in context; clue: expansion that explains the meaning	19.2
Identify words in context by synonym in adjacent sentence	#17	Understand vocabulary in context; clue: a synonym in parallel position in adjacent sentence	19.2
	#18	Understand vocabulary in context; clue: definition given as reference	19.2
Identify words in context using contrast or addition marker	#31	Understand discourse marker of addition	36.6
Identify words in context using similarity marker	#10	Understand vocabulary in context; clue: markers of comparison, similarity or parallelism	32.2
	#16	Understand vocabulary in context; clue: marker of example and examples that illustrate the meaning	28.5
Identify words in context using affixes	#11	Understand vocabulary in context; clue: word with affixes	19.1
Identify words in context using direct object of the verb	#5	Understand vocabulary in context; clue: complement or direct object of the verb	19.2