Políticas lingüísticas

Validation Study of Colombia's ECAES English Exam[1]

Alexis Augusto López
Gerriet Janssen
Universidad del los Andes
Bogotá, Colombia

In 2005, Colombia's Ministry of National Education developed a nationwide program to improve the teaching and learning of English. As one of this project's initial steps, the Ministry developed two English tests included in the Common European Framework of Reference (CEFR). This paper presents the findings from a study that examined the validity of one of these tests, the ECAES English Exam. Data for this study was gathered using content evaluation sessions with teachers and think-aloud protocols with students. Our findings suggest that this test may not be a valid measure of the students' general English language proficiency.

Key words: test validity, test usefulness, validation, validity evidence

Estudio de validación de la prueba de inglés de ECAES en Colombia
En el año 2005, el Ministerio de Educación Nacional (MEN) de Colombia desarrolló un programa a nivel nacional para mejorar y fortalecer la enseñanza y aprendizaje del inglés. Como una de sus primeras iniciativas de este proyecto, el MEN diseñó dos exámenes de inglés alineados con el Marco Común de Referencia Europeo. Este manuscrito presenta los hallazgos de un estudio que evaluó la validez de uno de esos exámenes, el Examen de Inglés de ECAES. Los datos para este estudio se recogieron a partir de unas sesiones de evaluación de contenido con los profesores y unos protocolos de pensar en voz alta con los estudiantes. Nuestros hallazgos sugieren que los resultados de este examen no son válidos para medir la suficiencia general en inglés de los estudiantes.

Palabras claves: validez de una evaluación, utilidad de una evaluación, validación, evidencia de validez

---

1 This research study is based on a paper presented at the American Association of Applied Linguistics 2009 Conference entitled, "Validation study of Colombia's ECAES English Exam." The research study was funded by the Fondo de Apoyo a Profesores Asistentes –FAPA– at Universidad de los Andes. The study started in February 2008 and ended in December 2009.

Étude de validation d'examen d'anglais E.C.A.E.S. de la Colombie
En 2005, le Ministère de l'Éducation Nationale (M.E.N.) de Colombie a développé un programme dans tout le pays ayant pour but d'améliorer et fortifier l'enseignement et l'apprentissage de l'anglais. Pour y parvenir, le M.E.N. a tout d'abord proposé deux épreuves d'anglais inscrites dans le Cadre européen commun de référence pour les langues du Conseil de l'Europe. Ce document présente les résultats d'une recherche qui a évalué la pertinence d'une des épreuves proposées, «l'examen d'anglais E.C.A.E.S.» Les données de cette étude ont été recueillies à partir de séances d'évaluation du contenu avec les enseignants et les protocoles de penser à haute voix avec les élèves. Nos recherches montrent que les résultats de cette épreuve ne sont pas en mesure de déterminer chez l'étudiant sa compétence générale en anglais.

Mots clés: Validité d'une évaluation, bénéfice d'une évaluation, validation, preuves de validité

INTRODUCTION

Figueras, North, Takala, Verhelst, and Van Avermaet (2005) have documented a current trend to align language programs and exams to the Common European Framework of Reference (CEFR) levels and descriptors most likely as a response to educational reform initiatives and to educational accountability systems. From an educational perspective, reform may be sought as an "expression of concern with how well schools are functioning and the quality of educational outcomes and/or student learning" (Chalhoub-Deville, 2008, p. 12). Furthermore, educational reform might signal a deep belief that "education lies at the heart of economic development, international competitiveness, and social harmony" (Chalhoub-Deville, 2008, p. 12).

Linn (2000) explains test-based accountability as the "engine" of educational reform. Tests are created and used for various reasons: attractiveness to stakeholders such as the public, politicians, and policymakers, cost-effectiveness. Another reason tests are used concerns tangibility: teachers and administrators can be held responsible for gains or losses described by test scores (Chalhoub-Deville, 2008; Linn, 2000). Many programs link course syllabi and exams to the CEFR hoping that student progress can be specifically documented, using a scale that is widely recognized and understood.

CEFR, educational reform, and test-based accountability are not limited solely to the European context; indeed, these three topics are currently important issues in many countries around the world, including Colombia. Law 115 of 1994 called for the acquisition of elements of conversation, reading, comprehension, and the capacity to express oneself in at least one foreign language (Ministerio de Educación Nacional, 2005). However, it is only in the Colombian Ministry of National Education's more recent 2004 mandate, *Programa Nacional de Bilingüismo 2004-2019* **(better known as** *Colombia Bilingüe***)**, that CEFR-based instruction and language testing has gathered force in Colombia (Ministerio de Educación Nacional, 2005). *Colombia Bilingüe* documentation rationalizes Colombian Spanish-English bilingualism as a reaction to global economic pressures: in times of globalization, the country needs to develop its citizens' capacity to be proficient in a foreign language. In this context, *Colombia Bilingüe* posits new standards of communicative competencies in a foreign language: English (Ministerio de Educación Nacional, 2005).

Colombia Bilingüe and the inclusion of the CEFR-aligned instruction and exams are designed to improve the teaching of English in Colombia. The governmental educational body ICFES (the Colombian Institute for the Promotion of Higher Education) writes that the use of the CEFR will "act as a source of information in the construction of evaluative indicators in service of the educational sector, so as to encourage the assessment of institutional processes, policy formulation and facilitate the decision making process in all levels of the educational system" (2009: para. 3).

By virtue of the CEFR and the exams created for *Colombia Bilingüe*, the Colombian government and other stakeholders intend to assess English language proficiency of those who are in their last year of undergraduate and high school academic programs (Ministerio de Educación Nacional, 2005). The Ministry of National Education also states that the CEFR levels and skill descriptions will strengthen and focus the teaching and learning of English within the country (Ministerio de Educación Nacional, 2005).

To gauge potential changes in English language proficiency arising from *Colombia Bilingüe*, the Ministry of National Education has included CEFR-based measurement standards in its 2005 mandate. According to this plan, by 2019 all graduating high school students should perform in English at a CEFR B1 level, while undergraduate university students will do so at a CEFR B2 level (Ministerio de Educación Nacional, 2005).

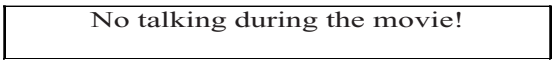
To measure whether or not graduates from high school and university have met the two above-stated *Colombia Bilingüe* goals, the Ministry of National Education has also mandated the design and implementation of two CEFR-based English proficiency exams (Ministerio de Educación Nacional, 2005). Under direction of ICFES and in collaboration with the British Council-Colombia and Cambridge ESOL, teams of local item writers created the *Examen de Estado* [Exam of the State] English Exam. The *Examen de Estado* English exam is designed to be taken by all high school students as part of a series of exams given during the last year of high school. The other exam similarly and simultaneously developed was the *ECAES* (Examen de Calidad para Educación Superior) English Exam. The ECAES English Exam is designed to be taken by undergraduates as a part of a battery of ECAES exams given during the last year of university. This study focused on the ECAES English Exam, as it directly relates to the university level English language students, professors, and other relevant stakeholders immediately available to this research team.

The ECAES English Exam

Students take the ECAES Exam during their last year at the university. Students take a battery of three exams, one of which is related to their field of study to prove proficiency in the subject area. The other two exams are the same for all students: a Spanish reading exam and the English exam. The ECAES English Exam can be considered a low-stakes test since it is only used to inform the public about the effectiveness of the new language education policy, and no important decisions are taken based on test scores.

The ECAES English Exam is a seven part, selected response exam (45 items). The exam only assesses the following skills: reading, vocabulary and grammar. Listening, speaking and writing are not assessed because of practicality reasons (difficulty in setting up adequate technology to assess listening and difficulty in training raters for speaking and writing). The following sections describe each part of the test. A full sample ECAES English Exam can be downloaded from the ICFES website <www.icfes.gov.co>. Below we provide a short description of each part of the exam and show sample items developed by the researchers to illustrate the content of the exam.

Part 1. This part includes five three-option multiple-choice items. It assesses the test-takers' ability to understand short notices or signs. The task requires test-takers to read the notice or sign and then identify where it would be seen by choosing the best option. Figure 1 shows a Part 1 sample item.



A. at a cinema*
B. at a restaurant
C. at a bank

Figure 1. Part 1 sample item

Part 2. This part includes five items matching questions with eight options. This part asks test-takers to match definitions to a list of options from one lexical category (e.g. things you can find in a kitchen, professions, and classroom materials). The questions are very short definitions similar to the ones that are found in a dictionary. Figure 2 shows a Part 2 sample item.

1. You use one of this to write a note. (C)
2. You use this to delete what you write. (A)

A. eraser
B. glue
C. pencil
D. ruler

Figure 2. Part 2 sample item

Part 3. This part includes five three-option multiple-choice items which ask test-takers to complete a conversation. All the items are short conversations between two speakers; the stem is something the first person says and the options are three possible responses. It assesses the test-takers' ability to understand and use the English language in everyday life endeavors. Figure 3 shows a Part 3 sample item.

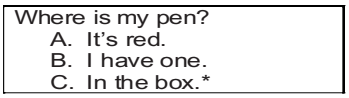


Figure 3. Part 3 sample item

Part 4. This part is a grammatically oriented, eight three-option items cloze exercise in a modified, authentic text. Test-takers have to read a short informational text. In the text there are some missing

words and test-takers have to choose the best option to fill in each of the gaps. The words could be verb forms, articles, prepositions, conjunctions, or pronouns among others. Figure 4 shows a Part 4 sample item.

Because of global warming, the North Pole could start to melt __1__ just a few centuries. As the North Pole melts, sea levels will rise __2__ several feet, causing beaches to flood and disturbing many plant and animal habitats.

1. A. in* B. on C. at
2. A. for B. by* C. from

Figure 4. Part 4 sample item

Part 5. This part is a seven three-option items, multiple-choice reading comprehension task. This part requires test-takers to read a short modified factual text and then answer some questions related to scanning relevant information in the text. Figure 5 shows a Part 5 sample item.

Because of global warming, the North Pole could start to melt in just a few centuries. As the North Pole melts, sea levels will rise by several feet, causing beaches to flood and disturbing many plant and animal habitats. As Earth gets warmer, cold areas will no longer be cold. Warm areas may become too hot for people to live or grow crops. Some scientists believe this will cause problems in the world's food supply.

What could happen if the North Pole starts melting?
A. The place where animals live will be affected.*
B. The beaches will disappear after a few centuries.
C. The plants in the area will grow taller.

Figure 5. Part 5 sample item

Part 6. This part has five four-option multiple-choice questions which challenge test-takers to answer reading comprehension questions based on a short modified authentic text. The questions are related to the author's purpose, attitude or opinion, inferential meaning, recalling details, and global meaning. Figure 6 shows a Part 6 sample item.

Because of global warming, the North Pole could start to melt in just a few centuries. As the North Pole melts, sea levels will rise by several feet, causing beaches to flood and disturbing many plant and animal habitats. As Earth gets warmer, cold areas will no longer be cold. Warm areas may become too hot for people to live or grow crops. Some scientists believe this will cause problems in the world's food supply.

What is the writer trying to do in the text?

A. Inform people about the dangers of global warming*
B. Criticize people who create global warming
C. Give his opinion about global warming
D. Invite people to protest against global warming

Figure 6. Part 6 sample item

Part 7. This part has ten four-option multiple-choice cloze exercises. It assesses both grammar and vocabulary. Test-takers are required to read a short modified, informational text. In the text there are some missing words and the test-takers to have to choose the best option that best fills each gap. The grammar items could be function words and the vocabulary items could be content words. Figure 7 shows a Part 7 sample item.

As Earth gets warmer, cold areas will no longer be cold. Warm areas may __1__ too hot for people to live __2__ grow crops. Some scientists believe this will cause problems in the world's food supply.

1.	A. transform	B. become*	C. convert	D. alter
2.	A. nor	B. yet	C. so	D. or*

Figure 7. Part 7 sample item

THEORETICAL FRAMEWORK

It is important to first highlight that both researchers of this study were members of the ECAES English Exam development team. During the test development process, we took an effect-driven approach (Lopez, 2008). This approach reflects Shohamy's (2001) view of critical language testing, wherein test developers "need to develop critical strategies to examine the uses and consequences of tests, monitor their power, minimize their detrimental force, reveal the misuses, and empower test takers" (p. 131). This stance towards language testing—which these two researchers share—has encouraged us to examine the validity of the ECAES English Exam.

Test Validity

The concept of validity has shifted over the years. Initially, validity was seen as a characteristic of a test, but now validity is seen as an argument that deals with how a test is used and how the results of a test are interpreted (Chapelle, 1999). Traditionally, validity had been defined as "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231) and it was usually subdivided into three different categories: content, criterion-related and construct validity (Bachman, 1990). Content validity referred to the extent to which a test reflected the content domain that it was intended to measure; criterion-related validity referred to the extent to which a test could be used to draw inferences regarding the criterion; and construct validity referred to the extent to which a test measured some psychological trait or theoretical concept it was intended to measure(Bachman, 1990). Messick argues that "content-related inferences are inseparable from construct-related inferences" (1988, p.38).

Messick (1989) proposed a framework of validity that includes what he calls consequential validity: the consequences of test score interpretation and use. He defines validity as an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions resulting from test scores. In this framework, Messick presents a unified concept of validity and argues that a unified validity framework could be constructed by distinguishing two interconnected facets of the unitary validity concept. Specifically, he argued that "one facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of the testing, being either interpretation or use" (p. 20).

According to Messick (1989), there are two major threats to construct validity: construct underrepresentation and construct-irrelevant variance. Construct underrepresentation occurs when the construct of a test or assessment is too narrow and fails to include important aspects of the construct. On the other hand, construct-irrelevant variance occurs when the test or assessment is too broad and construct contains extraneous variables that could make the text irrelevantly difficult (construct-irrelevant difficulty) or too easy (construct-irrelevant easiness) for some students. Today, the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation. Evidence may

be based on test content, response processes, internal structure, relations to other variables and consequences of testing (AERA, APA & NCME, 1999). By test content, we mean the specification of the boundaries of the construct domain that is being assessed in the test, the knowledge, skills or abilities that are revealed by each assessment task (Messick, 1989). Response processes refer to the match between the cognitive processes that test takers actually use when completing a measure and the process that they should use (AERA, APA & NCME, 1999). A test's internal structure is the way different parts of a test are related to each other (AERA, APA & NCME, 1999). On the other hand, associations with other variables refer to the match between the scores on a test with the scores on similar tests (AERA, APA & NCME, 1999). Finally, consequences of testing refer to the social consequences of using a particular test for a particular purpose (Messick, 1989). In fact, "tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores... A fundamental purpose of validation is to indicate whether these specific benefits are realized" (AERA, APA & NCME, 1999, p. 16). Below we describe an alternative validity criteria presented by Bachman and Palmer (1996), specific to language testing.

Test Usefulness

Although most discussions of validity remain grounded in the traditional framework, there have been proposals that address the need for additional, alternative validity criteria, such as Bachman and Palmer's (1996) conceptualization of test usefulness. In their model of usefulness, Bachman and Palmer (1996) also separate validity from test consequences. They approach the consequences of tests from the viewpoint of test usefulness. Their framework of test usefulness includes construct validity, reliability, authenticity, interactiveness, practicality, and impact. Bachman & Palmer (1996) define these test qualities as:

- Construct validity refers to "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (p. 21).
- Reliability refers to the "consistency of measurement" (p. 19)
- Authenticity refers to "the degree of correspondence of the characteristics of a given language test task to the features of a target language use (TLU) task" (p. 23).

- Interactiveness refers to "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (p. 25).
- Practicality refers to "the relationship between the resources that will be required in the design, development, and use of the test resources that will be available for these activities" (p. 36).
- Impact refers to the effect tests have on "society, educational systems, and upon the individuals within those systems" (p. 29).

Several types of evidence should be used to build a case for valid test use. In this study, we are presenting evidence based on some of Bachman and Palmer's (1996) test qualities: 1) evidence based on the construct, interactiveness, authenticity and impact of the test. This evidence is then used to form a validity argument–which cumulatively presents a case for or against the assessment potential inferences and assumptions (Kane, 1992). After presenting all the different types of evidence in favor or against the validity of a test, a validity conclusion is presented (Shepard, 1993).

Although the ECAES English Exam has been designed to provide English language proficiency measurements for purposes of accountability, the exam's validity has not been established. There have been a couple of studies conducted in Colombia, which have examined the impact or washback of the ICFES English Exam (Barletta Manjarres, 2005; Barletta Manjarres & May Carrascal, 2006), but we did not find any studies examining the validity or washback of the ECAES English Exam. Thus, we feel a study such as the one we present here may contribute to the academic discussions concerning *Colombia Bilingüe*, CEFR exams in Colombia, or the ECAES English Exam as we provide empirical evidence of the validity of *Colombia Bilingüe*'s new CEFR-based ECAES English Exam. This might become especially relevant if this test is used to inform the public about the effectiveness of Colombia Bilingüe, to evaluate English language programs at Colombian universities, and to measure students' English language proficiency. In order to examine our claims we set out to answer the following research questions:

a. To what extent does the construct of the ECAES English Exam adequately measure the levels of proficiency defined by the CEFR?

b. To what extent are the tasks on the ECAES English Exam similar to the tasks Colombian university students do in real life?
c. What impact does the ECAES English Exam have on Colombian university English programs?

METHODOLOGY

This section provides an overview of the methods employed to examine the validity of the ECAES English Exam. We used data collected from a content evaluation of the ECAES exam with teachers and think-aloud sessions with students taking this test. This section provides a description of the participants, the data collection instruments and the data analysis procedures.

Participants

Teachers. A group of 15 university English language teachers, from public and private institutions, participated in the content evaluation (described below) of the ECAES English Exam. Seven of the teachers worked in Bogota, four in Manizales, two in Pereira, one in Armenia, and one in Medellin. Participants were selected based on their availability to participate in the study, their familiarity with the CEFR, and their experience teaching English in Colombian universities. In individual sessions, each teacher was asked to describe what each item measured, code the items against the CEFR, describe the degree to which the item matched teaching goals from within their own program of study, and provide their perceptions about the test. More information is provided in the Data Collection section.

Students. A total of 13 university students from different universities in Bogota participated in this study. They were selected based on the following criteria: 1) university students in their final year of their programs, 2) students who had taken English courses at their respective universities, and 3) their availability and willingness to take the ECAES English exam. Students were identified by their English teachers as potential participants. We tried to get a balance of high-proficiency, intermediate, and low-proficiency students, as defined by current class levels. In individual sessions, students took the test and verbalized, explained and justified the strategies they used to complete each item.

Data Collection

We used two different data collection instruments to gather information about the validity of the ECAES English Exam: a content evaluation and think-aloud sessions. The purpose of these two instruments was to obtain information about the construct of the test, test use, teachers' and students' perceptions about the test, and test impact; in addition, these instruments documented the strategies test takers use to complete the test.

Content Evaluation Sessions. We asked 15 university English teachers to conduct a content evaluation of the ECAES English Exam. A content evaluation refers to "the judgments of experts concerning the ability that test items measure" (Chapelle, 1994, p. 168). The content evaluation sessions for this study were conducted individually and lasted approximately 2 hours each. All the participants were given the test, a form to complete (see Appendix A), and a list of CEFR level descriptors (see Appendix B).The purpose of the content evaluation was to ask EFL teachers to give their expert opinion about the skill(s) the items on the ECAES English Exam are measuring or the skill(s) the students have to use to complete each task, and to comment the items by highlighting their strengths and weaknesses. Teachers were also asked to align each item to the CEFR and the goals found within their particular university's curriculum. After the teachers completed this part, they were asked to comment the test items by providing their judgments on the stimulus, the directions, and the task expectations.

Think-aloud protocol. University students from different institutions in Bogota were asked to participate in a think-aloud session while they were completing the ECAES English Exam. The think-aloud session employed in this study broadly followed procedures similar to those proposed by Ericsson and Simon (1993). Think-aloud protocols are commonly used as a data collection method in problem solving, writing, and reading research (Stratman & Hamp-Lyons, 1994). The think-aloud protocol used in this study was piloted with two university students. This allowed us the opportunity to determine the quality of the probing questions. We also modeled the think-aloud protocol to the students by using tasks from another EFL test. After we modeled the think-aloud protocol, the students had an opportunity to practice the think-aloud protocol. This training session lasted approximately 30 minutes.

All the students who participated in the think-aloud session were asked to narrate in their language of preference anything that came to their mind while completing the ECAES English Exam. All the think-aloud sessions were conducted individually in a private office and were entirely video-recorded, so they could be transcribed and analyzed later. Immediately after completing each part of the test, we conducted retrospective interviews with the students to collect information about their perception of the different parts on the test and about how they felt completing them. Each think-aloud session and retrospective interview lasted approximately 100 minutes.

Data Analysis

The analysis and interpretation of all the items on the test provided evidence for the validity of the ECAES English Exam. The authors transcribed the think-aloud sessions and typed all the content evaluations forms. Then, we sorted and organized all the materials collected. Individually, we read all the data several times, then proceeded to use open coding to analyze the data (Miles & Huberman, 1994). Categories emerged directly from the data as repeated ideas or themes. This technique allowed us to use researcher triangulation to validate our coding and interpretations. The categories that emerged from the data were: evidence based on the construct of the exam, evidence based on the interactiveness of the test, evidence based on the authenticity of the test and evidence based on the impact of the test. Below we present information on each of these categories.

FINDINGS AND DISCUSSION

In this section, we present a validity argument in favor and against the ECAES English Exam. According to Chapelle, a "validity argument should present and integrate evidence and rationales from which a validity conclusion can be drawn pertaining to particular score-based inferences and uses of a test" (1999, p. 263). We frame our argument on evidence based on the construct, interactiveness, authenticity and impact of the ECAES English Exam.

Evidence Based on the Construct of the ECAES English Exam

According to the content analysis, test reviewers found that the ECAES English Exam only assesses some reading, grammar, and reading skills. Also, as it was stated in the description of the ECAES English Exam, other skills, such as listening, writing, and speaking are not assessed on this test. From the content evaluation sessions we found that, in general, teachers feel that the ECAES English Exam is not adequately aligned to the CEFR level descriptors. In Table 1, we present the perceived alignment between each part of the exam and the CEFR based on the number of reviewers who chose each answer in the scale. This finding suggests that the ECAES English Exam is partially aligned to the CEFR level descriptors.

Table 1. Alignment between the Exam and the CEFR level descriptors.

Part	Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned at All
1	0	0	11	4	0
2	0	0	9	6	0
3	0	0	6	8	1
4	0	0	13	2	0
5	0	6	7	2	0
6	0	10	5	0	0
7	0	0	13	2	0

Source: Teachers responses in the Content Evaluation form

Reviewers also revealed that the content of the test is not fully aligned to the content of their foreign language programs as they state that their programs are integrated and communicative in nature. From the content evaluation sessions we found that, in general, teachers feel that the ECAES English Exam is not adequately aligned to their university's English program. In Table 2, we present the perceived alignment between each part of the exam and their English programs based on the number of reviewers who chose each answer in the scale. This lack of alignment leads us to conclude that there is construct underrepresentation in the test.

Table 2. Alignment between the Exam and university English programs.

Part	Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned at All
1	0	0	0	2	13
2	0	0	1	2	12
3	0	0	2	3	10
4	0	8	4	2	1
5	7	7	1	0	0
6	8	6	1	0	0
7	2	7	4	2	0

Source: Teachers responses in the Content Evaluation form

As the test construct of the ECAES English Exam has not been clearly defined by the test developer, there is a resulting problem of construct validity due to construct underrepresentation. From what we have read from *Colombia Bilingüe*, we infer that this test assesses general English language ability. If this is the case, data from the content review suggests that there are relevant aspects missing in the focal construct (i.e. listening, speaking and writing). As a result, the test construct is not well-operationalized, nor is it sufficiently represented. As construct underrepresentation is one of the biggest threats to test validity (Messick, 1989), this creates one of the strongest arguments against this exam's validity argument from the perspective of language testing scholars.

Moreover, we found in the think-aloud sessions that there is construct-irrelevant variance in many of the items on the test, the other strongest argument against an exam's construct validity. Some of these extraneous variables in the tasks make some parts of the ECAES English Exam too difficult or too easy for some test-takers. For instance, many students in the think alouds used the tactic of word association to identify the correct response in Part 1, even if they did not understand the content of the text they were reading. Hence, for some students, the task becomes easier than it is supposed to be. Conversely, Part 2 requires students to match definitions with items from one lexical category. Since all items are from the same lexical category, students may require topical knowledge to complete this task. Thus, students with high English language proficiency may have difficulties with this task, depending on the lexical category presented and their familiarity with this category of words.

Evidence Based on the Interactiveness of the ECAES English Exam

Interactiveness of tasks considers which strategic competence, metacognitive strategies, and topical knowledge students use to complete the tasks on a test (Bachman & Palmer, 1996). In terms of strategic competence, we found in the think-aloud sessions a considerable amount of guessing, due to the selected response test format. This often resulted in students successfully selecting the correct response, especially if the guessing was based on applying test-taking strategies (e.g. eliminating options or finding key words). Even though guessing was present throughout the entire test, it was more evident in Parts 1, 2, 3, 4, and 7. Guessing is a major threat to the validity of this test and it represents a problem because individual item scores are taken as measures of how well the students know or understand the construct that is being measured (Fortus, Coriat & Fund, 1998). Likewise, we can consider this guessing as construct-irrelevant variance (Messick, 1989) in the sense that there are extraneous variables making the item easier than it is supposed to be.

Furthermore, we found that some tasks on the test require students to use other language skills to give a response. For instance, Part 1 is supposed to assess reading skills, while in reality it assesses lexical knowledge. In the think-aloud sessions, all the students simply matched key words on the sign with a key word in the options. If the sign said, "Please do not feed the animals," the students consistently focused on the word "animals" and then match this word to the word "zoo" in one of the options. There were five instances in which students guessed a correct response even though they did not understand the message on the sign. This leads us to believe that Part 1 is more a vocabulary task than the reading task it is intended to be, which brings us back to the fact that the construct of the test is not clearly defined.

In terms of topical knowledge, we found that Part 2 of the test requires students to have topical knowledge in the lexical category that is being assessed (e.g. things in a house, body parts, etc.). In the think-aloud sessions, some students stated that the task was a somewhat difficult because it required for them to have knowledge about things in a kitchen. These students said that to do well on this part of the test, they had to be lucky to get a lexical field they were familiar with. Some teachers commented in the content review that the topical knowledge

might favor particular groups of students. The performance on this part of the test does not necessarily reflect vocabulary breadth and depth (Nation, 2001). Some of the teachers that participated in the content review also argued that some of the lexical fields on this part had nothing to do with the lexical categories they worked in class.

In the think-aloud sessions, we found that students use few metacognitive strategies to complete the test tasks. For instance in Part 7, students only used word association to guess the meaning of unknown words. We did not see any evidence that students used other strategies such as guessing meaning from context or using word analysis. Likewise, in the two reading comprehension tasks (Parts 5 and 6), students only used a few strategies such as focusing on detail, recognizing words, associating words, guessing, inferring, or rereading. We did not see any evidence the students used more interactive strategies such as knowledge of grammar, using background knowledge, verifying what is already known, integrating new knowledge with known knowledge, or anticipating.

From the think-aloud sessions, we conclude that the majority of the items, except for most items on Part 6, do not require students to use a wide range of English language skills. In fact, most of the items on the test require students to only use lexical knowledge; thus, the test is not highly interactive. In general, the items on the ECAES English Exam require students to use minimal language skills or knowledge. Hence, we conclude again that this test does not necessarily provide a valid and useful measure of general English language proficiency.

Evidence Based on the Authenticity of the ECAES English Exam

Authenticity describes the degree to which the test takers perform tasks that are similar to the tasks they will have to do in real-life (Bachman & Palmer, 1996). From the findings in the content review and the think-aloud sessions, we found that the tasks on the ECAES English Exam are not very authentic. In the content review, teachers commented that the test tasks are not similar to the tasks students do in real life. For example, one of the reviewers explained that his students are used to reading longer texts which include more complex lexical items and grammatical structures. Another reviewer reported that her students are required to use more complex cognitive strategies when they read texts, such

as comparing and contrasting, building an argument, or establishing cause and effect.

This test's validity is also challenged by a lack of authentic language use in the tasks. By language use task we mean "an activity that involves individuals in using language for the purpose of achieving a particular goal in a particular situation" (Bachman & Palmer, 1996, p. 44). Also, the target language use domains in the test are not similar to target language use domains in real-life. Target language use (TLU) are "a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to extrapolate" (Bachman & Palmer, 1996, p. 44). About these two points, teachers commented that the language that is used and the tasks accomplished do not reflect language use or tasks in real life. Thus, we conclude that there is little correspondence between test performance and non-test language use. We feel that the test developer needs to conduct a needs analysis in Colombian universities to identify TLU tasks and then design test tasks that are appropriate and relevant to a greater percentage of Colombian test takers.

Furthermore, all the text-based items (Parts 4-7) use semi-authentic texts. These texts are based on authentic texts, but have been modified to make the level of the test appropriate for all students. These texts have been simplified in terms of length, lexical items, and grammatical structures. In general, teachers in the content review felt that these types of texts do not reflect the texts they use in their programs or texts students will encounter outside their English courses (e.g. on the Internet). Some students also commented during the think-aloud sessions that the texts were extremely short, which made it easier for them to select responses even if the questions were considered to be difficult. The short length of the texts allowed students to read the texts several times and facilitated the reading process or made it easier for them to locate the required information in the texts. This reduces the amount of processing the students have to do to answer the text-based questions (Fortus, Coriat & Fund, 1998).

Positively, Parts 4 and 7—assessing grammar and vocabulary skills—are more authentic than other parts of the test. These parts are discourse-based tasks and do not assess syntactical and lexical knowledge and skills in isolation. However, they are not authentic in the sense that

the discourse is incomplete and require students to gap fill. Some teachers felt that this activity is very common in foreign language contexts, but they are not something their students would be required to do outside a test environment.

Evidence Based on the Impact of the ECAES English Exam

Bachman and Palmer describe impact as being one of six factors of exam usefulness and write that "the impact of test use operates at two levels: a micro level, in terms of the individuals who are affected by the particular test use, and a macro level, in terms of the educational system or society" (1996, pp. 29-30). They further describe the impact on test takers as relating to "how relevant and appropriate the test scores [are] to the decisions to be made" (pp. 146-147) and to "how relevant, complete, and meaningful the feedback [is] that is provided to test takers" (p. 146). Bachman and Palmer describe test impact on teachers in terms of "how consistent the purpose of the test [is] with the values and goals of the teachers and the instructional program"; in terms of "how consistent the areas of language ability to measured [are] with those in teaching materials"; and in terms of "how consistent the characteristics of the test and test-tasks [are] with the characteristics of the teaching and learning activities" (p. 147). Even though the ICFES English Exam is not an exam within an instructional program, the test does measure what students are learning in their English programs. Thus, it is relevant to examine the correlation between the courses and the exam.

In terms of "the individuals affected by test use" (Bachman & Palmer, 1996, p.29), students during the think-aloud sessions described their perceived sense of the ECAES English Exam's minimum impact. They reported that the scores on the test are not currently used for jobs, graduate studies, or other important decisions. Many of them reported not taking this test seriously because they are not held accountable. They also reported that they are not compelled to prepare for this test or to focus on the content of the test. Finally, the researchers found that the scoring system does not provide any sort of meaningful feedback to students on their performance on the test beyond a number score.

In terms of teachers, we also found that the test has minimal impact. In the content review, teachers reported that they are not teaching to this test or using test preparation materials. They also claim that their

programs are not currently aligned to the content of the test and do not feel compelled to narrow the content of their courses to mirror the test. We also learned that some teachers are not well-informed about the ECAES English Exam. Four of them reported that they were not even aware that a new version of the test was developed in 2007.

Finally, we also found little impact on institutional language programs. All the teachers reported that their university uses tests (e.g. TOEFL, IELTS, PET, MELICET or in-house exams) in their accountability systems to describe and monitor student language proficiency, or for decision making purposes, but these systems do not include the ECAES English Exam. Likewise, teachers reported that they do not have or plan to align their English programs to the ECAES English Exam.

CONCLUSIONS

We have presented different types of validity evidence in this study. Taken together, all these different types of evidence provide a cumulative case for or against the appropriateness of score interpretation and use of the ECAES English Exam. Below we summarize the positive and negative cases for the test's validity.

Among the positive evidence for the validity of the ECAES English Exam we found that this test has no major negative effects on students, teachers, or programs; teachers are not teaching to the test; and some lexical and grammatical tasks are discoursed-based. Among the negative evidence against the validity of the ECAES English Exam we found that the construct of the test fails to include relevant aspects of general English language ability (e.g. listening, speaking, and writing) and many test items do not require students to use a wide range of areas of English language knowledge We found that the test is not fully aligned to the CEFR and that the content of the test is not aligned to the content of English language programs in Colombia in the sense that many test tasks are not aligned to the tasks used in university language programs. Furthermore, test tasks do not reflect authentic language use outside the test context and the texts within the test have been modified. In term of the items, the test uses only selected response items, many of which do not frequently challenge students to use higher thinking skills, and many items do not allow students to fully demonstrate their English language ability.

In light of the above, we find more compelling evidence that builds a case against the validity of this test in its current form. This suggests that 1) general English language proficiency cannot be accurately judged from this test; 2) we cannot make responsible generalizations about the test takers' English language ability beyond the testing situation; and 3) we cannot make responsible predictions about the test takers' ability to use the English language in real-life situations. The central problem in the validation argument for the ECAES English Exam in its current form is that it is being used to describe a student's English language level based on the CEFR. We should be able to infer from the test scores that a student at Level A2 is able to do everything that is included in the CEFR A2 level descriptors. However, this test does not provide information about many of the descriptors in the CEFR, mainly descriptors about listening, speaking and writing. Thus, any inferences we make about the students' listening, speaking or writing abilities based on the scores on the ECAES English Exam are inappropriate. More broadly speaking, this exam has not transparently exposed or delineated the purposes of its use (Hawthorne, 1997). Intended target language domains should be clearly identified, as well as the purposes of the exam.

Because the *Colombia Bilingüe* policy is politically charged, several issues become important in terms of critical language testing. Should the ECAES English Exam have greater validity, this could help alleviate the political pressures surrounding the creation and implementation of both *Colombia Bilingüe* and its related exams. With greater validity—especially construct validity—, perhaps university stakeholders would be more inclined to integrate this exam into their university curricula. Furthermore, future studies evaluating testing concepts such as test validity and test usefulness of Colombian exams should continue to be explored, especially if these studies were to be collaboratively conducted between the Ministry of National Education and local universities. Not only would these studies guarantee the creation of the most useful language tests, but the multilateral efforts would help appease political tensions and foster the most pro-active, forward-thinking language learning environment for our students.

REFERENCES

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Barletta Manjarres, N. (2005). Washback of the foreign language test of the state examination in Colimbia: A case study. *Arizona Working Papers in Second Language Acquisition and Teaching*, 12, 1-20.

Barletta Manjarres, N. and May Carrascal, O. (2006). Washback of the ICFES exam: A case study of two schools in the Departamento del Atlantico. *Íkala: Revista de Cultura y Lenguaje*, 11, 235-261.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall.

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

Chalhoub-Deville, M. (2008). *The social and educational impact of standards-based assessment in the USA*. Paper presented at ALTE Third International Conference, Cambridge, UK.

Ericsson, K. A., and Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (2nd ed.)*. Cambridge, MA: MIT Press.

Figueras, N., North, B., Takala, S., Verhelst, N. & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual, *Language Testing, 22(3)*, 261-279.

Fortus, R., Coriat R. & Fund, S. (1998). Prediction of item difficulty in the English Subtest of Israel's Inter-university psychometric entrance test. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 61-87). Mahwah, NJ: Lawrence Erlbaum Associates.

Hawthorne, L. (1997). The political dimension of English language testing in Australia. *Language Testing*, 14(3), 248-260.

ICFES (2009). *Características y guías de ECAES.* Retrieved March 3, 2009, from http://www.icfes.gov.co/index.php?option=com_content&task=view&id=568&Itemid=1061

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

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher*, 29(2), 4-16.

López, A. A. (2008). *Potential impact of language tests: Examining the alignment between testing and instruction*, Saarbrucken, Germany: VDM Publishing.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds.), *Test validity (pp. 33-45)*. Hillsdale, NJ: Lawrence Erlbaum Associate.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). New York: ACE/MacMillan.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook (2nd ed).* Thousand Oaks, CA: Sage.

Ministerio de Educación Nacional (2005) Colombia Bilingüe. *Altablero (37).* Retrieved June 14, 2008, from http://www.mineducacion.gov.co/1621/article-97495.html

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.

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

Stratman, J. F., and Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89-111). Thousand Oaks, CA: Sage.

INFORMATION ABOUT AUTHORS

Alexis Augusto López Mendoza
Ph.D. in Education and an M.A. in TESL from the University of Illinois at Urbana-Champaign. Professor in the Centro de Investigación y Formación en Educación –CIFE– at Universidad de los Andes and Director of the Centro de Evaluación. Research interests: educational assessment, language testing and biliteracy.
Correo electrónico: allopez@uniandes.edu.co

Gerriet Janssen
M.A. in Applied Linguistics and TESOL from the University of California at Los Angeles. Professor in the Department of Languages and Sociocultural Studies at Universidad de los Andes. Research interests: Academic writing for EFL Spanish-speaking students, curriculum development, teacher training and language testing.
Correo electrónico: gjanssen@uniandes.edu.co

Fecha de recepción: 14-08-2010

Fecha de aceptación: 8-11-2010

APPENDIX A
CONTENT REVIEW FORM
(Abbreviated Sample Form)

Reviewer: _____

Alignment to Common European Framework of Reference

In your opinion, which CEFR level descriptors (choose up to two) are being assessed in each part of the ECAES English Exam?

PART 1

Item #	CEFR Level Descriptor 1	CEFR Level Descriptor 2	Comments
1			
2			
3			

In your opinion, how well is Part 1 of the ECAES English Exam aligned with the CEFR?

Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned At All
O	O	O	O	O

In your opinion, how well is Part 1 of the ECAES English Exam aligned to your university's English program?

Fully Aligned	Adequately Aligned	Somewhat Aligned	Minimally Aligned	Not Aligned At All
O	O	O	O	O

In your opinion, which constructs (e.g. skills, abilities, concepts, knowledge) are being assessed in each part of the ECAES English Exam?

PART 1

Item #	Reading	Vocabulary	Grammar	Other
1				
2				
3				

In your opinion, what content is **NOT** assessed in the ECAES English exam that **SHOULD BE** assessed?

APPENDIX B
CEFR LEVEL DESCRIPTORS – READING
(Abbreviated List)

A1
Can understand familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.
I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.
Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.
Can understand short, simple messages on postcards.
Can recognize familiar names, words and very basic phrases on simple notices in the most common everyday situations.
Can get an idea of the content of simpler informational material and short simple descriptions, especially if there is visual support.
Can follow short, simple written directions (e.g. to go from X to Y).
A2
Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment).
I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.
Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
Can understand short simple personal letters.
Can understand basic types of standard routine letters and faxes (enquiries, orders, letters of confirmation etc.) on familiar topics.
Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables.