UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Colombian Applied Linguistics Journal
http://revistas.udistrital.edu.co/ojs/index.php/calj/
DOI: http://dx.doi.org/10.14483/udistrital.jour.calj.2014.2.a03

Research Article

Classical test theory and item response theory: Two understandings of one high-stakes performance exam[1]

Teoría clásica de la evaluación y teoría de respuesta al ítem: dos comprensiones de un examen avanzado de proficiencia

Gerriet Janssen, M.A.[2], Valerie Meier, M.A.[3], and Jonathan Trace, M.A.[4]

Citation / Para citar este artículo: Janssen, G., Meier, V., & Trace, J. (2014). Classical Test Theory and Item Response Theory: Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 167-184.

Received: 15-Nov-2013 / Accepted: 15-Jun-2014

Abstract

Language testing professionals and teacher educators have articulated the need for a broad variety stakeholders—including classroom teachers—to develop assessment literacy. In this paper, we argue that when teachers are involved in local assessment development projects, they can expand their assessment knowledge and skills beyond what is necessary for conducting principled classroom assessments. We further claim that a particular analytic approach, Rasch analysis, should be considered as one possible element of this expanded assessment literacy. To this end, we use placement exam data from one Colombian university to illustrate how analyses from item response theory perspectives (Rasch analysis) differ from, and can usefully complement classical test theory.

Keywords: assessment literacy, classical test theory, item response theory, language testing, Rasch analysis

Resumen

Evaluadores de lengua y formadores de maestros argumentan que los involucrados en el campo de la educación, incluyendo los maestros de aula, deben desarrollar un conocimiento profundo en el tema de la evaluación. Planteamos que los profesores, a la hora de estar involucrados en el desarrollo de proyectos de evaluación, puedan expandir sus conocimientos y habilidades para ir más allá de la evaluación tradicional del aula. Para alcanzar este fin, proponemos que la herramienta de análisis Rasch sea considerada como una parte de esta expansión de conocimiento. En este ensayo, a través de los datos obtenidos de un examen de clasificación de lengua aplicado en un contexto universitario colombiano, ilustramos cómo el análisis Rasch puede complementar la teoría clásica de la evaluación.

Palabras clave: evaluación de alfabetización, teoría de evaluación clásica, teoría de respuesta al ítem, evaluación de lenguas, análisis Rasch.

1 The research presented in this paper was supported by funding from Universidad de Los Andes, Departamento de Lenguajes y Estudios Socioculturales, and the University of Hawai'i, Mānoa, Graduate Student Organization.

2 Universidad de los Andes, Bogotá, Colombia and University of Hawai'i, Mānoa, Honolulu, United States. gjanssen@hawaii.edu
3 University of California, Santa Barbara, United States. valmeier@gmail.com
4 University of Hawai'i, Mānoa, Honolulu, United States. jtrace@hawaii.edu

Introduction

As high-stakes tests, including language tests, become ever more ubiquitous and influential, language assessment professionals have articulated the need for a broad variety of stakeholders to develop assessment literacy. Taylor (2009) describes assessment literacy as "an appropriate balance of technical know-how, practical skills, theoretical knowledge, and understanding of principles … all firmly contextualized within a sound understanding of the role and function of assessment within education and society" (p. 27). Teacher educators have recognized that appropriate assessment practices are integral to teaching and learning, even though these practices are often inadequately employed; this has prompted these educators to argue that developing assessment literacy be a central goal of pre-service teacher education and professional development (Popham, 2009). Popham (2009) argues that teachers' assessment literacy must encompass the skills and knowledge necessary to make defensible decisions about (a) the high-stakes tests increasingly (mis)used on behalf of standards-based accountability movements, as well as (b) classroom-based assessment that can be used to enhance teaching and learning. Regarding assessment literacy, Colombian scholars López Mendoza and Bernal Arandia (2009) have suggested that in order to support language learning, language teachers need to develop the competencies necessary to "develop, use, score, and interpret" classroom assessments.

While teachers undoubtedly need to understand how their classroom can be impacted by the summative uses of high-stakes testing and the formative uses of classroom assessments, some or perhaps many teachers find themselves involved in contexts that require an even greater range of knowledge and skills. This is particularly the case when teachers become involved in developing program-level, norm-referenced tests, such as placement exams. The construction, principled use, and systematic evaluation of such tests often require a more sophisticated set of conceptual and empirical tools than what is typically needed when planning and implementing classroom assessments; the responsibility for such tests often rests with local program insiders, including classroom teachers, rather than external testing experts.

We argue that involvement in local test development projects, particularly when they are a part of internally-motivated accountability efforts, can be an excellent catalyst for developing assessment literacy in terms of the knowledge Taylor (2009) described above: "technical know-how, practical skills, theoretical knowledge, and understanding of principles." For example, in the *15*(1) issue of CALJ, Janssen and Meier (2013) described how local stakeholders made gains in all of these areas when they participated in an "iterative, self-reflective, test development process [that] provide[d] opportunities for professional development and deeper engagement in accountability projects" (p. 100). Since most test development processes for placement exams employed at the institutional level are necessarily iterative in that these processes typically consist of phases of trialing and operationalization (Bachman & Palmer, 2010, pp. 144–145; Kane, 2013), the multiple iterations of test development and analysis provided local stakeholders with repeated opportunities to make gains in their assessment literacy.

In this paper, we would like to propose that projects concerning placement tests specifically provide language teachers the opportunity to further extend their assessment literacy because the high-stakes nature of placement tests requires items that both "fit" and "function" with the intended test taker population (see the section on CTT below)[5]. We propose here that

5 Special attention is also required by assessment developers when establishing a validity argument for test score interpretations, a topic that is far outside the bounds of this paper's scope. This critically important topic can be explored in canonical works such as Bachman and Palmer (2010), Chapelle (2012) or Kane (2006,2013).

item response theory analyses complements the basic CTT techniques presented in Janssen and Meier (2013): descriptive statistics, estimates of reliability, and other measures of classical test theory. Item response theory provides powerful analytical tools that, even in their most basic applications, can be a valuable option in the analysis of local, high-stakes tests. To this end, the present paper seeks (a) to provide readers with an introduction to test analyses from item response theory perspectives and (b) to answer the following research questions: How can basic item response theory analyses be used to evaluate the performance of a norm-referenced placement test, and how do the results of such an analysis compare with those of classical test theory on one Colombian high-stakes placement test?

We begin this paper with a brief overview of the theories behind classical test theory (CTT) and item response theory (IRT) analyses and then address our research questions by using data from one Colombian high-stakes placement exam. Hopefully, we will successfully transmit our enthusiasm for this approach with language teachers, administrators, and others involved in local test development and program evaluation efforts, so that they will be encouraged to apply item response theory analyses to their own projects and enhance their assessment literacy.

Literature Review: Classical Test Theory (CTT) and Item Response Theory (IRT)

CTT and its Use in Test Analysis

As the name would imply, Classical Test Theory (CTT) is one traditional way of understanding test scores. CTT is thought to be classical in that it is "well-established, having resisted the erosion of time" (Muñiz, 2003, p. 192), a quantitative approach that had its start in the early 20th century; still today, CTT's principles are "alive and well" in language assessment (Brown, 2013, p. 2)[6]. A central CTT concept considers a test measurement's *reliability*: measurements taken today should be nearly equivalent to one taken tomorrow, and there should be little *variance* or *error* in the scores. More specifically, CTT posits that underlying any *observed score* on a test is the test taker's *true score*. This true score would be very close a test taker's average score if he or she could hypothetically take the same exam a very large number of times (obviously discounting any practice effects). The *true score variance* would be the variation in these true scores, which would happen even though true scores are conceived of as being free from measurement error. Each observed score has its own variance (*observed score variance*), which is a cumulative result of problems in the environment, exam administration, scoring, poor test items, or examinee-related factors (Brown, 2013, p. 4). The difference between the true score variance (i.e., how much the scores might vary when free of measurement error) and the observed score variance is called the *error variance*. This relationship gives us equation 1). below, the cornerstone of CTT.

Observed Score Variance =
True Score Variance + Error Variance.

Given this basic relationship, CTT moves forward to focus on a variety of reliability measures that are available to language testers for assessing the consistency of their assessment instruments (cf., Cronbach's alpha, KR20, KR21, split-half reliability). These different reliability measures have been described in encyclopedia entries (cf. Brown, 2013, pp. 3–19; Sawaki, 2013), being elegantly summarized in Brown's Table 1

6 Today, CTT is considered to be one specific case within a larger framework called Generalizability Theory. Interested readers looking for an introduction to G-theory should refer to Bachman (2004, pp. 176–190), Ferlazzo (2003), or Marcoulides and Ing (2013); canonical treatments of this topic include Brennan (2001) and Shavelson and Webb (1991).

(2013, pp. 19). Further, in-depth coverage of these topics is offered in several canonical books on testing (cf. Bachman, 2004, pp. 153–170; Brown, 2005, pp. 177–181; Crocker & Algina, 1986, pp. 105–152).

Grounded in this understanding of reliability, CTT also provides measures for the analysis of the individual test items. Two basic measures of item analysis are item facility (IF) and item discrimination (ID). IF—also called *item difficulty* and labeled as *p* (see Crocker & Algina, 1986, p. 311)—is the measure of the percentage of students who answered a test item correctly: how easy the item was for the specific test population. In norm-referenced tests such as most placement tests, IF values should fall within the range of .30 (relatively difficult; 30% of the test takers answered the item correctly) to .70 (relatively easy; 70% of the test takers answered the item correctly) (Brown, 2005); the mean IF value should be approximately .50, so as to maximize distribution of the test takers into different classifications. The IF statistic can be said to describe the degree to which a test "fits" the population it is being used with; IF values that widely differ from those suggested above would be evidence of a test not "fitting" the local population.

The other test item statistic, ID, is a measurement of the degree to which an item separates the more proficient test-takers from the less proficient test takers; ID values are a proxy for the degree to which a test item is "functioning." Ideally, proficient test takers will answer an item correctly while less proficient test-takers will not, which means that the assessment is functioning well. The ID statistic is calculated by subtracting its IF value for a predetermined percentage of the lowest performing test takers from the IF value of the same percentage of the top performing test takers (Bachman, 2004, p. 125; Brown, 2005, pp. 68–71). Crocker and Algina (1986) present different percentages for calculating ID values; we follow Brown (2005) and use the top third and lower third. Ebel (1979) has presented a set of guidelines for interpreting ID values; Ebel's guidelines are thought to be field standards and are reported in Brown (2005) as well as in Crocker and Algina (1986). ID values have a possible range between -1.0 and +1.0, with +1.0 representing that the top percentage of test takers always answer the item correctly while the bottom third always answer the same item incorrectly. In an opposite fashion, a test item with an ID value of -1.0 would have the lower percentage of test takers always answering the item correctly, while the most able percentage of test takers always do NOT answer the same item correctly—a test item that truly is NOT functioning well for classification purposes! Ebel suggests that items with ID values of .40 and higher are considered excellent; .30–.39 are considered to be reasonably good, but potentially requiring modification; .20–.29 are considered marginal, with substantial revision being needed; and .19 and below are considered poor, and should be rejected or reworked.

CTT in One Previous Study

One study from the Colombian context that considers the use of CTT tools in the evaluation of an assessment instrument is Janssen and Meier's (2013). This article's appendix presents a selection of specific IF statistics for the exam these authors studied (p. 112). One can calculate for the items that this appendix displays that the vocabulary (VO) and grammar (GR) items did not fit the test taker population well, as IF values ranged from 0.69–0.96, with the average IF value for these two sections being 0.81, notably above Brown's recommended ranges. However, the reading comprehension (RC) questions fit the population much better, with IF values ranging from 0.42–0.72 (with the exception of one outlying IF value of 0.23) and the average IF value for this section being 0.55.

The appendix in Janssen and Meier (2013, p. 112) also presents a selection of specific ID statistics for the exam they studied. One can

calculate for the items that this appendix displays that the vocabulary (VO) and grammar (GR) items generally function adequately well with this test taker population in terms of separating proficient from non-proficient test-takers, as the ID values ranged from 0.05–0.55, with the average IF value for these two sections being 0.35. Here, it is worth noticing that the ID values for three items were quite low, which had the effect of lowering the average ID values to this still acceptably good value. The reading comprehension (RC) questions functioned much better with this specific population; indeed, the ID values for the RC items ranged from 0.31–0.79 (with the exception of one outlying IF value of -0.17) and the average ID value for this section was 0.51. Janssen and Meier (2013) recommended that these four outlying items described above be considered for omission from the test item pool.

CTT: Limitations

Despite its usefulness, CTT has several important limitations that have led researchers to look for complementary approaches. Bachman (2004) describes five shortcomings of CTT, but here we are primarily concerned with one: item analysis from CTT perspectives "is essentially sample-based descriptive statistics" (Bachman, 2004, p. 139). This means that, for example, IF and ID values are only representative of the specific sample of examinees from which they were calculated, so that making generalizations across different groups of examinees—or across different test formats—may not be possible. Because of its dependence on a specific sample, it is difficult for CTT to handle the more complex assessment situations that occur with great regularity, such as measuring test taker performance at different points in time (pre/ post); using different test forms which contain different items of different difficulty; or having raters assign scores to different elements of a performance exam. Still, CTT successfully completes the essential task of basic item analysis in a test development protocol for a homogenous population: it "determine[s] flaws in test items … evaluate[s] the effectiveness of distracters … and determine[s] item statistics for use in subsequent test development work" (Hambleton & Dirir, 2003, p. 189). Though CTT provides a variety of easy-to-use tools which can be applied for a basic description of how a specific sample of test takers performed on a specific test, more complex analytic approaches are required for many language assessment situations. IRT-based analyses are one family of such analytical tools.

Item Response Theory (IRT) and Its Use in Test Analysis

At its core, Item Response Theory (IRT) addresses CTT's limitation of using descriptive units that are not comparable between different assessments or between different points within the same assessment. To examine this last thought, consider what a 0.10 difference in IF values on one assessment represents: all the scientist knows from comparing items with IF values of 0.45 and 0.55 is that 10% more test takers completed the second item correctly than the first, which is also the case for the items with IF values 0.10 and 0.20. What is not known is the relative difficulties of the items: It cannot be said that the first item is 10% more difficult than the second item. IRT analyses, however, do give us a way to exactly quantify the differences between item difficulties and even between test taker performances.

To quantify the differences between two item difficulties (or two test taker performances) IRT uses as its metric a *derived measure*, a measure comprised of two fundamental measurements. A derived measure that all readers should be somewhat familiar with is the concept of *density*, the combination of the fundamental measurements *mass* and *volume*. In a similar way, IRT analyses use a derived measure based on the probability that a test taker will correctly

answer an item of a certain level of difficulty. This derived metric is what makes the family of IRT models so powerful, and it also allows for the inclusion of many different relevant *facets* of the testing situation into the statistical model. Among other things, facets such as item difficulty, prompt difficulty, rubric category difficulty, test taker ability, or rater severity can be included in one model and can be directly compared using a single unit called a *logit*.

In this paper we will address one of the simpler forms of IRT modeling, which can be easily used with tests that produce dichotomous data (e.g., multiple choice items). This basic analysis calculates the probability for a correct response based on the relationship between an item's difficulty and a test taker's ability (Bond & Fox, 2007). In this model, test takers have a 50% chance of answering an item correctly when both their ability level and the difficulty level of the item are equal. When changes occur in either item difficulty or person ability, the probability shifts accordingly (i.e., less person ability or more item difficulty will lead to a lower chance of success, and vice-versa). Based on these probabilities, item difficulty and person ability measures are calculated as logits and arranged along a true interval scale.

Using a probability model based on an interval scale makes it possible to understand how items perform independently of a specific sample, and one major benefit of IRT analyses is the ability to generalize findings about data that is considered to be unidimensional (Ellis & Ross, 2014, p. 1270). For example, consider a test for which some of the items were working but others required revision. Using only CTT, we could identify problematic items, revise these items, and re-administer the test, but unless the new sample of test takers is nearly identical to the old one, we are likely to end up with different item statistics for those items that were not revised, making comparisons difficult if not impossible. Using IRT analyses—when data is thought to be unidimensional—items within one exam administration or across multiple administrations, which may include items that are different across the two tests, can be directly compared. Using IRT, it is possible to exactly quantify the difference in item difficulties in logit units. We can predict, for instance, how revised items might have performed for a sample of test takers who did not encounter them, as well as how items shared between the two groups performed in relation to the combined samples[7].

Because it constructs a probability-based model, IRT analyses require unidimensionality, as has been alluded to in the above paragraphs. What this means is that all of the items on a test or test section should measure a common factor, construct, or latent trait (e.g., reading comprehension, pragmatic competence) in order for assumptions about difficulty and ability to be justified. More complex IRT models, well beyond the scope of this paper (see van der Linden and Hambleton, (1997) or Ostini and Nering (2006) for elaborate descriptions of these more complex models), permit measuring different latent traits simultaneously; nevertheless, each of these constructs of its own right should be unidimensional. This is to say that different constructs on an assessment instrument can each be unidimensional, even if the instrument includes multiple constructs (Henning, Hudson, & Turner, 1985; Wright & Linacre, 1989)[8]. As a final comment on

7 This can be done within a single analysis, or spread across multiple analyses, though comparisons across more than one analysis need to incorporate some sort of standardization step in the form of anchor items. Anchor items are common items to both analyses that have a fixed difficulty value determined in advance to define or "anchor" the scale for comparisons.

8 This comes with one major caveat, however, as the requirement for unidimensionality can still force limitations upon our interpretations of a test. Items that do appear to be testing an alternate construct from the majority of the test can incorrectly be labeled as misfitting the model when they are in fact functioning appropriately by CTT terms. In this case, the model is unable to account for the variance introduced by these items and so misidentifies them as problematic. While the goal is to fit the data to the model, the oversensitivity of unidimensionality can lead to a test that is ultimately too narrow and restricted to be of practical and authentic use.

unidimensionality, it is important to highlight that without unidimensionality for the latent trait(s) or construct(s) being measured, the probability model will fail. This is because without unidimensionality for each latent trait, we cannot meaningfully order ability and difficulty along the same scale. This is reflected in a high degree of misfit within the model, which is measured principally by a statistic called *infit mean square*.

Methodology

So that readers would be able to easily compare the CTT and IRT approaches, we reanalyzed the data presented in Janssen and Meier ([2013) according to both CTT and IRT approaches. The university in question generously permitted us to study this data. To respect the test takers and protect their identities, only data that were released for research purposes were used in this study; all data were stripped of identifying information before their release and analysis by these researchers.

Participants

Scores were collected from two placement test administrations for which dichotomously scored item-level data (i.e., 0/ 1, incorrect/ correct) was available for individual test takers (n = 190). This is admittedly a convenience sample; however, as there was no noticeable difference in the total exams scores between this sub-group and the larger pool of PhD applicants who have taken this test, this sample was taken to adequately reflect the larger population of test takers at this university.

Test Instruments

The reading test comprised one section of a three-part placement exam given to incoming PhD students at one Colombian university[9]. The reading test is a computer-based, timed exam (70 minutes), consisting of 78 multiple-choice questions that target three language constructs: grammar, reading comprehension, and vocabulary. The reading comprehension questions are passage-based, while the grammar and vocabulary sections of the exam may or may not be contextualized within paragraph-length texts. The distribution of passage-based and independent items across test constructs is presented in Table 1, a modified version of Janssen and Meier's Table 2 (2013, p. 106).

Rasch Analysis

Rasch analysis—one of the simplest analyses in the IRT family—was conducted on the reading subtest data using *Winsteps* v3.70.0.1 (Linacre, 2010). Within the data set, 62 instances of missing values were found for 15 examinees across

Table 1. *Reading Test Structure*

Construct	Total parts	Passage-based	Independent	Number of Questions (approx.)
Grammar	6	4	2	25
Reading Comp.	3	3	0	25
Vocabulary	7	5	2	25

9 The speaking and writing sections require the use of multi-faceted Rasch modeling, which is a more complex elaboration of the basic Rasch model presented in this paper.

various items. As Rasch analysis can account for missing data in its model without any adjustment, these values were retained and coded with the value of "N" within the input file so that *Winsteps* could recognize these as missing data compared to valid responses. To illustrate what a Rasch analysis input file looks like, we have included a condensed version of our input in Appendix A; for readers interested in pursuing Rasch analysis, Linacre (2012) includes myriad examples of input files that can be adapted to the needs of particular testing situations.

Results

While even a basic Rasch analysis produces a vast amount of potentially useful information, in this section we concentrate on introducing three key results: summary statistics of person and item measures; individual person and item fit statistics; and the vertical ruler. These are not only central when understanding how well a test is performing but also are relatively easy to grasp for those new to Rasch analysis.

Initial Analysis

Summary statistics for the initial analysis are displayed in Table 2 for both test items and persons. Descriptive statistics—mean, standard deviation (*SD*), maximum (Max), and minimum (Min)—are displayed for each column, which, from left to right, show raw scores; logit measures produced by Rasch analysis; the standard error of these measures; and mean-square (MNSQ) and standardized *z*-score (ZSTD) values for the two types of fit statistics reported by the model, infit and outfit (described in more detail in the following paragraph). At the bottom of the table is the person-separation reliability estimate, which can be interpreted the same way as Cronbach alpha. This reliability estimate is appropriately high (.93), signifying that the Rasch measures in the model that quantify test taker ability are separating test takers of different abilities 93% of the time. Table 2 also indicates ways in which parts of the test are not functioning optimally.

Fit. The question of model fit is one of utmost importance; without good model fit, no other

Table 2. *Summary Fit Statistics for Reading Test with 78 Items*

				Infit		Outfit	
	Raw Score	**Measure**	**Model Error**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Persons							
Mean	60.50	1.86	0.37	0.98	0.10	0.95	0.10
SD	12.20	1.24	0.12	0.13	0.80	0.38	1.10
Max	76.00	4.31	0.75	1.53	4.40	2.40	4.00
Min	28.00	-0.74	0.26	0.71	-2.20	0.22	-2.00
Items							
Mean		0.00	0.23	1.00	0.00	0.96	-0.10
SD		1.19	0.07	0.14	1.40	0.50	1.70
Max		3.71	0.59	1.61	4.80	3.40	6.40
Min		-2.93	0.17	0.66	-4.60	0.37	-3.70

Person-separation reliability = .93

statistics produced by the Rasch model are worth considering, as the model is thought to not work. Misfitting persons or items are those that do not conform to expected response patterns based on the model that the Rasch analysis has produced. Instances of misfitting persons might arise when a less-proficient test taker answers very difficult items correctly (due, for example, to lucky guessing or cheating) or when a more-proficient test taker answers very easy questions incorrectly (due, for example, to carelessness). Instances of misfitting items might arise due to problems with quality (i.e., items that are poorly worded) or multidimensionality (i.e., items that develop a different theoretical construct). For both persons and items, misfit occurs when the observed response patterns vary from their expected patterns in such an erratic way that accurate predictions cannot be made—and the placement of the person or item within the model cannot be done accurately. Conversely, persons and items can also be identified as overfitting when they are too perfectly consistent, as Rasch models assume that there will naturally be some amount of variation. Overfitting items do not degrade measurement and so are typically retained, but misfitting items are more problematic. High numbers of misfitting persons (more than 5% of the sample) can suggest problems with the test as a whole; a small number of misfitting persons is acceptable, though additional sources of information about these test takers' abilities should be sought if the test is used as the basis for high-stakes decisions. For the purpose of analysis, severely misfitting items should be removed so that they do not distort the Rasch model; for the purposes of test development, the content of misfitting items should be carefully scrutinized to determine if revision or removal is appropriate.

Table 2 displays MNSQ and ZSTD values for both infit and outfit statistics. While both can used to estimate fit, here we discuss only infit statistics for reasons described in the previous paragraph[10]. Identifying misfit is accomplished in much the same way we might identify an outlier in other forms of analysis: we can look at the distance an item or person is from the mean MNSQ infit value and judge whether or not this distance represents an acceptable amount of variation or if it is an anomaly in the data. While there are no "hard-and-fast rules" for making these judgments (Bond & Fox, 2007, p. 242), a number of useful guidelines do exist. One rule of thumb is that MNSQ values (in this study, for individual persons and individual items) greater than 1.30 or less than 0.70 signal misfit and overfit, respectively (Bond & Fox, 2007). A second guideline is to look for MNSQ values of greater than two standard deviations away from the mean MNSQ in either direction, which is perhaps the more applicable guideline as it related directly to the distribution of the data (McNamara, 1996). Accordingly, in the current study, misfitting persons have infit MNSQs of greater than 1.24 and less than 0.72 (M = 0.98; SD = 0.13), and misfitting items have infit MNSQs of greater than 1.28 and less than 0.72 (M = 1.00, SD = 0.14). With these ranges of fit values in mind, one next observes the fit statistics for each individual item and test taker (see Tables 3 and 4, respectively), and one can make the determination of whether each item and test taker fits within the model. Following a similar methodology, ZSTD scores can also be used to interpret fit; however, they are sensitive to sample size and might be less reliable in certain instances (Linacre, 2012).

Table 3 displays a partial output of item fit statistics for the reading subtest, ordered according to descending infit MNSQ values. This table focuses on the items that fit less-well, and for the sake of brevity we have omitted the vast

10 Outfit mean squares are unweighted and sensitive to outliers, while infit mean squares are weighted towards typical observations and are more sensitive to unexpected "inlying" responses (Linacre, 2012). Bond and Fox (2007) suggest that outfit can be oversensitive to identify data as misfitting, and so infit statistics often provide a more accurate measure.

majority of the reading test's items, which fit the model well. The first column displays the item number, followed by the number of responses for this item in the data set (Count), the difficulty of each item measured in logits (Measure), the standard error of measurement (*SEM*), and infit statistics. Based on the criteria of misfitting items being those with MNSQ values more than two *SD*s from the mean (*M* = 1.00, *SD* = 0.14, from Table 2), three items have MNSQ values above 1.28, which indicates misfit (items 78, 69, and 10, with infit MNSQ values of 1.61, 1.39, and 1.37, respectively, shaded in grey in Table 3). According to the same criteria, one item has an MNSQ value below .72 and is overfitting (item 49, infit MNSQ = 0.66, also shaded in grey in Table 3). All other items appear to be functioning within the expectations of the model.

One next should consider the actions to be taken in light of the above data. As stated earlier, overfitting items can be safely retained. The misfitting infit MNSQ values are not so large as to suggest misfit that would degrade measurement (i.e., values above 2.0; see Linacre, 2012, p. 553, for more details), yet it can be worthwhile to omit such items and re-run the analysis to see what effect this has on key results. Additionally, though it is beyond the scope of this paper, the content of the misfitting items should be examined to identify the source of misfit and determine what revisions, if any, would be appropriate. It is interesting to note that CTT analyses also signaled similar results for these misfitting test items: the one test item flagged by IRT as being most widely misfitting (78), was found in CTT analyses to be markedly more difficult than the test section average difficulty (IF = 0.23 and 0.79 respectively); furthermore, the item discrimination value for this item was -0.17, indicating that less-proficient test-takers were slightly more able to answer this difficult question correctly than more-proficient test-takers.

Person fit statistics are displayed in a similar way in Table 4, again ordered by descending infit MNSQ values. Person ID occupies the first column, followed by the number of observed responses for that person (Count), test-taker ability measured in logits (Measure), standard error of

Table 3. Partial Output of Item Fit Statistics, 78 Items

				Infit	
Item	**Count**	**Measure**	**SEM**	**MNSQ**	**ZSTD**
78	181	3.71	0.21	1.61	4.80
69	187	1.69	0.17	1.39	4.60
10	190	-0.01	0.20	1.37	3.10
64	190	1.23	0.17	1.28	3.40
48	190	-0.58	0.23	1.28	1.80
… *the most properly fitting items would be listed here* …					
51	190	-0.14	0.21	0.83	-1.50
77	180	1.44	0.17	0.82	-2.40
66	189	0.24	0.19	0.81	-2.00
49	190	0.71	0.18	0.66	-4.60

Note. The misfitting and overfitting items have been shaded grey.

measurement (*SEM*), and infit statistics. Person fit can be interpreted in much the same way as was described for item fit, with values greater than two *SD*s above or below the mean signaling as misfit or overfit, respectively. Based on these criteria, there are six test takers with MNSQ values above 1.24 ($M = 0.98$; $SD = 0.13$, from Table 2) who can be identified as misfitting, and one test taker with a MNSQ value below 0.72 who can be identified as overfitting. Unlike items, which can be relatively easily removed from a test when there is evidence of misfit, persons cannot be so simply excluded from a test. As the number of misfitting persons is comparatively small—only six out of 190, or about 3%—there is little reason to be concerned with their effect on the analysis at this stage.

Person and Item Measures. While the spread of both persons and items were normally distributed, as is expected in a placement exam, the distribution of test taker abilities is not well matched by the distribution of item difficulties. A study of Table 2 reveals that the mean person ability is noticeably higher ($M = 1.86$) than that of item difficulty, which is set at 0.00 by default in the model. This indicates that the test as a whole was relatively easy for examinees. Had the test been well matched to the population, the mean estimate of person ability would have been closer to 0.00 (Bond & Fox, 2007). Moreover, while the maximum value for person ability is 4.31 logits, the most difficult item on the test is only 3.71 logits (again, see Table 2). This result indicates that there are no items in this test section appropriately matched to the students at the highest ability levels.

This discrepancy between person ability and item difficulty measures is perhaps better represented in the vertical ruler (Figure 1), a graphic visualization produced within Rasch analyses that presents the interval scale along which persons and items have been plotted according to their logit measure. If a test is well matched to the population, the range of test taker abilities will be complemented by items of

Table 4. Person Fit Statistics for 190 Examinees

				Infit	
ID	**Responses**	**Measure**	**SEM**	**MNSQ**	**ZSTD**
185	35	-0.27	0.26	1.53	4.4
81	58	1.34	0.29	1.32	2.1
117	42	0.29	0.26	1.31	2.7
123	38	-0.07	0.26	1.27	2.5
75	61	1.6	0.3	1.26	1.6
113	44	0.33	0.26	1.26	2.3
103	49	0.67	0.26	1.2	1.8
… *the most properly fitting persons would be listed here* …					
160	76	4.31	0.75	0.75	-0.2
7	75	3.85	0.62	0.75	-0.4
133	75	3.85	0.62	0.75	-0.4
17	74	3.52	0.54	0.71	-0.6

Note. The misfitting and overfitting persons have been shaded grey.

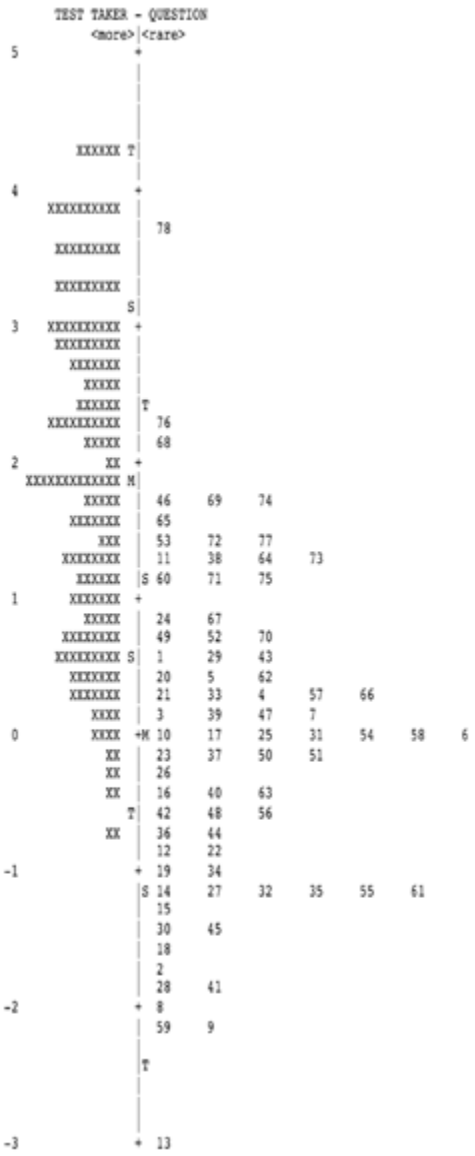

Figure 1. Vertical ruler of person ability and item difficulty for 78 items. Each X represents one person.

commensurate difficulty, such that test takers and items line up along the length the vertical ruler. In Figure 1, persons are shown on the left side of the axis by ability, with each "X" corresponding to one person, while items are arranged by difficulty along the right side by item number; the higher the position on the vertical ruler, the greater the test taker's ability or the item's difficulty. Again there is a clear mismatch between test taker ability and item difficulty, with the vast majority of test takers falling above the midpoint of the scale (0.00 logits) and items being spread more evenly around the mean. From Figure 1, it is easy to see that there is only one single item (78) above 2.50 logits, whereas there are a large number of persons with ability measures greater than 2.50. One can also see many items have difficulty measures below -1.00 logits, yet there are no

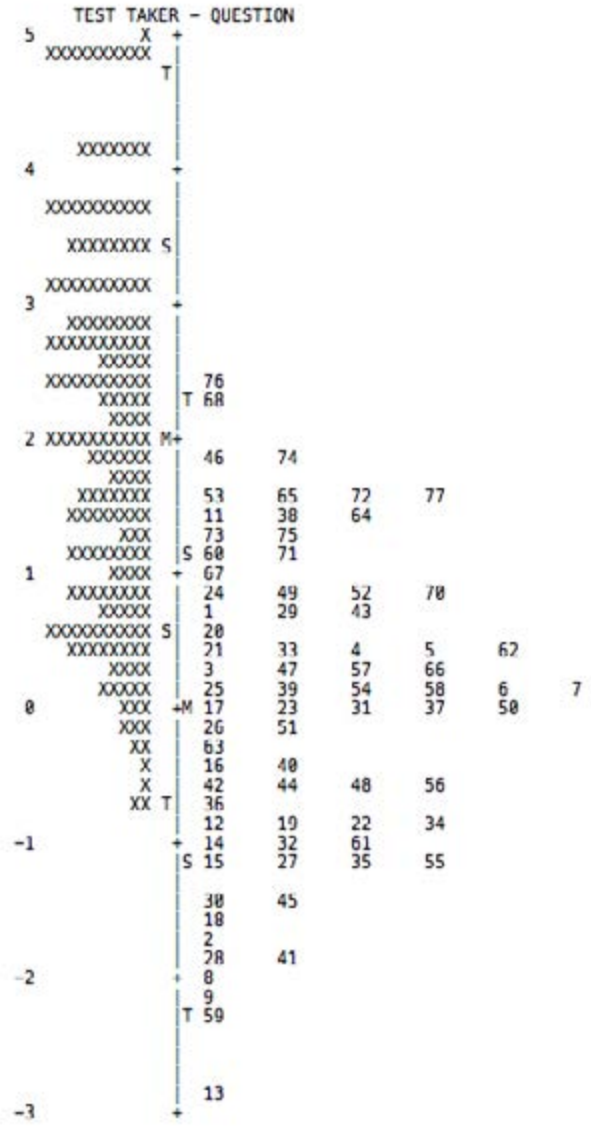

Figure 2. Vertical ruler of person ability and item difficulty for 75 items. Each X represents one person.

test takers whose ability measures are that low. The vertical rulers produced by Rasch analyses are one of its important benefits: a non-expert user can literally see at first blush the degree to which the test-taker ability levels match the items' difficulty. This evidences how an assessment instrument can be repurposed to include a variety of easier or more difficult items, depending on the test takers' abilities.

Follow-Up Analysis, Misfitting Items Removed

Based on these initial results, a second Rasch analysis was conducted with the three misfitting items removed ($k = 75$). This was done to confirm the degree to which the original analysis was affected by the misfitting items. Summary statistics for the revised model are shown in Table 5, along with a revised vertical ruler (Figure 2).

Removing the misfitting items from the initial analysis produced a slight increase in the person-separation reliability estimate of .94. Not surprisingly, though, removing misfitting items did not noticeably affect the basic mismatch between test taker ability and item difficulty. In fact, because one of the items removed (78) was the most difficult item in the initial analysis, there is an even larger discrepancy between the number of examinees with high ability and the number of

Table 5. Summary Fit Statistics for the Reading Test with 75 Items

				Infit		Outfit	
	Raw Score	**Measure**	**Model Error**	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**
Persons							
Mean	58.90	2.00	0.40	1.00	0.10	0.96	0.10
SD	12.10	1.37	0.19	0.11	0.80	0.45	0.90
Max	74.00	4.87	1.02	1.52	4.20	3.28	3.70
Min	27.00	-0.71	0.26	0.72	-2.60	0.19	-2.20
Items							
Mean		0.00	0.23	1.00	0.10	0.96	-0.10
SD		1.14	0.08	0.12	1.20	0.53	1.40
Max		2.37	0.59	1.33	3.90	4.46	5.00
Min		-2.88	0.17	0.66	-4.60	0.37	-3.20

Person-separation reliability = .94

Table 6. Item Fit Statistics for 75 Items

				Infit	
Item	**Count**	**Measure**	**SEM**	**MNSQ**	**ZSTD**
64	189	1.59	0.17	1.33	3.90
63	190	1.36	0.17	1.32	3.70
47	190	-0.52	0.24	1.32	2.00
67	189	2.24	0.17	1.27	3.20
69	186	1.21	0.18	1.16	1.90
… *the most properly fitting items would be listed here* …					
75	180	1.55	0.18	0.85	-1.90
50	190	-0.08	0.21	0.83	-1.50
60	189	-1.02	0.28	0.83	-0.90
65	189	0.31	0.20	0.82	-1.90
48	190	0.79	0.18	0.66	-4.60

Note. The misfitting and overfitting persons have been shaded grey.

items of appropriate difficulty. In terms of misfit and overfit, infit MNSQ statistics continue to indicate that there continue to be a small number of items that misfit, which can be seen in the detailed item fit statistics output (see Table 6). It is worth noting that these misfitting items in this new analysis were nearly misfitting in the original analysis (compare Tables 3 and 6). Thus we can conclude that improving the performance of this test will require more than just the revision of a few misfitting items; it will involve a systematic replacement of easy items with more difficult items that correspond to the ability levels of the PhD students applying to this university.

Discussion

We began this paper emphasizing our belief in—and many scholars' call for—promoting more assessment literacy by program teachers and other stakeholders, and it is only fitting that we conclude focusing on the same theme. More specifically, the "know-how, practical skills, theoretical knowledge, and understanding of principles … all firmly contextualized within a sound understanding of the role and function of assessment within education and society" that Taylor (2009, p. 27) calls for becomes vitally important when one considers the different uses that tests may have, and that test developers may be the only line of defense test takers may have in assuring that the tests being used to evaluate them are fair, appropriate, and valid. We hope we have highlighted the importance of two theoretical frameworks—CTT and IRT—that can be used to understand the quality of test items for the test taker populations being assessed, and we hope that program teachers begin to inform themselves from a variety of perspectives about the quality of the instruments they are designing and employing.

Furthermore, in this paper we revealed how basic IRT analysis could be used to evaluate the performance of a norm-referenced placement test, and how these results compared with those of classical test theory methodologies. The results of Rasch analysis presented in the previous section mirrored the most central findings reported by Janssen and Meier (2013) using CTT: that the current version of the reading portion of the placement exam is not well matched to the prospective PhD students whose academic English reading ability it is designed to measure, though the items generally function well. The results of both types of analysis additionally suggest that this test would benefit from the elimination of several items that are too easy and the inclusion of a greater number of more difficult items.

Janssen and Meier (2013) based their conclusions on measures of central tendency and dispersion, which indicated that the reading test section was broadly speaking too easy for the population sample; they also used IF values to identify specific items which were and were not of suitable difficulty (p. 109) and ID values to ascertain which items were and were not effectively differentiating between more and less able test takers (p. 109). The conclusions in this current analysis were reached based on the discrepancy between item difficulty measures and test taker ability measures which, unlike descriptive statistics and item analysis values based on raw scores, are plotted on a single interval scale and thus are directly comparable. Importantly, while logit measures and standard errors for each individual item or test taker can be reported in tabular form (e.g., Tables 3, 4, 6), Rasch analysis handily produces a vertical ruler that efficiently summarizes the relationship between item difficulty and test taker ability. In our experience, we have found that non-specialist but interested stakeholders such as program administrators can more intuitively grasp the implications of a vertical ruler describing test items and test takers than they can the import of a table of IF and ID values, especially since values for item difficulty values and test-taker ability are reported in integer units. Thus although learning

Rasch analysis requires a bit more initial effort, this effort is repaid when it comes time to sharing findings of a test analysis. This is one of the benefits of using IRT analyses.

Moreover, Rasch analysis provides additional information not available through CTT, such as fit statistics, which can flag test items and test takers that require additional review. Item fit statistics can alert test developers to problems with item quality, while person fit statistics can alert those responsible for making decisions about test takers when additional sources of information should be collected in order to make defensible inferences about test takers' abilities (i.e., for misfitting test takers). Thus, while we are not suggesting that IRT analyses supplant CTT, we suggest that even the basic output presented in this paper make important contributions to understanding and evaluating test performance.

Conclusion

While the Rasch analysis results further confirm CTT results, they also provide useful additional resources, including (a) a single graphic, the vertical ruler, which neatly captures the relationship between item difficulty and test taker ability and can be used to clearly and efficiently communicate these finding to other test stakeholders; and (b) and the identification of misfitting items. Moreover, while we did not use Rasch analysis to compare the performance of test items across different groups of examinees, in the literature review we suggested that this was a major advantage to the sample-independent nature of the IRT approach. In this particular instance, the misfitting items we identified could be revised and their performance analyzed across groups of different examinees provided two test forms were linked through a common set of anchor items. While this next step is beyond the scope of this paper, we hope this brief introduction to the possibilities of Rasch analysis has demonstrated the value of this analytic approach and perhaps inspired those involved in the development of local, high-stakes exams to extend their assessment literacy by delving more deeply into the topic.

References

Bachman, L. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press.

Bachman, L. & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*, (2nd ed.). New York, NY: Routledge.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Brown, J. D. (2013). Classical theory reliability. In A. Kunnan (Ed.), *Companion to language assessment*, Vol. 3. Hoboken, NJ: Wiley Blackwell.

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29(1)*, 19–27. doi:10.1177/0265532211417211.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*, (1st ed.). Belmont, CA: Wadsworth Group/ Thomson Learning.

Ebel, R. L. (1979). *Essentials of educational measurement*, 1st edition. Upper Saddle River, NJ: Prentice Hall.

Ellis, D., & Ross, S. (2014). Item response theory in language testing. In A. Kunnan (Ed.), *Companion to language assessment*, Vol. 3. Hoboken, NJ: Wiley Blackwell.

Ferlazzo, F. (2003). Generalizability theory. In Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 425–429). London, UK: Sage Publications.

Hambleton, R., & Dirir, M. (2003). Classical and modern item analysis. In Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 188–192). London, UK: Sage Publications.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2(2)*, 141–154.

Janssen, G., & Meier, V. (2013). Establishing placement test fit and performance: Serving local needs. *Colombian Applied Linguistics Journal, 15(1)*, 100–113.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education / Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1)*, 1–73. doi:10.1111/jedm.12000.

Linacre, J. M. (2010). *Winsteps* (Version 3.70.0.1). Chicago, IL: MESA Press.

Linacre, J. M. (2012). *A user's guide to Winsteps [software manual]*. Chicago, IL: MESA Press.

López Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *PROFILE, 11(2)*, 55–70.

Marcoulides, G., & Ing, M. (2013). The use of Generalizability Theory in language assessment. In A. Kunnan, (Ed.), *The companion to language assessment, Vol. 3* (pp. 1207–1223). New York, NY: John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla014

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

Muñiz, J. (2003). Classical test theory. In Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment*. London, UK: Sage Publications.

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48(1)*, 4-11. doi:10.1080/00405840802577593

Sawaki, Y. (2013). Classical test theory. In A. Kunnan (Ed.), *The companion to language assessment. Vol. 3*. Hoboken, NJ: Wiley Blackwell.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. London, UK: Sage.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21-36. doi:10.1017/S0267190509090003

Van der Linden, W., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

Wright, B. D., & Linacre, J. M. (1989). *Observations are always ordinal; Measurements, however must be interval (MESA Research Memorandum No. 44)*. MESA Psychometric Laboratory. Retrieved from: www.rasch.org/memo44.htm

Appendix A

Sample Winsteps Input File

```
Title= Reading test
Name1=1
Item1=10
NI=78
CODES=01
PERSON=TEST TAKER
ITEM=QUESTION
T1I#=1
T1P#=1
&END
1
2
3
4
5
 (continues enumerating all 78 items)
END NAMES
120505001111111111111111111111111111111111111111101111111111111111111111111111111
120505002111111111111111111111111111111111111111111111111111111111111111101111111
120505003111111111111111111111111111111111111111111101111111111111111111111111111
120505004111111111111111111111111111111111111111111111111111111111101111111111111
(continues providing one line of data for each of 190 test takers)
```