A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

2010, AER, 9, 010116-1, 10.3847/AER2010024

Do Concept Inventories Actually Measure Anything?

Colin S. Wallace
University of Colorado at Boulder, Boulder, Colorado, 80309
Janelle M. Bailey
University of Nevada, Las Vegas, Las Vegas, Nevada, 89154

Received: 08/19/10, Accepted: 09/17/10, Published: 12/3/10

Abstract

Although concept inventories are among the most frequently used tools in the physics and astronomy education communities, they are rarely evaluated using item response theory (IRT). When IRT models fit the data, they offer sample-independent estimates of item and person parameters. IRT may also provide a way to measure students' learning gains that circumvents some known issues with Hake's normalized gain. In this paper, we review the essentials of IRT while simultaneously applying it to the Star Properties Concept Inventory. We also use IRT to explore an important psychometrics debate that has received too little attention from physics and astronomy education researchers: What do we mean when we say we "measure" a mental process? This question leads us to use IRT to address the provocative question that constitutes the title of this paper: Do concept inventories actually measure anything?

1. INTRODUCTION

Researchers in the fields of astronomy education research (AER) and physics education research (PER) frequently use concept inventories to measure students' knowledge. A concept inventory is an approximately 20–25 question multiple-choice test designed to probe students' understandings of a single topic (Bailey 2009; Sadler *et al.* 2010), sometimes called a construct (Wilson 2005). The items and answer choices on a concept inventory are selected based on research into students' common reasoning difficulties (Bailey 2009). Current concept inventories cover a variety of constructs, including forces (Hestenes, Wells, and Swackhamer 1992), electricity and magnetism (Maloney *et al.* 2001; Ding *et al.* 2006), lunar phases (Lindell 2001), the greenhouse effect (Keller 2006), light and spectroscopy (Bardar *et al.* 2007), and star properties (Bailey 2007). See also Sadler *et al.*'s (2010) work on developing and validating a pool of 211 items for K-12 astronomy and space science concept inventories.

But do these concept inventories give us the information we need? Do they provide us with the information we think we are getting? Typically, item and test quality, as well as students' knowledge, are evaluated using sample-dependent statistics from an unfalsifiable model known as classical test theory (CTT). Students' learning gains are often determined by comparing their scores on a concept inventory they took before and after instruction. Yet such comparisons are not without problems. In Sec. 2, we highlight some issues surrounding traditional CTT analyses of and learning gains computed using concept inventories.

Item response theory (IRT) offers an alternate route around some of these issues. A few PER studies employ IRT (Ding and Beichner 2009; Lee *et al.* 2008; Marshall, Hagedorn, and O'Connor 2009; Pek and Poh 2000; Planinic 2006; Planinic, Ivanjek, and Susac 2010; Wang and Bao 2010) or IRT-inspired methods (Morris *et al.* 2006). IRT also has been used to analyze concept inventories in chemistry (Herrmann-Abell, DeBoer, and Roseman 2009), statistics (Allen 2007), and geology (Libarkin and Anderson 2005). Yet, despite their potential, IRT models are not commonly used by our community (but see Sadler 1998 for an exception).

In this paper, we undertake an IRT analysis of an astronomy concept inventory, the *Star Properties Concept Inventory v3* (SPCI; Bailey 2007). Our data consist of the matched pre- and post-instruction responses of 334 students who took introductory astronomy for nonscience majors (hereafter ASTRO 101) in Spring 2005 at a large university in the southwestern United States. We have three goals for this paper:

1) introduce IRT to members of the AER community who may be unfamiliar with its details and uses;
2) apply IRT to the SPCI to learn more about the SPCI and to exemplify how future studies may use IRT; and
3) explore what “measuring” a person’s trait, such as knowledge about star properties, truly entails.

This last point is the subject of a vigorous debate within the psychometrics community, one which raises some important questions for researchers attempting to measure students’ astronomical knowledge. Yet this debate has not received much attention within the AER and PER communities (although it is briefly touched upon by Ding and Beichner 2009). We highlight some of these questions and their implications for test development and interpretation near the end of this paper.

This paper is organized as follows. Section 2 elaborates on some of the problems with traditional concept test analyses and motivates the use of IRT. Section 3 is a general introduction to three common IRT models; readers already familiar with IRT may skip this section. Section 4 contains the results of our analysis of the SPCI. Section 5 discusses IRT calculated gains. In Secs. 6 and 7, we evaluate how well IRT models fit the SPCI data and whether or not the underlying assumptions of IRT hold, respectively. Section 8 is an argument for why one might choose one IRT model over another. Section 9 is a summary of the paper and our conclusions.

2. ISSUES WITH TRADITIONAL CONCEPT TEST ANALYSES

Why should one take the time to apply IRT to a concept inventory? Or, stated another way, what is lacking in our current approaches to concept inventories? In this section, we highlight issues with 1) using CTT to judge item and test quality and student achievement and 2) traditional calculations of learning gain.

2.1. Issues with CTT

CTT postulates that a student p’s observed score (X_p) differs from her true score (T_p) by a certain amount of error (E_p),

$$X_p = T_p + E_p \tag{1}$$

(Lord and Novick 1968). From this simple model, a number of elegant statistics are derived (Lord and Novick 1968). Some of the most important statistics include the following.

1) Estimates of reliability: These estimate how much of the variation in observed scores are due to variation in test takers’ true scores (Lord and Novick 1968; Thompson 2003). One of the most popular reliability estimates is Cronbach’s α. To maximize the test’s reliability, Cronbach’s α should be as close to one as possible (Borsboom 2005), although values greater than 0.70 are generally considered acceptable (George and Mallery 2009).
2) P-values: An item’s P-value is the fraction of examinees who correctly answer the item (Lord and Novick 1968; Crocker and Algina 1986). This is a common measure of an item’s difficulty in CTT. P-values should lie around 0.5 to maximize reliability (Ding and Beichner 2009), although some studies accept items with P-values as low as 0.10 or 0.20 and as high as 0.80 or 0.90 (Bardar *et al.* 2007; Ding *et al.* 2006; Maloney *et al.* 2001).
3) Point-biserials: A point-biserial is the correlation between an examinee’s score on a single, dichotomously scored item and her total score (Lord and Novick 1968; Crocker and Algina 1986). Point-biserials are frequently used to judge the discriminatory power of items (e.g., how well do individual items separate students who are high on the construct from students who are low on the construct). Convention suggests point-biserials should be greater than or equal to 0.20 (Ding and Beichner 2009), although researchers often set their own criteria (e.g., the developers of the *Light and Spectroscopy Concept Inventory* flagged items with point-biserials smaller than 0.30 and larger than 0.70; Bardar *et al.* 2007).

Concept inventory developers use these statistics to ensure the quality of their tests (e.g., Maloney *et al.* 2001, Ding *et al.* 2006, Bardar *et al.* 2007, and Sadler *et al.* 2010).

While these statistics do provide useful information, they are also highly sample-dependent (Hambleton and Jones 1993; Thompson 2003). For example, we computed Cronbach's α for the SPCI for two groups: We first used only students' pre-instruction responses, and then we used only students' post-instruction responses. For the pre-instruction group, α=0.45. For the post-instruction group, α=0.72. The variation in these values for Cronbach's α is likely due to the fact that it depends on the total score variance; it is thus sensitive to the homogeneity of the test-taker population (Thompson 2003). More heterogeneous groups commonly yield higher values of Cronbach's α, as our example demonstrates. The fact that Cronbach's α is low for the pre-instruction group is not necessarily an indictment of the SPCI. It merely reflects the fact that the pre-instruction group is very homogeneous in their (lack of) knowledge about star properties, which underscores our point about the sample-dependence of CTT statistics.

The same sample-dependence can be seen if we look at the *P*-values and point-biserials for each item on the SPCI. Table 1 demonstrates that one gets very different values for the item's *P*-values and point-biserials depending on the group one examines. These results are not surprising. We do not expect these statistics to be invariant. *P*-values, for instance, necessarily depend on what students know. Once students learn more about the construct being tested, the items' *P*-values must change. These data do, however, emphasize the sample-dependent nature of CTT statistics. These data also present a warning to researchers: One cannot claim that a concept inventory is reliable or that an item has adequate difficulty and discrimination by simply quoting the CTT statistics calculated from an earlier study on a different group of examinees. If one wants to use CTT to judge item and test quality, then one must recalculate the CTT statistics for each group one tests.

Table 1. CTT statistics for the SPCI, calculated using students' pre-instructional responses only, then using post-instructional responses only. Items are presented in order of ascending *P*-value within each group

Pre-instruction			Post-instruction		
Item	*P*-value	Point Biserial	Item	*P*-value	Point Biserial
22	0.05	0.10	2	0.17	0.17
5	0.10	0.14	13	0.17	−0.02
18	0.10	0.12	3	0.18	0.02
12	0.14	0.21	22	0.22	0.42
3	0.17	0.13	5	0.32	0.27
19	0.20	0.26	18	0.32	0.27
2	0.21	−0.08	8	0.36	0.23
15	0.23	0.16	12	0.40	0.45
20	0.23	0.14	19	0.42	0.32
10	0.24	0.14	9	0.49	0.13
13	0.28	0.09	15	0.51	0.41
17	0.29	0.16	20	0.51	0.28
23	0.30	0.07	6	0.60	0.23
11	0.31	0.19	23	0.60	0.22
8	0.34	0.06	10	0.62	0.32
9	0.36	0.17	11	0.64	0.34
7	0.37	0.17	21	0.67	0.38
21	0.39	0.08	16	0.70	0.15
6	0.39	0.07	1	0.73	0.29
1	0.42	0.20	17	0.74	0.42
16	0.56	0.16	7	0.75	0.30
4	0.63	−0.04	4	0.81	0.29
14	0.81	0.24	14	0.93	0.24

Another problem with CTT-based analyses is their focus on students' total scores. In CTT, one strives to reduce the error in measuring a student's true score so that her observed score is as close to her true score as possible. But what does this true score actually represent? It is defined as the expectation value of all the observed scores a student would have earned if she took the test multiple times under the same conditions (which means we have to brainwash her so that she cannot remember her previous answers to the test; see also Lord and Novick 1968 and Borsboom 2005). Even if we ignore the inherent fictitiousness of this thought experiment, we may still wonder what the true score actually measures. Borsboom (2005) points out that CTT proposes no relationship between a student's true score and the amount of the construct she possess (which we will henceforth call her *ability* in keeping with the nomenclature of IRT). A student's score is due to some unspecified (at least in CTT) combination of her ability and the properties of the items, such as how difficult they are. What does a score tell us about her ability *independent of the items*? Perhaps very little. A student may do well on a test because she has a high ability, because the test's items are easy, or because she has a high ability and the test's items are easy. If we only use CTT, then we may not actually measure students' abilities.

We note one final weakness with CTT: The model cannot be falsified (Lord 1980). One either accepts Eq. (1) or not. CTT offers no way to test whether or not it is a valid description of what happens when students take a test.

Of course, CTT is not worthless. It has several advantages (Hambleton and Jones 1993). Its statistics are easily computed, and they may be calculated with only a modest number of examinees. It can help detect poor items through relatively straightforward procedures. It is easy for a data set to meet the weak assumptions of CTT. Researchers must weigh these positives against CTT's negatives when they are deciding whether and how to use CTT.

That being said, IRT models, when they fit the data, offer several advantages over CTT. First, its parameters are sample-independent (Hambleton and Jones 1993; Whitely and Dawis 1974). This means that one can estimate item parameters independent of the population of examinees. Thus, judgments of item and test quality may still be made, even if the pilot test population is not representative of the population of interest. Students' abilities also can be estimated independent of the specific items they take. IRT can disentangle students' abilities from item properties to provide test-independent measures of ability. IRT models are also falsifiable (Embretson and Reise 2000). One can test whether or not a given model accurately describes how students respond to individual items. In Secs. 3, 4, 6, and 7 below, we highlight these advantages of IRT over CTT in the context of our analysis of the SPCI.

2.2. Issues with Learning Gains

Concept inventories are frequently used to measure gain—that is, how much a student or a group of students improves on a construct over time. One way to measure gain is to administer a concept inventory twice, once before and once after instruction, and then subtract a student's pre-instruction (pre-test) score X_0 from her post-instruction (post-test) score X_f. Yet differences between post- and pre-test scores may be problematic measures of gain. Bereiter (1963) noted that such gains decrease in reliability as the correlation between pre- and post-test scores increases, exhibit a spurious negative correlation with pre-test scores, and assume that a given difference $X_f - X_0$ (e.g., 15 points) implies the same difference in ability regardless of the student's initial score. Cronbach and Furby (1970) recommend avoiding gain calculations altogether, although Rogosa and Willett (1983) describe when gain scores can be reliable. In general, AER and PER studies rarely use $X_f - X_0$ as a measure of gain.

In AER and PER, a more common expression is Hake's (1998) formula for the normalized gain,

$$\langle g \rangle = \frac{X_f - X_0}{M - X_0}, \tag{2}$$

where M is the maximum number of points for the test. Gains calculated via Eq. (2) frequently are interpreted as the ratio of the amount by which a student improves on a test to the maximum amount by which she could have improved (Hake 1998; Prather *et al.* 2009).

Note that Hake's formula still relies on the difference between the post-test and pre-test scores. It also relies on the difference between the maximum possible score and the pre-test score. If one wishes to find the average gain for a group of students, one may either calculate and average $\langle g\rangle$ for every student or plug the group averages of X_f and X_0 into Eq. (2). Bao (2006) showed that one may get different answers depending on which procedure one uses. Most studies adopt the latter procedure (e.g., Hake 1998 and Prather *et al.* 2009). But before we worry about how to calculate normalized gains, we must address a more fundamental question: Are these mathematical and statistical operations even sensible to perform?

To better understand this question, we must look at Stevens's (1946) work on scale types. Stevens defines measurement as "the assignment of numerals to objects or events according to rules." The scale you are working with is determined by how numbers are assigned. If numbers are assigned merely as labels, then one has a nominal scale (e.g., group 1, group 2, etc.; Vogt 2007). If the assignment maintains some empirical ordering between the objects one is measuring, then one has an ordinal scale (e.g., class rank; Vogt 2007). If the numbers are assigned such that they preserve an empirical ordering *and* such that differences between numbers represent meaningful empirical differences between the objects, then one has an interval scale (e.g., calendar years; Vogt 2007). If the assignment is done such that one has all the properties of an interval scale and ratios between numbers are also meaningful, then one has a ratio scale (e.g., age; Vogt 2007). Stevens noted that many measurements in the physical sciences are on ratio scales, while the social and behavior science frequently have ordinal measures.

Stevens also notes that only certain mathematical and statistical operations are permitted for each scale type. For example, subtracting two numbers is only sensible if one has an interval or ratio scale. Imagine studying two stars: one, Population I; the other, Population III. The numbers are on a nominal scale since they are simply labeling two different groups of stars. Subtracting one number from the other yields a meaningless number. Likewise, taking the mean could lead us to conclude that the average star in our sample is Population II—except this statement is also meaningless (and wrong). While we can manipulate numbers all we want, some manipulations will not provide any sensible information, depending on the scale type.

This is relevant for gain calculations because Wright and Linacre (1989) argued that raw scores are typically ordinal. By looking at raw scores, we can readily order students by the number of questions each correctly answers. Such an ordering lets us rank students according to the amount of the construct each possesses. For example, a student who scores a perfect 23 points on the SPCI likely knows more about the properties of stars than a student who only answered 20 items correctly. However, we cannot say that a student needs the same increase in her knowledge of star properties to move from a score of 15 to a score of 18 points as she would to move from a score of 20 to 23 points. *Equal differences in raw scores do not necessarily correspond to equal intervals in students' abilities* (Bereiter 1963; Wright 1997). Thus, raw scores do not necessarily form an interval scale for students' abilities.

Yet Hake's formula requires that we subtract scores and, if we are finding a group's average gain, calculate means. These operations are only appropriate if differences between raw scores are meaningful—that is, we must have an interval scale (Stevens 1946). But raw scores are not manifestly interval measurements of ability (Wright and Linacre 1989). Planinic, Ivanjek, and Susac (2010) noted that "Hake's normalized gain… may also be influenced by the nonlinearity of raw scores expressed as percentages." Hake's normalized gain, being constructed from ordinal data, may be at most an ordinal measure of learning gain.

This ordinal nature is further supported when we consider what happens when $\langle g\rangle < 0$. While $\langle g\rangle$ has an upper bound of one, there is no corresponding lower bound (Marx and Cummings 2007). This means that the difference between two gains does not have the same meaning across the scale, as must be the case for interval and ratio scales. Marx and Cummings (2007) proposed redefining $\langle g\rangle$ when post-test scores are lower than pre-test scores, but this further calls into question the nature of $\langle g\rangle$'s scale.

Additionally, what are we to make of the fact that Hake's formula multiplies the difference $X_f - X_0$ by $(M - X_0)^{-1}$? At first, this seems to account for Bereiter's (1963) criticism that $X_f - X_0$ corresponds to different changes in ability for different X_0. Yet the multiplying factor $(M - X_0)^{-1}$ increases as the pre-test score increases. This means that given two students who exhibit the same $X_f - X_0$, a higher normalized gain score will be assigned to the student with the highest pre-test score (Brogt *et al.* 2007). Hake did not justify why an improvement of, say, ten points corresponds to a higher gain in ability for students with high, rather than low, pre-test scores. As we show in Sec. 5, such an improvement in raw scores, in some cases, actually denotes a larger increase in ability for students with low pre-test scores.

We must clarify an important point: We are not claiming that gain scores calculated via Hake's formula are useless. Indeed, they have played a pivotal role in drawing attention to the ineffectiveness of traditional lectures (Hake 1998; Prather *et al.* 2009). We merely urge caution in their use and interpretation. Hake's gain may provide adequate ordinal rankings of students or groups of students. But sometimes we want more than ordinal comparisons. Sometimes we want to know "how much" better a student or a group of students is compared to another. For example, Hake (1998) made the statement "it appears that the present interactive engagement courses are, on average, more than twice as effective… as traditional courses" (p. 66) when he found that the average normalized gain for interactive engagement courses in his analysis was more than twice as large as for traditional courses. This statement implicitly assumes more than just an ordinal ranking. Conclusions about how much more gain students or groups achieved compared to others require measures on interval or ratio scales.

IRT offers the potential to place students' abilities on an interval scale (Embretson and Reise 2000). Using IRT, we may be able to make meaningful comparisons of gain by looking at the differences in post- and pre-instruction abilities (Embretson and Reise 2000). Section 5 revisits gain from an IRT perspective.

3. IRT BASICS

Section 2 motivated examining concept inventories with IRT. This section is a pedagogical review of the basics of IRT. We first expatiate the three simplest IRT models for dichotomously scored items: the Rasch, two parameter logistic, and three parameter logistic models. These models contain one, two, and three parameters, respectively, that are adjusted in order to provide the best fit between the model and students' response data. After presenting these three models, we discuss the assumptions of IRT and how, when those assumptions hold, we can obtain sample-independent estimates of item parameters of students' abilities. This section concludes with an overview of how IRT parameters are estimated in practice.

3.1. The Rasch Model

The simplest IRT model is called the *one parameter logistic* (1PL) or *Rasch model* (Lord and Novick 1968; Rasch 1960; Hambleton and Jones 1993; Harris 1989; Embretson and Reise 2000; Whitely and Dawis 1974) and can be written as

$$P(X_{pi} = 1 | \theta_p, b_i) = \frac{\exp[\theta_p - b_i]}{1 + \exp[\theta_p - b_i]}. \tag{3}$$

This equation represents the probability that a person p will correctly answer a dichotomously scored item i. X_{pi} represents the person's response to the item; it equals one when the person gives the right answer and zero when the person gives the wrong answer. The probability depends on two factors: The person's ability θ_p and the difficulty of the question b_i. This probability can be interpreted in one of two ways: Either it represents the probability of selecting at random a person with an ability θ_p from a larger population (the *random sampling interpretation*), or it represents the percentage of the number of times a person of ability θ_p will give the correct answer if she is brainwashed and retested without changing any other conditions (the *stochastic subject interpretation*). See Borsboom (2005) and Holland (1990) for discussions of the strengths and weaknesses of each interpretation.

Whence comes Eq. (3)? Following Linacre (2005) and Masters (2001), it may be motivated as follows. (For more formal derivations, see Fischer 1995). Imagine two students A and B with abilities θ_A and θ_B. Each student's ability is a number representing how much of the construct they possess. In the context of the SPCI, θ_p represents how much a student p knows about the properties of stars. A student's ability can range anywhere from $-\infty$ to ∞, and if $\theta_A > \theta_B$ then student A has more ability than student B. Since θ_A and θ_B are considered innate (though not necessarily unchangeable; see Sec. 5) properties of students A and B, the difference $\theta_A - \theta_B$ should be constant regardless of the specific items we use to measure θ_A and θ_B. This last statement is a manifestation of the principle of *specific objectivity* (Rasch 1960), a discussion of which we postpone until Sec. 8 below. In IRT, a person's ability determines the probability that she will correctly answer an item i. If both student A and student B respond to the same item i, then we can imagine that A has a probability of correctly answering the item P_{Ai} and B has a probability of correctly answering the item P_{Bi}. How do P_{Ai} and P_{Bi} relate to θ_A and θ_B? The difference $P_{Ai} - P_{Bi}$ cannot equal $\theta_A - \theta_B$ since the latter can take any value from $-\infty$ to ∞, while the former is restricted to lie between 0 and 1. However, the odds D_{pi} a person p correctly answers item i,

$$D_{pi} = \frac{P_{pi}}{1 - P_{pi}}, \tag{4}$$

ranges from 0 to ∞. Taking the natural logarithm of the odds yields a quantity whose value can lie anywhere between −∞ to ∞ (Linacre 2005). This suggests the following relationship:

$$\theta_A - \theta_B = \ln[D_{Ai}] - \ln[D_{Bi}], \tag{5}$$

which can be rewritten as

$$\theta_A - \theta_B = \ln\left[\frac{P_{Ai}}{1 - P_{Ai}}\right] - \ln\left[\frac{P_{Bi}}{1 - P_{Bi}}\right]. \tag{6}$$

In other words, the difference in ability between A and B equals the difference in log odds units (logits) that they correctly answer an item i. This relationship means that abilities in IRT are measured in logits.

Now imagine two items m and n on a test measuring a single construct. Item m's difficulty is represented by the difficulty parameter b_m, and item n's difficulty is represented by b_n. Just like students' abilities, the items' difficulties can take any value between −∞ to ∞, and if $b_m > b_n$, then item m is a harder item than item n. An item's difficulty is considered to be an intrinsic property of that item, much like a person's ability is an intrinsic property of that person. This means that the difference in difficulties $b_m - b_n$ should not depend on the specific students who answer items m and n (this statement is another manifestation of specific objectivity). Mimicking the reasoning that led to Eq. (6), we find the following relationship between $b_m - b_n$, the probability P_{pm} that a person p correctly answers item m, and the probability P_{pn} that p correctly answers item n:

$$b_m - b_n = \ln\left[\frac{P_{pm}}{1 - P_{pm}}\right] - \ln\left[\frac{P_{pn}}{1 - P_{pn}}\right]. \tag{7}$$

Equation (7) states that the difference in difficulty between items m and n, as measured in logits, equals the difference in the log odds that a person p correctly answers either item.

Equations (6) and (7) indicate that abilities and item difficulties are measured on the same scale. This means a person A's ability can be directly compared to an item m's difficulty to determine the probability that A correctly answers m. Equation (6) specifies θ_A to some constant K_1:

$$\theta_A = \ln\left[\frac{P_{Am}}{1 - P_{Am}}\right] + K_1. \tag{8}$$

Equation (7) likewise specifies b_m to a constant K_2,

$$b_m = \ln\left[\frac{P_{Am}}{1 - P_{Am}}\right] + K_2. \tag{9}$$

In order for both Eqs. (8) and (9) to be true, $K_1 = b_m$ and $K_2 = \theta_A$. Therefore,

$$\theta_A - b_m = \ln\left[\frac{P_{Am}}{1 - P_{Am}}\right]. \tag{10}$$

Equation (10) shows that when $b_m = \theta_A$, a person of ability θ_A has even odds ($D_{Am} = 1$ or $P_{Am} = 50\%$) of correctly answering item m. Equation (10) is equivalent to Eq. (3) for the case where $p = A$ and $i = m$.

Item characteristic curves (ICCs) are a graphical way to view the relationships between abilities, item difficulties, and the probability of giving a correct response. An item's ICC is a plot of the probability of correctly answering that item as a function of respondent ability (Hambleton and Jones 1993; Harris 1989). Figure 1 displays the ICCs for items 16 and 17 on the SPCI. For both items, the probability of a correct response is a monotonically increasing function of ability. Note that item 17 must be a harder item than item 16: For any given ability, the probability of correctly answering item 17 is always lower than the probability of correctly answering item 16. The value of an item's difficulty parameter b_i can be found by identifying the ability value at which the curve's inflection point occurs; in the Rasch model, the inflection point always happens when the probability of a correct response equals 50%. An item's ICC demonstrates that, in the Rasch model, the probability of a person's response is controlled by two factors: the person's ability and the item's difficulty.

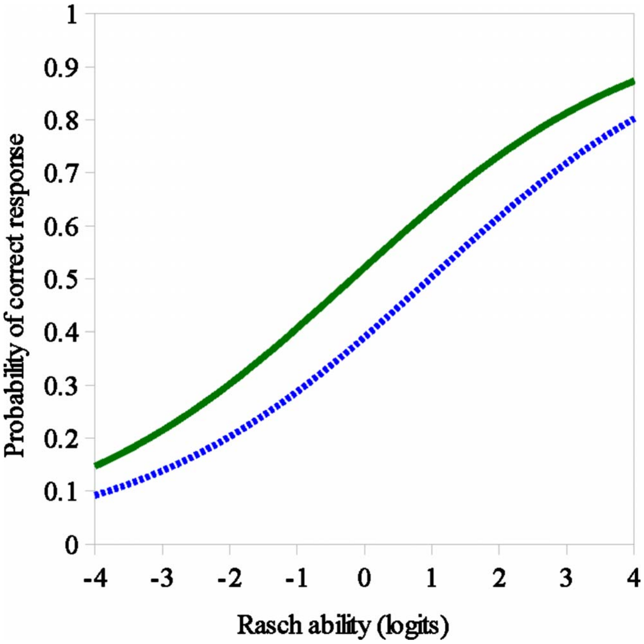

Figure 1. The Rasch model ICCs for items 16 (solid green curve) and 17 (dotted blue curve)

3.2. The Two Parameter Logistic Model

The Rasch model assumes that the student's performance on an item is based solely on her ability and the item's difficulty. However additional item parameters may be added to Eq. (3). In the *two parameter logistic* (2PL) model, a value representing the discrimination of the item a_i is included (Lord and Novick 1968; Hambleton and Jones 1993; Harris 1989):

$$P(X_{pi} = 1 | \theta_p, a_i, b_i) = \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}. \tag{11}$$

This discrimination parameter can have any value between 0 and ∞. Note that, like item difficulty, item discrimination in IRT is not the same as item discrimination in CTT. To understand why a_i reflects an item's discrimination, look at Figure 2. It shows the 2PL ICCs for items 16 and 17 on the SPCI. Item 17 has a larger value for a_i than item 16, so item 17's ICC has a steeper slope than the ICC of item 16. Steeper ICC slopes correspond to greater values of the discrimination parameter (Hambleton and Jones 1993; Harris 1989). One can imagine that in the limit $a_i \rightarrow \infty$, the slope becomes infinite. This corresponds to an item for which people below a certain ability have no chance of answering correctly; people above that ability have a 100% probability of giving the correct answer. In the limit $a_i \rightarrow 0$, an item's ICC becomes a horizontal line. This would mean that all respondents, regardless of ability, would have the same probability of answering the item. Thus, how a person answers the item would tell nothing about her ability. This is why items with higher values of a_i are considered to better discriminate between two people of different abilities.

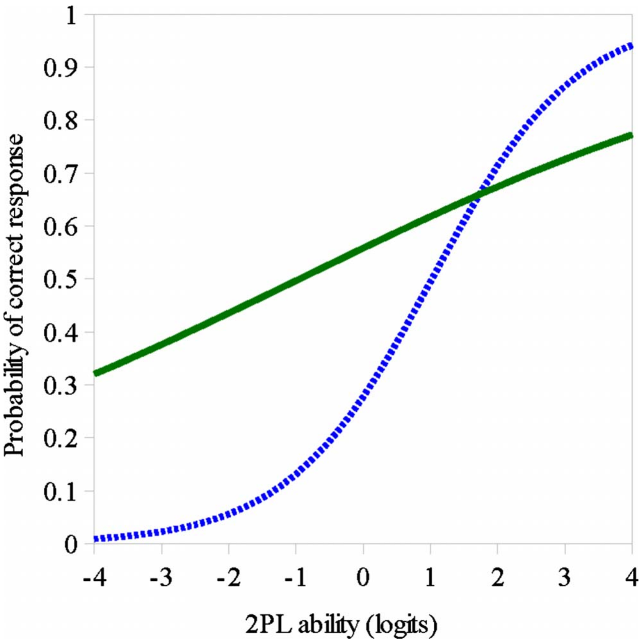

Figure 2. The 2PL model ICCs for items 16 (solid green curve) and 17 (dotted blue curve)

Note that Eq. (3) for the Rasch model implies that all items have a discrimination of unity. In practice, this requirement is sometimes relaxed: While all items must have equal discriminations in the Rasch model, the specific value of the discrimination parameter can have values other than one. This is the approach we took in our analysis of the SPCI (see Sec. 4 below).

3.3. The Three Parameter Logistic Model

The *three parameter logistic* (3PL) model adds a third item parameter (Lord and Novick 1968; Hambleton and Jones 1993; Harris 1989; Lord 1980). This parameter, c_i, is often referred to as the guessing parameter. It takes into account items for which even people of extremely low abilities have a nonzero probability of giving the correct answer (Hambleton and Jones 1993; Harris 1989). The 3PL model is written as

$$P(X_{pi} = 1|\theta_p, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}. \tag{12}$$

Figure 3 shows the 3PL ICC for items 16 and 17. The guessing parameter's effect is most obvious if one compares the ICCs for item 17 for the 2PL (Figure 2) and 3PL (Figure 3) models. The guessing parameter adds a lower asymptote to the probability of correctly answering the item (Hambleton and Jones 1993; Harris 1989; Lord 1980).

3.4. Assumptions of IRT

IRT makes two key assumptions. First, the test is assumed to be unidimensional. That is, it only measures abilities on a single construct (Embretson and Reise 2000; Whitely and Dawis 1974; Lord 1980; Kyngdon 2008). This assumption is consistent with the goals of many concept inventories (Bailey 2009). Second, IRT assumes local independence. This means that all correlations between examinees' responses should be entirely explained by their abilities; no other factor should cause any correlations in item responses (Embretson and Reise 2000; Whitely and Dawis 1974; Kyngdon 2008). Unlike CTT, we can test whether or not the data support these assumptions. Ideally, we would check these assumptions first before proceeding with a discussion of our results. However, one of the goals of this paper is to provide a pedagogical explanation of the methods and interpretations of IRT models. We have found such a pedagogical treatment is clearer to IRT novices if we first discuss the results of our analysis and then describe how to check IRT's assumptions. We thus postpone any further discussion of these assumptions until Sec. 7.

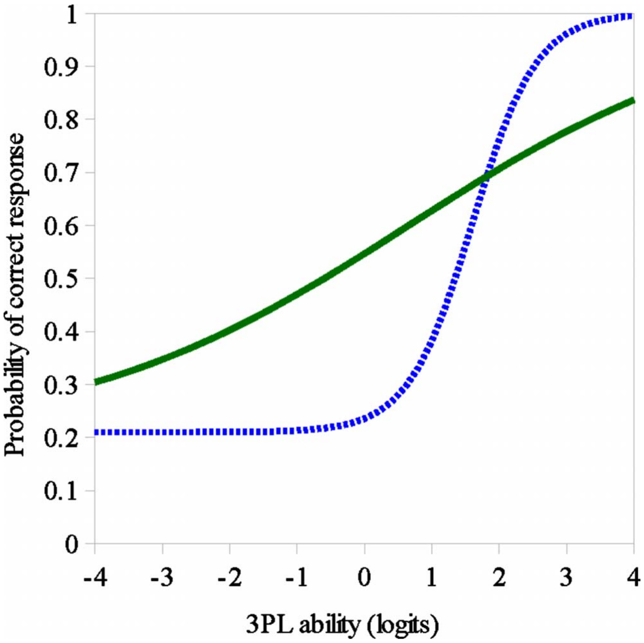

Figure 3. The 3PL model ICCs for items 16 (solid green curve) and 17 (dotted blue curve)

3.5. Parameter Invariance

When IRT's assumptions hold and the model fits the data, then estimates of students' abilities do not depend on the specific items administered, and estimates of item parameters do not depend on the abilities of respondents. This property is known as *parameter invariance* (Hambleton and Jones 1993; Rupp and Zumbo 2006).

To understand parameter invariance, imagine constructing a table like Table 2 for a set of students' responses to a five-item test. For simplicity, assume that the Rasch model fits the observed pattern of responses (e.g., only an item's difficulty and a student's ability determine the probability she gives the correct answer). Each column represents one of the six possible total test scores, and each row corresponds to a different item. In the Rasch model, total test score is a sufficient statistic for ability (Andersen 1977). That is, total test scores contain all the information that the data provides about students' abilities (Lord 1980; Embretson and Reise 2000; Wright 1997). Similarly, the number of people who correctly answer an item is a sufficient statistic for item difficulty (Embretson and Reise 2000). Each cell represents the proportion of students with a given total score (ability) who correctly answer a given item. If we adopt the random sampling interpretation of the Rasch model probability, then the proportion correct represented in each cell is an estimate of the probability that a person with that ability (total score) will correctly answer that item. We can transform these probabilities into log odds. To estimate the ability associated with each total score, we can average the log odds for each column. Item difficulties are likewise estimated by averaging the log odds across each row and multiplying by -1 (so that easier items have smaller difficulty values). These item difficulty estimates may be improved by adjusting their values such that each represents its deviation from the mean item difficulty (Embretson and Reise 2000). Ability estimates may be adjusted by the same factor as item difficulty estimates (Embretson and Reise 2000). This heuristic estimation technique originally was used by Rasch (1960) and has been subsequently used for illustrative purposes by others (Embretson and Reise 2000; Whitely and Dawis 1974).

How does this example motivate parameter invariance? Notice that we estimated item difficulties by averaging the log odds across total test scores. Each cell, which represents the probability that a person with that total score (ability) will give the right answer, contributes equally to this estimate. This means that item difficulty estimates do not depend on the number of people we have in each total score (ability) group. A similar argument applies to how we estimated abilities (Embretson and Reise 2000; Whitely and Dawis 1974).

Here is another way to think about this example. Imagine we only looked at students with low abilities (e.g., total scores of 0, 1, or 2), or we only looked at students with high abilities (e.g., total scores of 3, 4, or 5). If the Rasch model fits, then the log odds in the cells should not change because the log odds only depend on respondents' abilities and item difficulties, which are conceptualized as intrinsic properties of the respondents and items, respectively. When we estimate item difficulties by averaging the log odds for each item across abilities, we probably will get different values depending on whether we use the low ability students or the high ability students. However, our estimates based on these two groups should be related to one another by some linear transformation. In general, IRT parameters are invariant only up to some linear transformation (Rupp and Zumbo 2006). This is due to the fact that IRT models do not specify a scale: The researcher may anchor scores either to items (e.g., by setting the mean item difficulty to be zero; Embretson and Reise 2000) or to respondents (e.g., by specifying that abilities should have a mean of zero and a standard deviation of one; see Embretson and Reise 2000, Harris 1989, or Rupp and Zumbo 2006). In the above example, we adjusted the item difficulties such that they reflect their deviation from the mean item difficulty; this means we anchored the scores to the items. This is why we should get different estimates for item difficulties depending on whether we look at low or high ability students—but those estimates should be linearly related to one another. A similar argument can be made for ability estimates made using different subsets of items.

Table 2. A table of proportions, such as the one shown here, can motivate parameter invariance. Each item receives its own row, and each total score receives its own column. Each cell represents the proportion of respondents with that total score who correctly answer that item. These proportions are used as estimates of the probability that a respondent with a given score will correctly answer a given item

	Total Score					
Item	**0**	**1**	**2**	**3**	**4**	**5**
1	P_{01}	P_{11}	P_{21}	P_{31}	P_{41}	P_{51}
2	P_{02}	P_{12}	P_{22}	P_{32}	P_{42}	P_{52}
3	P_{03}	P_{13}	P_{23}	P_{33}	P_{43}	P_{53}
4	P_{04}	P_{14}	P_{24}	P_{34}	P_{44}	P_{54}
5	P_{05}	P_{15}	P_{25}	P_{35}	P_{45}	P_{55}

We can also think about parameter invariance if we imagine fitting ICCs to each item. To fit an ICC, imagine you have a list of students' responses to a single item. The students are ordered in that list according to their abilities. If we divide this list into bins of students, where each bin contains approximately the same number of students, then we can calculate the average ability within each bin and the proportion of students in each bin who correctly answered the item. Each proportion is an estimate of the probability that a student whose ability is equal to the bin's average ability will give the correct answer. We can then plot a point for each bin on a graph whose axes are identical to those of the ICCs shown in Figs. 1–3 above. If the model fits, then we should be able to fit the same ICC to the item regardless of whether we base our fit off of low ability students, high ability students, or both. If we look only at low ability students, then the data should trace out the lower part of the ICC. If we look only at high ability students, then the data should trace out the upper part of the ICC. The two ICCs may be offset from one another since the item's difficulty was estimated using two different populations. However, a linear transformation should make the two ICCs coincident and place them on the same scale.

The above example was for the Rasch model. What about the 2PL and 3PL models? The total test score is only a sufficient statistic for the Rasch model (Andersen 1977; Embretson and Reise 2000). In the 2PL model, the total test score depends both on a student's ability and the specific items she correctly answers; items with higher discriminations are weighted higher (Lord 1980). There is no sufficient statistic for the 3PL model due to the presence of the guessing parameter (Lord 1980). Nevertheless, parameter invariance still holds for both the 2PL and the 3PL models as long as they fit (Hambleton and Jones 1993).

This last point is critical: *Parameter invariance is only true if an IRT model fits the data* (Hambleton and Jones 1993; Whitely and Dawis 1974). If the model does not fit, then there is no guarantee of parameter invariance, which negates many of the statistical advantages of IRT relative to CTT. In Sec. 6, we discuss some methods for judging model fit.

3.6. Estimating IRT Parameters

The example parameter estimation procedure described in Sec. 3.5 is crude and can only be used if one adopts the Rasch model. In practice, IRT parameters are estimated via computer programs that maximize the likelihood of the observed response patterns. There are several maximum likelihood procedures one can use, and many IRT software packages allow one to choose between them. A joint maximum likelihood (JML) procedure alternates between estimating item parameters by assuming abilities are known and estimating abilities by assuming item parameters are known, with each iteration using the improved estimates from the previous step. Conditional maximum likelihood (CML) estimation only works for the Rasch model because it takes advantage of the fact that, in the Rasch model, total test score is a sufficient statistic for estimating abilities (Andersen 1977; Embretson and Reise 2000). Thus, abilities can be estimated by maximizing the likelihoods of the observed total scores, and difficulties can be estimated by maximizing the likelihoods of the number of people who correctly answer each item. In marginal maximum likelihood (MML) estimation, one assumes the sample of respondents is drawn from a population with a certain distribution of abilities (usually the standard normal distribution). Using this population distribution, one can integrate over abilities to obtain the marginal likelihood function, which is then maximized to estimate the item parameters. Once the item parameters are estimated, respondents' abilities can be estimated. Such abilities estimates may be done using a simple maximum likelihood estimation (MLE). However, if one has already assumed a population distribution, one can maximize the posterior distribution, which is just the product of the likelihood function and the hypothesized population distribution. If one uses the mode of the posterior distribution as the best estimate of θ_p, then the ability estimates are said to be found using a maximum *a posteriori* (MAP) method. If instead one uses the mean of the posterior distribution, then the method is called expected *a posteriori* (EAP). The details of these estimation procedures are described elsewhere (e.g., Baker and Kim 2004).

Each estimation technique has its own benefits and handicaps. Embretson and Reise (2000) discuss the advantages and disadvantages of these estimation procedures in detail; we highlight some of the salient points here. MML, MAP, and EAP estimations are dependent on choosing an appropriate prior distribution. However, unlike CML and MLE, they can provide ability estimates for examinees who correctly answer all or none of the test's items. CML is also only applicable to the Rasch model. However, CML does not require any assumptions about prior distributions, and neither does JML. JML is straightforward to program, but its parameter estimates are inconsistent: Adding more examinees to the sample population does not improve item parameter estimates. JML, MAP, and EAP estimates may also be biased (the expected value of θ_p does not always equal the true value of θ_p). Given the different assumptions, benefits, and handicaps of these various methods, researchers should always report which estimation procedure they employ when using IRT.

4. IRT ANALYSIS OF THE SPCI

With the background of Sec. 3 in mind, what do we get when we apply IRT to the SPCI? In this section, we apply the Rasch, 2PL, and 3PL models to the SPCI. Model parameters were estimated using the BILOG-MG software (Zimowski *et al.* 1996). We used an MML estimation procedure to find the item parameters since JML is inconsistent and since CML only applies to the Rasch model (Embretson and Reise 2000). We used these item parameters to construct the ICCs shown in Sec. 3 above. Abilities were estimated using an EAP estimation approach since it utilizes the prior distribution used by MML and it produces smaller standard errors than MLE (Embretson and Reise 2000). BILOG-MG estimated the model parameters from the matched pre- and post-instruction responses of 334 students.

We used BILOG-MG's multigroup capabilities (Zimowski *et al.* 1996) to estimate the item parameters using both students' pre-test and post-test responses. BILOG-MG anchored the scale by setting the mean ability of the pre-instruction population to 0 and the standard deviation to 1.

The estimated parameters for the Rasch, 2PL, and 3PL models are shown in Tables 3–5 below. As we alluded to in Sec. 3.2, all the items have the same nonunity value for the discrimination parameter (a_i=0.461) when we apply the Rasch model. Additionally, c_i=0 for all items in the Rasch and 2PL models since neither model has a guessing parameter.

Note that we do not report any 3PL item parameters for items 3 and 13. BILOG-MG was unable to estimate parameters for these items. Consequently, we excluded them from our 3PL analysis of the SPCI.

Table 3. Rasch model item parameters for the SPCI. SE stands for standard error

Item	a_i	a_i's SE	b_i	b_i's SE	c_i	c_i's SE
1	0.461	0.010	0.399	0.189	0	0
2	0.461	0.010	4.786	0.213	0	0
3	0.461	0.010	4.933	0.217	0	0
4	0.461	0.010	−1.180	0.197	0	0
5	0.461	0.010	4.396	0.227	0	0
6	0.461	0.010	1.192	0.179	0	0
7	0.461	0.010	0.549	0.191	0	0
8	0.461	0.010	2.677	0.178	0	0
9	0.461	0.010	1.885	0.176	0	0
10	0.461	0.010	1.824	0.194	0	0
11	0.461	0.010	1.341	0.190	0	0
12	0.461	0.010	3.663	0.219	0	0
13	0.461	0.010	4.139	0.190	0	0
14	0.461	0.010	−3.404	0.262	0	0
15	0.461	0.010	2.498	0.198	0	0
16	0.461	0.010	−0.185	0.182	0	0
17	0.461	0.010	0.968	0.198	0	0
18	0.461	0.010	4.396	0.226	0	0
19	0.461	0.010	3.100	0.200	0	0
20	0.461	0.010	2.482	0.192	0	0
21	0.461	0.010	0.834	0.187	0	0
22	0.461	0.010	5.796	0.281	0	0
23	0.461	0.010	1.612	0.184	0	0

Why are items 3 and 13 problematic? Look again at the CTT statistics in Table 1. Items 3 and 13 are among the hardest items post-instruction. Additionally, their point-biserials actually decrease from pre- to post-instruction. The point-biserials of other items increase from the pre-test to the post-test. These CTT statistics show that items 3 and 13 are answered incorrectly by many students and higher ability students do not do much better, if at all, on these items.

The fact that few students at any ability level correctly answered items 3 and 13 explains why BILOG-MG could not find 3PL parameters for these items. In the 3PL model, there are three item parameters that might explain why high and low ability students have approximately the same probability of correctly answering an item. The item might be so difficult that the probability a high ability student will chose the correct answer is approximately the same as the probability that a low ability student will give the right answer. Alternatively, the item might not be very discriminating. Finally, the guessing parameter could be high enough that even low ability students can guess the correct answer at the same frequency at which high ability students select the correct answer. Some combination of these three possibilities is also a possible explanation. Without additional data, BILOG-MG simply cannot determine the values for the discrimination, difficulty, and guessing parameters for items 3 and 13. This is not an issue with the Rasch and 2PL models since they have fewer item parameters.

As an interesting aside, one of us (Bailey) already suspected that item 3 was problematic before running this IRT analysis. In fact, it was removed from versions of the SPCI that have been administered since this pilot study. Item 13 was a surprise, however. In the most recent SPCI data (on which we plan to report in a future publication), there was a class in which almost every student incorrectly answered item 13 on the post-test. Clearly, we must re-examine and revise item 13.

Figure 4 graphically compares the item difficulty parameters for each of the three IRT models. In general, the estimates of b_i for each item are similar for the Rasch, 2PL, and 3PL models. Three notable exceptions are the 2PL difficulties for items 2, 3, and 13, which are much higher than the Rasch and 3PL (for item 2 only) difficulty estimates. These are the same three items for which students of high ability have roughly the same probability of giving the right answer as students of low ability. The 2PL model accounts for this by assigning a high value to the difficulty parameters for these items.

Table 4. 2PL model item parameters for the SPCI. SE stands for standard error

Item	a_i	a_i's SE	b_i	b_i's SE	c_i	c_i's SE
1	0.546	0.063	0.572	0.166	0	0
2	0.129	0.033	13.004	3.050	0	0
3	0.130	0.035	13.377	3.295	0	0
4	0.394	0.064	−1.390	0.401	0	0
5	0.494	0.057	4.506	0.348	0	0
6	0.325	0.047	1.388	0.252	0	0
7	0.599	0.062	0.724	0.152	0	0
8	0.168	0.036	5.134	0.923	0	0
9	0.214	0.039	2.818	0.454	0	0
10	0.648	0.062	1.784	0.151	0	0
11	0.586	0.060	1.411	0.157	0	0
12	0.630	0.062	3.386	0.221	0	0
13	0.082	0.023	16.203	4.336	0	0
14	0.602	0.122	−2.505	0.560	0	0
15	0.565	0.062	2.473	0.202	0	0
16	0.247	0.045	−0.947	0.493	0	0
17	0.935	0.099	1.022	0.112	0	0
18	0.490	0.057	4.526	0.345	0	0
19	0.452	0.053	3.343	0.284	0	0
20	0.432	0.050	2.741	0.243	0	0
21	0.477	0.059	0.961	0.180	0	0
22	0.689	0.074	5.053	0.307	0	0
23	0.418	0.049	1.811	0.212	0	0

Table 5. 3PL model item parameters for the SPCI. SE stands for standard error

Item	a_i	a_i's SE	b_i	b_i's SE	c_i	c_i's SE
1	0.905	0.160	1.235	0.324	0.194	0.077
2	1.061	0.449	5.111	0.769	0.164	0.020
3	···	···	···	···	···	···
4	0.589	0.094	−0.157	0.475	0.206	0.094
5	1.369	0.309	3.292	0.167	0.090	0.021
6	0.763	0.200	2.395	0.366	0.262	0.069
7	0.902	0.114	1.050	0.238	0.121	0.058
8	1.195	0.483	4.241	0.467	0.298	0.027
9	0.535	0.156	3.627	0.535	0.249	0.066
10	1.222	0.213	1.929	0.179	0.136	0.043
11	1.142	0.215	1.797	0.219	0.175	0.053
12	1.864	0.358	2.736	0.105	0.099	0.019
13	···	···	···	···	···	···
14	0.812	0.134	−1.589	0.388	0.173	0.084
15	2.037	0.503	2.546	0.114	0.193	0.025
16	0.406	0.075	0.650	0.641	0.199	0.090
17	2.122	0.523	1.602	0.147	0.210	0.039
18	1.176	0.256	3.328	0.180	0.079	0.022
19	1.317	0.315	2.926	0.159	0.157	0.029
20	0.740	0.121	2.456	0.233	0.093	0.043
21	1.161	0.276	2.015	0.250	0.291	0.054
22	1.867	0.426	3.424	0.142	0.039	0.011
23	0.807	0.161	2.117	0.292	0.167	0.062

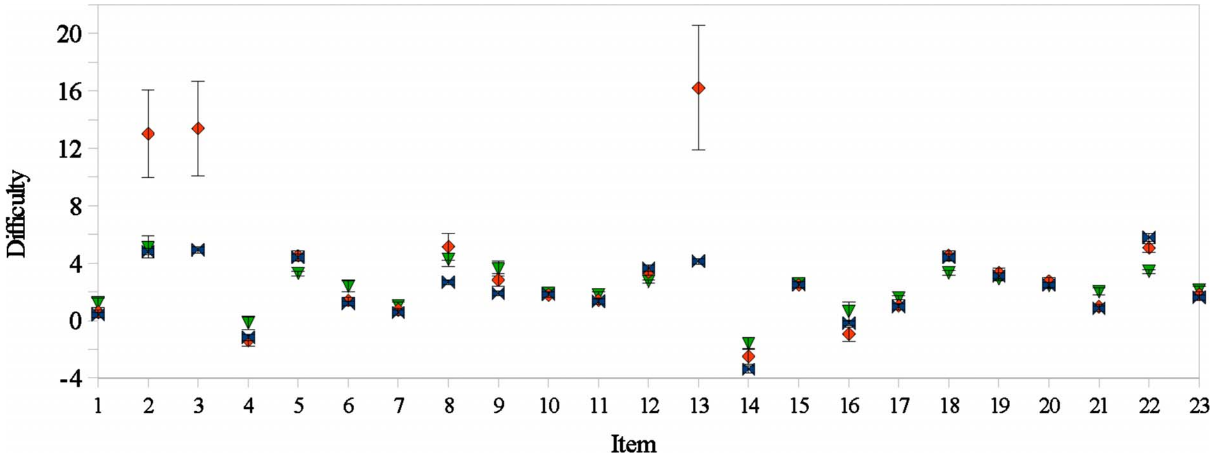

Figure 4. A comparison of the difficulty (b_i) parameters for each item for the Rasch (blue double triangles), 2PL (red diamonds), and 3PL models (green triangles). Error bars represent standard errors

Figure 5 compares the 2PL and 3PL discrimination parameters for each item. For each item, the 3PL value for a_i is larger than the 2PL value. Why? Remember that the 3PL model also includes a guessing parameter. The 3PL model can thus attribute at least some of the correct responses of low ability students to an item to guessing. In contrast, since the 2PL model does not account for guessing, low ability students who nevertheless give a correct answer necessarily reduce the discriminatory capability of an item. By including the effects of guessing, the 3PL model makes items seem more discriminating than they would appear in the 2PL model.

Figure 6 shows the 3PL guessing parameters for each item. Many items have guessing parameters in the range of 0.20 to 0.25. This makes sense, as six items (items 2, 9, 10, 16, 18, and 21) have four answer choices while the rest have five. If students simply guess on each question, we expect them to choose the correct answer 25% (for four choices) or 20% (for five choices) of the time. However, this raises an important issue. On concept inventories, we do not want students to simply guess an answer. Instead, we want to write items such that high ability students are likely to pick the correct answer and low ability students are likely to pick one of the distractors. If low ability students are frequently drawn to one or more distractors, then they should choose the correct answer at a lower rate than one might expect by chance (Sadler *et al.* 2010). The fact that so many items have guessing parameters in the 0.20–0.25 range may indicate that many items on the SPCI do not have appealing distractors.

Figure 7 shows the ability estimates (both before and after instruction) for the Rasch, 2PL, and 3PL models as a function of the percent correct on the SPCI. Note that there is a one-to-one correspondence between percent correct and Rasch model ability estimates. This is a manifestation of the fact that the total test score is a sufficient statistic for the Rasch model (Andersen 1977; Embretson and Reise 2000). Since the total test score is not a sufficient statistic for either the 2PL or 3PL ability estimates, a variety of ability estimates may

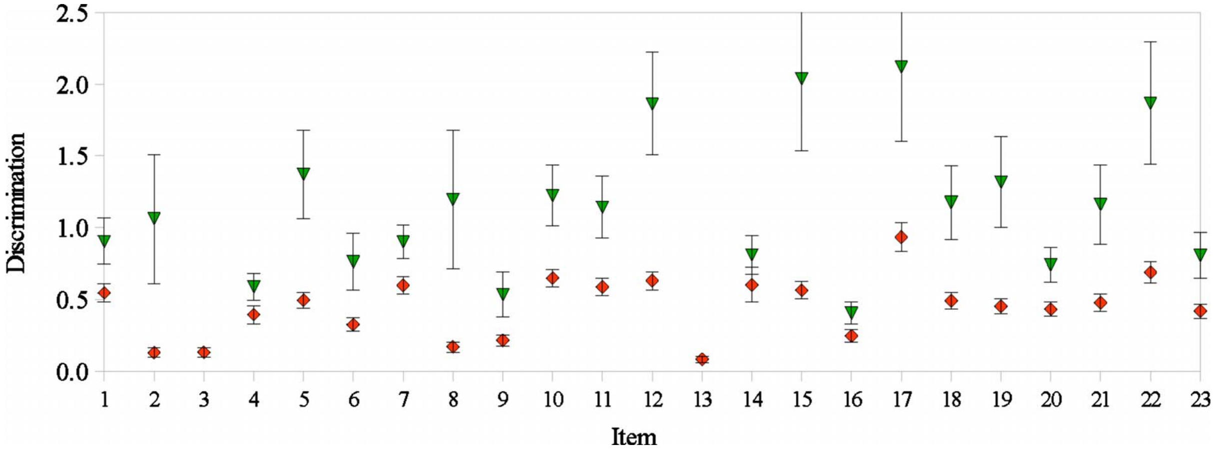

Figure 5. A comparison of the discrimination (a_i) parameters for each item for the 2PL (red diamonds) and 3PL models (green triangles). Error bars represent standard errors

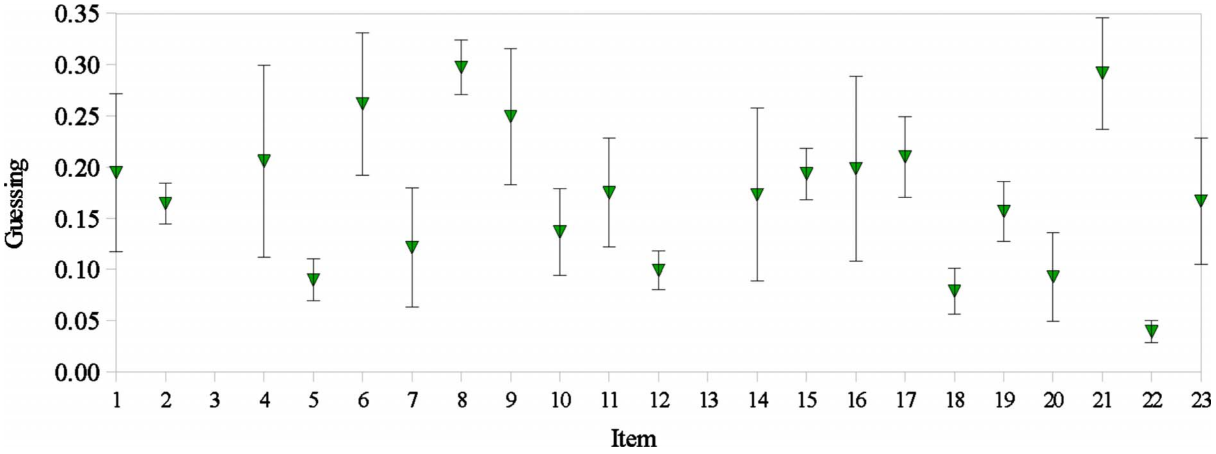

Figure 6. A comparison of the guessing (c_i) parameters for each item for the 3PL model. Error bars represent standard errors

correspond to a single percent correct score. Another way of thinking about this is that 2PL and 3PL ability estimates depend on which items a student correctly answers; in these models, items with larger discriminations are given larger weights in the estimation of abilities.

Figure 7 also shows that the ability estimates for any of the three models are nonlinear functions of the percent correct. This means a change of 10% correct corresponds to a different change in ability depending on where one starts. We revisit this issue when we discuss IRT gain calculations in Sec. 5 below.

How accurate are the ability estimates in Figure 7? We can answer this question by examining the standard errors associated with each ability estimate, shown as the error bars in Figure 7. Another way to answer this question is to look at plots of the standard errors as a function of ability (Hambleton and Jones 1993). This is shown for the Rasch, 2PL, and 3PL models in Figs. 8–10. These figures also show the test information function, which is the reciprocal of the standard error plot (Hambleton and Jones 1993; Lord 1980). Where the standard error is lowest (and the test information highest) is where the SPCI gives the best estimates of ability (Hambleton and Jones 1993).

The differences in the plots in Figures 8–10 can be explained by the effects of the discrimination and guessing parameters (Embretson and Reise 2000). In the 2PL and 3PL models, items can have different discriminations from one another. This allows us to better estimate students' abilities based on how they answer

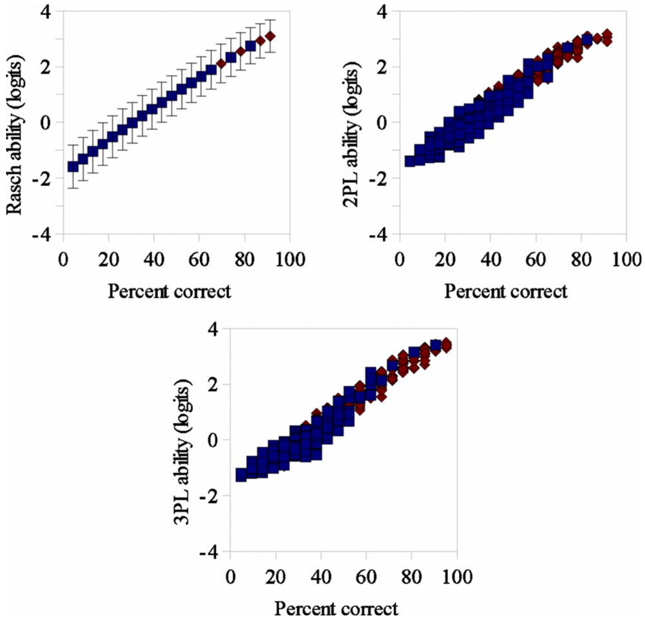

Figure 7. Rasch model (upper left), 2PL (upper right), and 3PL (bottom) ability estimates as a function of the percent correct on the SPCI both pre- (blue squares) and post-instruction (red diamonds). Error bars represent the standard errors. The error bars are suppressed in the 2PL and 3PL graphs for clarity

the most discriminating items, which is why the test information peaks in Figure 9 (the 2PL model) and 10 (the 3PL model) are higher than in Figure 8 (the Rasch model). The guessing parameter has the opposite effect: Nonzero guessing parameters mean that even students with very low abilities have a nonzero probability of correctly answering items, which reduces our ability to estimate abilities based on total test scores. This effect is the most pronounced at the extremes of the distribution. This is why the test information curve is so low for abilities less than 0 logits in Figure 10. Figures 8–10 demonstrate how the different item parameters influence our capacity to estimate students' abilities.

Figures 8–10 underscore an important point: Test items should be chosen such that the test information curve peaks at the ability around which most examinees are clustered (Hambleton and Jones 1993; Lord 1980). For example, Figure 10 implies that the SPCI can best estimate the 3PL abilities of students whose abilities are around 2.8 logits. If we administer the test to a population of students with abilities near −2.8 logits, then we may not get very accurate 3PL ability estimates due to the high standard error at around that point. If

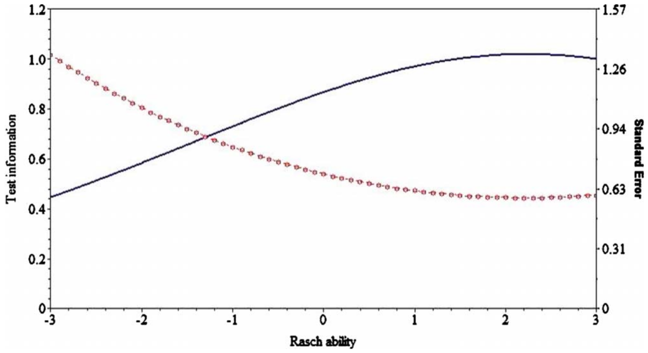

Figure 8. The standard error (dotted red curve) and test information (solid blue curve) as a function of ability for the Rasch model

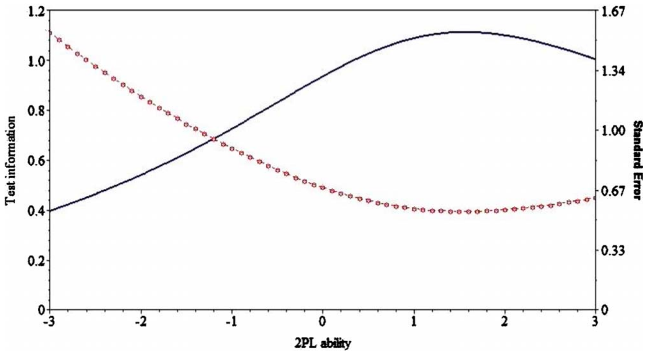

Figure 9. The standard error (dotted red curve) and test information (solid blue curve) as a function of ability for the 2PL model

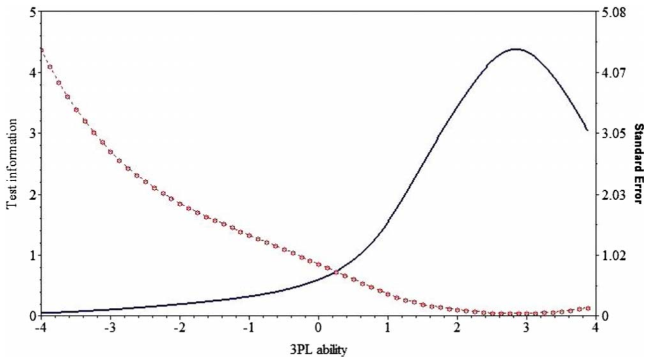

Figure 10. The standard error (dotted red curve) and test information (solid blue curve) as a function of ability for the 3PL model

the test information curve peaks at a higher ability than most examinees possess, then the test designer should include additional easier items in order to better estimate examinees' abilities; conversely, if the test information curve peaks at a lower ability than most examinees posses, then the test designer should include additional harder items. Note that this contradicts a CTT principle of test construction that says one should choose items with P-values around 0.50 in order to maximize test reliability (Ding and Beichner 2009).

If one uses a Rasch model, then one can visually check whether or not the items' difficulties fall around the same logit values as examinees' abilities by constructing a Wright map. See Pek and Poh (2000), Planinic (2006), and Planinic, Ivanjek, and Susac (2010) for examples of Wright maps from PER studies. A Wright map for the SPCI data is shown in Figure 11. On the left is a histogram of respondents' abilities both before and after instruction. Since abilities and item difficulties are both measured in logits, we placed each item from the SPCI at the logit value of its difficulty on the right. Figure 11 shows that students' post-instruction abilities have a larger mean and standard deviation than their pre-instruction abilities. Figure 11 also shows that many items are located at logit values larger than many students' abilities, both before and after instruction.

One can also use the Wright map to quickly estimate the probability a student with a given ability will correctly answer an item (Wilson 2005). One need only read off an ability value and an item difficulty from the Wright map and apply Eq. (3). For example, the peak of the pre-instruction distribution appears to lie near 0 logits. Students in this part of the distribution have an approximately 50% chance of correctly answering item 16 since its difficulty value also places it near 0 logits. Note that we could not make this quick estimation if we instead created a Wright Map from our 2PL or 3PL data. In either case, we would need to know the values of other item parameters (a_i and c_i) in order to calculate probabilities. This is one reason why one does not, in general, construct Wright maps when using the 2PL and 3PL models.

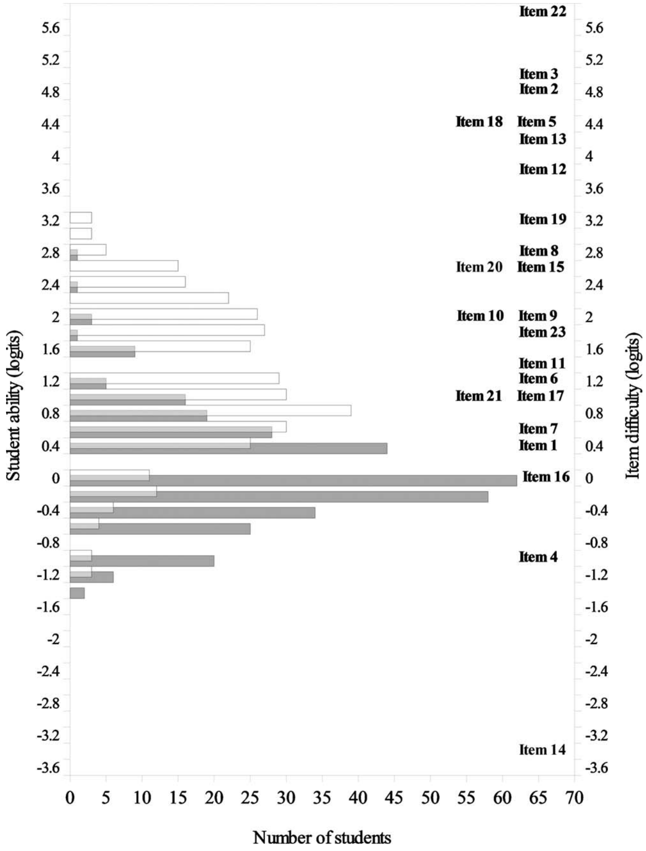

Figure 11. The Wright map for the Rasch model ability and item difficult estimates for the SPCI. Each bin is 0.2 logits wide and is labeled by the upper value of the bin (e.g., bin 1 includes logit values between 0.8 and 1). Pre-instruction ability estimates are shown in grey while post-instruction abilities are shown in white

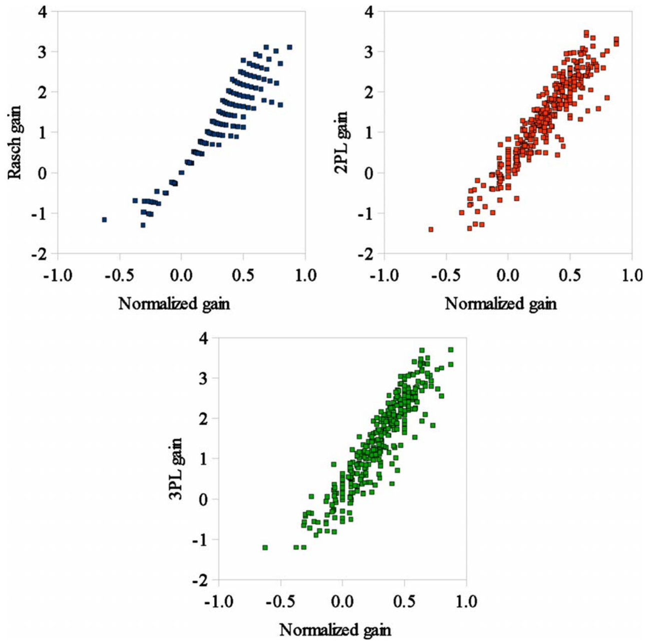

Figure 12. IRT calculated gains as a function of Hake's normalized gain. The upper left panel shows Rasch model gains, the upper right panel shows 2PL gains, and the bottom panel shows 3PL gains

5. LEARNING GAINS IN IRT

Researchers use concept inventories to measure learning gains. As we discussed in Sec. 2.2, these measures of learning gain are frequently obtained directly from the raw pre- and post-test scores. IRT models convert raw scores into measures of ability. If IRT-estimated abilities fall on an interval scale (which is not always the case—see Sec. 8 below), then we can compute gains by subtracting students' pre-instruction abilities from their post-instruction abilities (Embretson and Reise 2000). Given that IRT postulates a nonlinear relationship between raw scores and abilities, how do IRT computed gains compare with more traditional gain calculations?

Figure 12 compares IRT calculated gains ($\theta_{\text{post}} - \theta_{\text{pre}}$) to gains calculated using Hake's (1998) normalized gain formula. Regardless of which IRT model one chooses, Figure 17 shows that there is not a one-to-one correspondence between Hake's normalized gain and changes in IRT ability. A given value for Hake's normalized gain may correspond to multiple IRT gains.

In Sec. 2.2, we alluded to the fact that higher values for Hake's gain do not always correlate with greater changes in ability. Here is one concrete example: Say student A has a raw pre-test score of 1 on the SPCI, while student B has a raw pre-test score of 18. After instruction, student A's SPCI score improves to 4 and student B's score improves to 21. Both students improved by three points (4%). According to Hake's formula, student B has the higher gain: $g_B = 1.57$, while $g_A = 0.14$. If we instead look at the differences in their pre- and post-instruction Rasch model abilities, then we see student A's ability increased by 0.81 logits, while student B's ability increased by 0.55 logits. The discrepancy in who achieved the largest gain may be explained by the fact that Hake's gain is biased toward high pre-test scores (Brogt *et al.* 2007). Student B has the higher pre-test score, and Hake's formula assigned student B a larger normalized gain than student A, even though student A's ability increased by more logits than student B's ability. With this example in mind, we must ask the provocative question: Which is the better measure of students' learning gain?

6. ASSESSING MODEL FIT

As Secs. 4 and 5 show, IRT models provide substantial information about a test's items and test takers. Much of this information presumes that the IRT model one uses actually fits the data. In this section, we examine how well the Rasch, 2PL, and 3PL models fit our SPCI data. Some PER IRT studies explicitly address the need to check model fit (Lee *et al.* 2008; Planinic 2006; Planinic, Ivanjek, and Susac 2010; Wang and Bao 2010), while others do not (Pek and Poh 2000; Marshall, Hagedorn, and O'Connor 2009). The fact that we can test the applicability of a given IRT model is one advantage IRT has over CTT (Embretson and Reise 2000).

Table 6. χ^2 *p*-values for the Rasch, 2PL, and 3PL models. All χ^2 *p*-values <0.05 are bolded and italicized

Item	Rasch	2PL	3PL
1	***0.0001***	***0.0145***	***0.0061***
2	***0.0000***	***0.0090***	***0.0095***
3	***0.0000***	0.3025	⋯
4	0.6956	0.2579	0.6896
5	0.2426	0.8185	0.7599
6	0.7423	0.2674	0.2622
7	***0.0002***	0.1961	***0.0000***
8	***0.0000***	0.4108	0.1714
9	***0.0056***	0.2497	0.9367
10	***0.0002***	***0.0169***	0.5795
11	***0.0017***	0.5726	0.4697
12	***0.0000***	***0.0072***	0.7515
13	***0.0000***	0.0195	⋯
14	***0.0021***	0.0551	***0.0000***
15	***0.0006***	***0.0211***	0.7848
16	***0.0052***	***0.0271***	***0.0422***
17	***0.0000***	***0.0271***	0.9307
18	0.1942	0.2831	0.9866
19	0.1661	***0.0151***	0.7300
20	0.1916	0.1728	***0.0000***
21	***0.0013***	***0.0251***	***0.0013***
22	***0.0000***	0.3375	0.7108
23	0.1677	0.9547	0.4482

BILOG-MG checks the model fit for each item by performing a χ^2 test. To calculate the χ^2 statistics, BILOG-MG first bins examinees based on their estimated abilities. BILOG-MG then calculates the proportion of examinees within each bin that correctly answer a given item. Finally, it compares this observed proportion to the proportion correct predicted by the IRT model. (N.B.: This procedure uses the random sampling interpretation of the probability in IRT models; see Holland 1990 and Borsboom 2005). The null hypothesis for this test is that the IRT model describes the observed response frequencies for each subgrouping of ability. As with all χ^2 tests, the null hypothesis is rejected for an item when the *p*-value associated with the item's χ^2 value (not to be confused with the *P*-value described in Sec. 2 as a CTT estimate of item difficulty) is small. This means items whose observed response patterns fit the model will have large *p*-values. Convention suggests flagging items as exhibiting model misfit when they have *p*-values <0.05; when the *p*-value is this low, it suggests that the deviations of the observed response pattern from the IRT model cannot be explained by chance. Note that this is opposite of most tests of statistical significance in which one wants $p < 0.05$ in order to reject a null hypothesis that two samples are not different.

Table 6 shows the χ^2 *p*-values for each item on the SPCI for the Rasch, 2PL, and 3PL models. BILOG-MG used both pre- and post-test responses to calculate these fit statistics. These *p*-values are bolded and italicized whenever they are smaller than the 0.05 threshold. Table 6 shows that the Rasch model has the largest number of misfitting item (16 items), followed by the 2PL model (9 items), and then the 3PL model (7 items).

BILOG-MG also produces fit plots: These are graphical representations of how well the observed responses to an item fit the model (Harris 1989). Figure 13 is an example of a fit plot. It shows the 2PL ICC for item 23. The dots are the observed proportion of respondents in an ability bin who gave the right answer. The error bars represent the standard error of the ICC at a given logit value. The 2PL model fits the observed responses to item 23 well since all dots fall within the error bars in Figure 13.

Fit plots can reveal the degree of misfit for each item. For example, consider Figure 14. The χ^2 *p*-value indicates the Rasch model does not fit the pattern of observed responses for item 7. Figure 14 shows that this misfit is due to one point; overall, the amount of misfit does not appear to be too severe. In fact, when we examined the fit plots for all items we found that many of the items with χ^2 *p*-values <0.05 possess only a slight

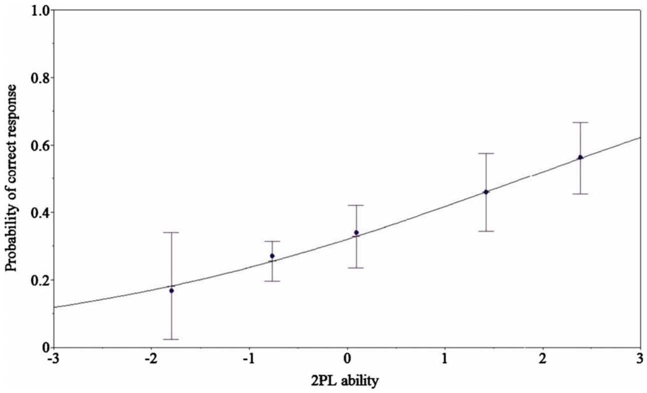

Figure 13. The 2PL fit plot for item 23. The curve is the 2PL ICC. Each dot represents the proportion of respondents in an ability bin who gave the correct answer. The error bars represent the standard error of the ICC

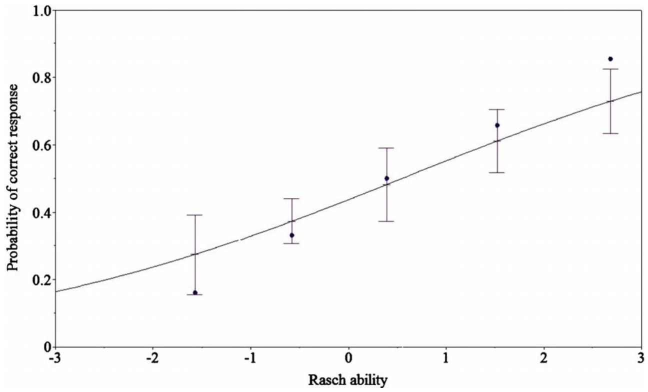

Figure 14. The Rasch model fit plot for item 7. The curve is the Rasch model ICC. Each dot represents the proportion of respondents in an ability bin who gave the correct answer. The error bars represent the standard error of the ICC

amount of misfit. We could not draw this conclusion without looking at the graphs BILOG-MG generates.

Figure 15 shows another fit plot, this time for item 5 with parameters estimated using the 3PL model. BILOG-MG assigns this item a χ^2 p-value of 0.7599, despite the fact that the model does not seem to fit the data at all. What is going on? BILOG-MG does not choose ability bins such that each bin has a roughly equal number of respondents. We hypothesize that BILOG-MG selected some ability bins with few or no students. This would throw off the χ^2 values BILOG-MG calculates. Regardless of the explanation, Figure 15 further underscores the need to look at fit plots; if one only looks at χ^2 p-values, then one would have no idea that there was any problem with this item.

BILOG-MG's fit statistics have other problems. For example, they are known to exhibit Type I errors (Orlando and Thissen 2000). While there are other methods for ascertaining fit, IRT fit statistics, in general, are underdeveloped. Many researchers are working to better understand the strengths and limitations of various IRT fit statistics (e.g., Orlando and Thissen 2000; Wu and Adams 2010), although much work remains to be done. Furthermore, researchers who exclusively use the Rasch model often use different fit statistics than the ones we describe here; see Wilson (2005) for a review of these statistics and Wu and Adams (2010) for important warnings on their misuse. For the time being, we advise researchers who use IRT to consider multiple approaches (such as looking at both χ^2 p-values and fit plots) to judge model fit.

7. TESTING THE ASSUMPTIONS OF IRT

In addition to judging model fit, IRT users should also examine whether or not the two underlying assumptions of IRT—unidimensionality and local independence—also hold. Only some of the PER and AER IRT studies discuss checking these assumptions (Ding and Beichner 2009; Marshall, Hagedorn, and O'Connor 2009; Planinic, Ivanjek, and Susac 2010; Sadler 1998; Wang and Bao 2010). If these assumptions do not hold, then many

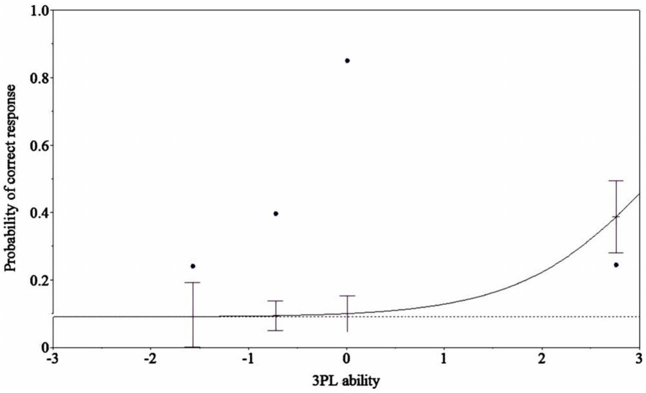

Figure 15. The 3PL fit plot for item 5. The curve is the 3PL ICC. Each dot represents the proportion of respondents in an ability bin who gave the correct answer. The error bars represent the standard error of the ICC

of IRT's potential benefits, such as parameter invariance, will not be realized. In this section we outline one method for testing unidimensionality and one method for testing local independence. See Embretson and Reise (2000) and references therein for overviews of alternative approaches.

We use Bejar's (1980) approach for evaluating unidimensionality. Bejar's technique relies on estimating item parameters twice. One estimation uses students' responses to every item, as usual. We call these the total test-based estimates. We then estimate item parameters a second time using only a subset of the test's items. These are the subtest-based estimates. One should choose items for the subtest that one believes probe a different content area from the other items (Bejar 1980).

If the item is truly unidimensional, then a plot of the subtest-based item parameter estimates versus the total test-based item parameter estimates should lie near a line with a slope of 1 and an intercept of 0 (Bejar 1980). Why? If the points depart significantly from this line, then the subtest-based item parameter estimates do not equal the total test-based item parameter estimates. But different item parameter estimates for an item imply different probabilities for passing that item. Bejar noted that "[t]his is inadmissible because it implies that performance on that item depends on which items are included in the test, which contradicts the assumption that a single trait explains performance on all items" (Bejar 1980, p. 284).

Within the SPCI, there are natural divisions of the items by content. Out of the 23 items, 13 items cover star properties (including stellar masses, the temperature-color relationship, and the mass-lifetime relationship), five cover fusion, and five cover star formation (Bailey 2007). We selected the 13 star properties items (items 3, 5, 7, 9, 10, 13, 16–18, and 20–23) as our subtest for Bejar's unidimensionality test.

Figures 16–18 show the subtest-based item difficulty estimates plotted versus the total test-based item difficulty estimates for the Rasch, 2PL, and 3PL models, respectively. While similar plots could be made for the discrimination and guessing parameters, Bejar (1980) warned that these item parameters are often not estimated as accurately as item difficulties. Such plots might confound departures from unidimensionality with parameter estimation accuracy issues.

Figures 16–18 each show two lines. The solid lines represent a line with a slope of 1 and an intercept of 0. The dashed lines represent a line fitted to the data (what Bejar calls the "principle axis"). These two lines are separated by angles of 39.6°, 38.7°, and 38.3° for the Rasch, 2PL, and 3PL models, respectively. The fact that the solid and dashed lines in Figures 16–18 have different slopes suggests the SPCI is not unidimensional. In the future, we may reanalyze the SPCI using one of the many multidimensional IRT models, which we do not discuss here (but see Ackerman, Gierl, and Walker 2003 and Briggs and Wilson 2003 and references therein).

We used Yen's (1984) Q3 statistic to evaluate the local independence assumption. Yen's Q3 statistic looks at the difference between the observed and model-predicted responses to each item and then correlates these residuals across respondents by item (Yen 1984). Tables 7–9 show Yen's Q3 statistic for each item pair on the SPCI for the Rasch, 2PL, and 3PL models, respectively. Yen and Fitzpatrick (2006) recommend flagging all item pairs for which her eponymous statistic $\geq|0.20|$. Every such value is bolded in Tables 7–9. In general, most entries in Tables 7–9 are well below this threshold, indicating that local independence is a valid assumption for most item pairs on the SPCI.

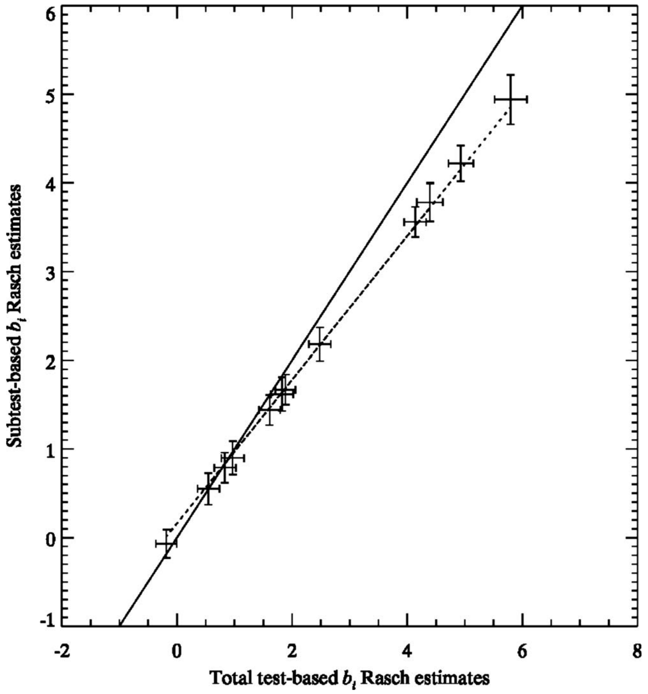

Figure 16. The Rasch model difficulties (b_i) of the 13 star properties items estimated without the other ten items versus the difficulties estimated with the other ten items. Error bars represent standard errors. The solid line is where the points should lie if unidimensionality holds, while the dashed line represents the line on which the points actually lie

A few entries in Tables 7–9 are flagged, however. Item pairs with correlated residuals $\geq|0.20|$ include items 5 and 22, 7 and 20, 10 and 17, 12 and 15, 12 and 19, and 16 and 17. Most of these results make sense. Items 5 and 22 ask students to infer the relative lifetimes of stars given their masses, while items 10 and 17 ask the reverse. Items 7 and 20 both probe the relationship between a star's temperature and color. Items 12, 15, and 19 all refer to fusion. The only item pair whose flagged Q3 statistic is not immediately explainable is

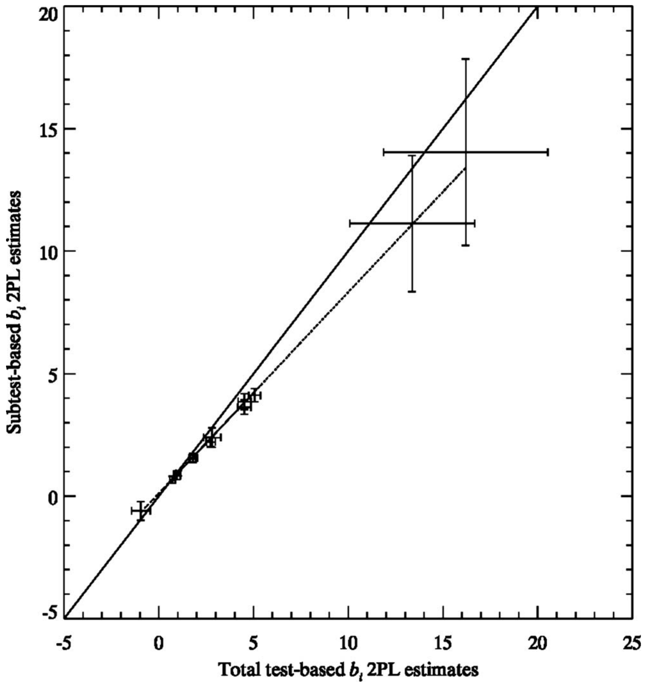

Figure 17. The 2PL difficulties (b_i) of the 13 star properties items estimated without the other ten items versus the difficulties estimated with the other ten items. Error bars represent standard errors. The solid line is where the points should lie if unidimensionality holds, while the dashed line represents the line on which the points actually lie

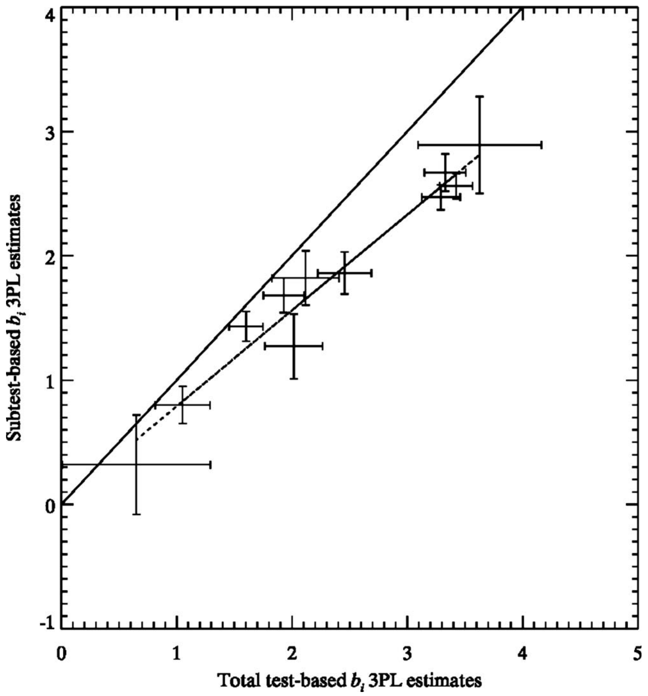

Figure 18. The 3PL difficulties (b_i) of the 13 star properties items estimated without the other ten items versus the difficulties estimated with the other ten items. Error bars represent standard errors. The solid line is where the points should lie if unidimensionality holds, while the dashed line represents the line on which the points actually lie

items 16 and 17. Item 16 asks students to compare the luminosity of two stars, while item 17 asks how a star's lifetime is related to its mass. Aside from this one instance, our cursory analysis of the content of the flagged items reveals ready explanations for their high Q3 statistics.

Our analysis above demonstrates one method for testing the assumption of unidimensionality and one method for testing the assumption of local independence. Researchers employing IRT models should use these or alternative methods to check whether or not the data supports IRT's assumptions before drawing any final conclusions from their IRT analysis.

8. COMPARING THE RASCH, 2PL, AND 3PL MODELS

In Secs. 4–7 we examined the SPCI using the Rasch, 2PL, and 3PL models. But amidst all the mathematical and statistical machinery operating in the body of this paper, let us not lose sight of our objective: We want to measure students' abilities. Have we succeeded?

We have certainly attached a lot of numbers to the SPCI's items and the ASTRO 101 students in our sample. If, like Stevens (1946), we define measurement as "the assignment of numerals according to rules," then everything we have done counts as measurement. Which model provides the best measures? Is it the model that best fits the data—in this case, the 3PL model? More fundamentally, is Stevens's definition an adequate conceptualization of measurement?

Social science researchers frequently look to the physical sciences to better understand the process of measurement. After all, the physical sciences possess theories of remarkable predictive power and generality. These theories are built on the foundation of measurement. Wright—who was a physicist by training and a psychometrician by practice—outlined a number of requirements for measures. According to Wright (1997), measures must be

1) unidimensional,
2) sample-independent,
3) invariantly comparable, and
4) additive.

Do the numbers generated by IRT models meet these requirements? We consider each of these in turn.

Table 7. Yen's Q3 statistic for each pair of items. Item parameters are estimated from the Rasch model. Cells with values ≥|0.20| are highlighted

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	1.00	−0.07	−0.06	0.06	−0.03	−0.04	−0.06	−0.05	−0.11	−0.06	0.03	0.05	0.03	−0.06	0.04	−0.11	−0.06	0.02	0.00	−0.03	−0.06	−0.06	−0.02
2	−0.07	1.00	−0.03	−0.06	0.02	−0.03	−0.03	−0.04	−0.04	−0.10	−0.06	−0.01	0.01	0.00	0.05	−0.04	−0.07	0.05	0.00	−0.08	−0.01	0.00	−0.02
3	−0.06	−0.03	1.00	−0.04	0.03	0.01	−0.06	−0.06	0.11	−0.04	−0.05	0.04	0.00	−0.01	0.03	−0.01	−0.06	−0.02	0.07	−0.04	−0.01	−0.02	−0.08
4	0.06	−0.06	−0.04	1.00	−0.05	−0.04	−0.11	0.00	−0.08	−0.06	−0.02	−0.09	−0.10	−0.07	−0.11	−0.10	−0.07	−0.08	−0.08	−0.06	−0.10	−0.02	0.02
5	−0.03	0.02	0.03	−0.05	1.00	−0.01	−0.02	−0.06	−0.02	0.08	0.05	0.10	−0.12	−0.06	0.08	−0.08	0.14	0.17	0.09	0.03	−0.01	**0.39**	−0.04
6	−0.04	−0.03	0.01	−0.04	−0.01	1.00	−0.07	−0.04	−0.02	−0.05	−0.04	−0.05	−0.03	0.00	−0.07	−0.10	−0.02	−0.03	−0.04	−0.11	−0.02	−0.02	−0.05
7	−0.06	−0.03	−0.06	−0.11	−0.02	−0.07	1.00	−0.09	−0.12	−0.08	−0.01	−0.08	−0.03	0.06	−0.10	0.03	−0.04	0.03	−0.07	**0.37**	0.09	−0.02	−0.02
8	−0.05	−0.04	−0.06	0.00	−0.06	−0.04	−0.09	1.00	0.02	−0.02	0.01	−0.02	0.02	0.08	−0.05	0.01	−0.06	−0.07	−0.03	−0.03	−0.03	−0.02	−0.06
9	−0.11	−0.04	0.11	−0.08	−0.02	−0.02	−0.12	0.02	1.00	−0.08	−0.03	−0.06	−0.07	0.01	−0.02	0.07	−0.03	−0.11	0.02	−0.01	0.00	0.00	−0.10
10	−0.06	−0.10	−0.04	−0.06	0.08	−0.05	−0.08	−0.02	−0.08	1.00	0.01	0.06	−0.07	0.02	0.08	−0.10	**0.43**	0.08	0.07	0.06	−0.12	0.09	0.05
11	0.03	−0.06	−0.05	−0.02	0.05	−0.04	−0.01	0.01	−0.03	0.01	1.00	0.09	−0.09	0.06	0.10	−0.17	0.06	0.01	0.08	0.01	−0.07	0.01	0.00
12	0.05	−0.01	0.04	−0.09	0.10	−0.05	−0.08	−0.02	−0.06	0.06	0.09	1.00	−0.02	0.05	**0.48**	−0.17	0.06	0.19	**0.22**	0.02	−0.01	0.14	0.02
13	0.03	0.01	0.00	−0.10	−0.12	−0.03	−0.03	0.02	−0.07	−0.07	−0.09	−0.02	1.00	0.04	−0.08	−0.01	−0.09	−0.08	−0.04	−0.08	−0.01	−0.05	−0.03
14	−0.06	0.00	−0.01	−0.07	−0.06	0.00	0.06	0.08	0.01	0.02	0.06	0.05	0.04	1.00	0.03	−0.02	−0.01	0.08	0.05	0.05	0.03	0.05	0.06
15	0.04	0.05	0.03	−0.11	0.08	−0.07	−0.10	−0.05	−0.02	0.08	0.10	**0.48**	−0.08	0.03	1.00	−0.18	0.00	0.14	0.20	−0.02	−0.08	0.09	−0.05
16	−0.11	−0.04	−0.01	−0.10	−0.08	−0.10	0.03	0.01	0.07	−0.10	−0.17	−0.17	−0.01	−0.02	−0.18	1.00	**−0.20**	−0.11	−0.13	0.00	0.08	−0.11	−0.08
17	−0.06	−0.07	−0.06	−0.07	0.14	−0.02	−0.04	−0.06	−0.03	**0.43**	0.06	0.06	−0.09	−0.01	0.00	**−0.20**	1.00	0.07	0.00	0.01	−0.12	0.12	0.09
18	0.02	0.05	−0.02	−0.08	0.17	−0.03	0.03	−0.07	−0.11	0.08	0.01	0.19	−0.08	0.08	0.14	−0.11	0.07	1.00	0.08	0.09	0.01	0.19	−0.01
19	0.00	0.00	0.07	−0.08	0.09	−0.04	−0.07	−0.03	0.02	0.07	0.08	**0.22**	−0.04	0.05	0.20	−0.13	0.00	0.08	1.00	0.02	−0.04	0.17	−0.03
20	−0.03	−0.08	−0.04	−0.06	0.03	−0.11	**0.37**	−0.03	−0.01	0.06	0.01	0.02	−0.08	0.05	−0.02	0.00	0.01	0.09	0.02	1.00	−0.02	0.09	0.00
21	−0.06	−0.01	−0.01	−0.10	−0.01	−0.02	0.09	−0.03	0.00	−0.12	−0.07	−0.01	−0.01	0.03	−0.08	0.08	−0.12	0.01	−0.04	−0.02	1.00	0.01	−0.02
22	−0.06	0.00	−0.02	−0.02	**0.39**	−0.02	−0.02	−0.02	0.00	0.09	0.01	0.14	−0.05	0.05	0.09	−0.11	0.12	0.19	0.17	0.09	0.01	1.00	−0.01
23	−0.02	−0.02	−0.08	0.02	−0.04	−0.05	−0.02	−0.06	−0.10	0.05	0.00	0.02	−0.03	0.06	−0.05	−0.08	0.09	−0.01	−0.03	0.00	−0.02	−0.01	1.00

Table 8. Yen’s Q3 statistic for each pair of items. Item parameters are estimated from the 2PL model. Cells with values ≥|0.20| are highlighted

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	1.00	−0.04	−0.02	0.07	−0.05	−0.03	−0.08	−0.02	−0.07	−0.12	0.00	0.02	0.07	−0.08	0.01	−0.07	−0.16	0.00	−0.01	−0.04	−0.06	−0.09	−0.03
2	−0.04	1.00	−0.02	−0.03	0.03	−0.01	0.00	−0.02	−0.01	−0.08	−0.03	0.01	0.01	0.02	0.07	−0.01	−0.04	0.06	0.02	−0.06	0.02	0.02	0.00
3	−0.02	−0.02	1.00	−0.01	0.05	0.04	−0.02	−0.04	0.13	−0.02	−0.03	0.06	0.01	0.00	0.05	0.02	−0.04	−0.01	0.09	−0.02	0.02	−0.01	−0.05
4	0.07	−0.03	−0.01	1.00	−0.05	−0.03	−0.11	0.03	−0.05	−0.09	−0.03	−0.10	−0.06	−0.06	−0.12	−0.05	−0.12	−0.08	−0.08	−0.06	−0.08	−0.03	0.02
5	−0.05	0.03	0.05	−0.05	1.00	−0.01	−0.04	0.00	0.05	0.03	0.01	0.06	−0.09	−0.09	0.05	−0.09	0.11	0.16	0.08	0.03	−0.03	**0.38**	−0.05
6	−0.03	−0.01	0.04	−0.03	−0.01	1.00	−0.05	−0.01	0.01	−0.05	−0.03	−0.04	0.00	0.01	−0.06	−0.06	−0.03	−0.02	−0.02	−0.09	0.00	−0.02	−0.03
7	−0.08	0.00	−0.02	−0.11	−0.04	−0.05	1.00	−0.05	−0.08	−0.15	−0.04	−0.12	0.02	0.04	−0.14	0.06	−0.15	0.00	−0.08	**0.35**	0.10	−0.05	−0.03
8	−0.02	−0.02	−0.04	0.03	0.00	−0.01	−0.05	1.00	0.06	0.02	0.05	0.06	0.03	0.10	0.01	0.02	−0.01	−0.01	0.04	0.02	0.01	0.06	−0.02
9	−0.07	−0.01	0.13	−0.05	0.05	0.01	−0.08	0.06	1.00	−0.04	0.01	0.02	−0.05	0.04	0.05	0.08	0.01	−0.04	0.09	0.05	0.04	0.08	−0.06
10	−0.12	−0.08	−0.02	−0.09	0.03	−0.05	−0.15	0.02	−0.04	1.00	−0.07	−0.02	−0.04	−0.04	0.01	−0.10	**0.35**	0.04	0.03	0.03	−0.16	0.03	0.02
11	0.00	−0.03	−0.03	−0.03	0.01	−0.03	−0.04	0.05	0.01	−0.07	1.00	0.04	−0.05	0.02	0.05	−0.14	−0.05	−0.03	0.06	−0.01	−0.08	−0.03	−0.02
12	0.02	0.01	0.06	−0.10	0.06	−0.04	−0.12	0.06	0.02	−0.02	0.04	1.00	0.02	0.00	**0.45**	−0.16	−0.02	0.15	**0.20**	0.00	−0.04	0.08	0.01
13	0.07	0.01	0.01	−0.06	−0.09	0.00	0.02	0.03	−0.05	−0.04	−0.05	0.02	1.00	0.06	−0.05	0.03	−0.04	−0.05	−0.02	−0.06	0.02	−0.02	0.00
14	−0.08	0.02	0.00	−0.06	−0.09	0.01	0.04	0.10	0.04	−0.04	0.02	0.00	0.06	1.00	−0.02	0.01	−0.09	0.05	0.02	0.03	0.03	0.01	0.05
15	0.01	0.07	0.05	−0.12	0.05	−0.06	−0.14	0.01	0.05	0.01	0.05	**0.45**	−0.05	−0.02	1.00	−0.17	−0.08	0.11	0.18	−0.03	−0.10	0.04	−0.06
16	−0.07	−0.01	0.02	−0.05	−0.09	−0.06	0.06	0.02	0.08	−0.10	−0.14	−0.16	0.03	0.01	−0.17	1.00	−**0.23**	−0.12	−0.12	0.00	0.12	−0.11	−0.07
17	−0.16	−0.04	−0.04	−0.12	0.11	−0.03	−0.15	−0.01	0.01	**0.35**	−0.05	−0.02	−0.04	−0.09	−0.08	−**0.23**	1.00	0.03	−0.04	−0.04	−0.18	0.08	0.05
18	0.00	0.06	−0.01	−0.08	0.16	−0.02	0.00	−0.01	−0.04	0.04	−0.03	0.15	−0.05	0.05	0.11	−0.12	0.03	1.00	0.08	0.09	−0.01	0.17	−0.01
19	−0.01	0.02	0.09	−0.08	0.08	−0.02	−0.08	0.04	0.09	0.03	0.06	**0.20**	−0.02	0.02	0.18	−0.12	−0.04	0.08	1.00	0.02	−0.04	0.16	−0.02
20	−0.04	−0.06	−0.02	−0.06	0.03	−0.09	**0.35**	0.02	0.05	0.03	−0.01	0.00	−0.06	0.03	−0.03	0.00	−0.04	0.09	0.02	1.00	−0.02	0.08	0.01
21	−0.06	0.02	0.02	−0.08	−0.03	0.00	0.10	0.01	0.04	−0.16	−0.08	−0.04	0.02	0.03	−0.10	0.12	−0.18	−0.01	−0.04	−0.02	1.00	−0.02	−0.02
22	−0.09	0.02	−0.01	−0.03	**0.38**	−0.02	−0.05	0.06	0.08	0.03	−0.03	0.08	−0.02	0.01	0.04	−0.11	0.08	0.17	0.16	0.08	−0.02	1.00	−0.03
23	−0.03	0.00	−0.05	0.02	−0.05	−0.03	−0.03	−0.02	−0.06	0.02	−0.02	0.01	0.00	0.05	−0.06	−0.07	0.05	−0.01	−0.02	0.01	−0.02	−0.03	1.00

Table 9. Yen's Q3 statistic for each pair of items. Item parameters are estimated from the 3PL model. Cells with values ≥|0.20| are highlighted

Item	1	2	4	5	6	7	8	9	10	11	12	14	15	16	17	18	19	20	21	22	23
1	1.00	−0.04	0.06	−0.04	−0.03	−0.09	−0.01	−0.07	−0.12	−0.01	0.03	−0.11	0.03	−0.06	−0.14	0.01	0.00	−0.05	−0.06	−0.09	1.00
2	−0.04	1.00	−0.02	0.00	−0.02	0.01	−0.04	−0.02	−0.09	−0.04	−0.02	0.02	0.06	−0.01	−0.04	0.03	−0.01	−0.08	0.02	−0.03	−0.04
4	0.06	−0.02	1.00	−0.04	−0.03	−0.15	0.04	−0.06	−0.10	−0.04	−0.10	−0.10	−0.09	−0.04	−0.13	−0.09	−0.06	−0.08	−0.09	−0.01	0.06
5	−0.04	0.00	−0.04	1.00	−0.01	−0.04	−0.07	0.02	−0.03	−0.03	−0.14	−0.15	−0.06	−0.02	0.07	0.01	−0.05	−0.08	−0.03	**0.22**	−0.04
6	−0.03	−0.02	−0.03	−0.01	1.00	−0.06	−0.01	0.01	−0.06	−0.03	−0.05	0.00	−0.06	−0.05	−0.03	−0.02	−0.02	−0.09	0.00	−0.03	−0.03
7	−0.09	0.01	−0.15	−0.04	−0.06	1.00	−0.05	−0.08	−0.17	−0.07	−0.15	−0.02	−0.15	0.07	−0.16	0.00	−0.09	**0.35**	0.09	−0.06	−0.09
8	−0.01	−0.04	0.04	−0.07	−0.01	−0.05	1.00	0.05	−0.01	0.03	−0.02	0.09	−0.04	0.06	−0.04	−0.07	−0.02	−0.01	0.01	−0.03	−0.01
9	−0.07	−0.02	−0.06	0.02	0.01	−0.08	0.05	1.00	−0.05	0.00	−0.03	0.02	0.01	0.09	0.01	−0.08	0.06	0.03	0.04	0.03	−0.07
10	−0.12	−0.09	−0.10	−0.03	−0.06	−0.17	−0.01	−0.05	1.00	−0.11	−0.12	−0.09	−0.04	−0.06	**0.34**	−0.03	−0.02	−0.03	−0.17	−0.06	−0.12
11	−0.01	−0.04	−0.04	−0.03	−0.03	−0.07	0.03	0.00	−0.11	1.00	−0.03	−0.03	0.02	−0.12	−0.06	−0.07	0.03	−0.06	−0.09	−0.10	−0.01
12	0.03	−0.02	−0.10	−0.14	−0.05	−0.15	−0.02	−0.03	−0.12	−0.03	1.00	−0.06	**0.38**	−0.10	−0.09	−0.02	0.06	−0.14	−0.05	−0.17	0.03
14	−0.11	0.02	−0.10	−0.15	0.00	−0.02	0.09	0.02	−0.09	−0.03	−0.06	1.00	−0.04	0.01	−0.13	0.01	−0.01	−0.03	0.01	−0.04	−0.11
15	0.03	0.06	−0.09	−0.06	−0.06	−0.15	−0.04	0.01	−0.04	0.02	**0.38**	−0.04	1.00	−0.10	−0.10	0.00	0.10	−0.13	−0.10	−0.10	0.03
16	−0.06	−0.01	−0.04	−0.02	−0.05	0.07	0.06	0.09	−0.06	−0.12	−0.10	0.01	−0.10	1.00	**−0.20**	−0.07	−0.06	0.04	0.13	−0.04	−0.06
17	−0.14	−0.04	−0.13	0.07	−0.03	−0.16	−0.04	0.01	**0.34**	−0.06	−0.09	−0.13	−0.10	**−0.20**	1.00	−0.02	−0.09	−0.10	−0.17	0.02	−0.14
18	0.01	0.03	−0.09	0.01	−0.02	0.00	−0.07	−0.08	−0.03	−0.07	−0.02	0.01	0.00	−0.07	−0.02	1.00	−0.05	0.00	−0.01	−0.03	0.01
19	0.00	−0.01	−0.06	−0.05	−0.02	−0.09	−0.02	0.06	−0.02	0.03	0.06	−0.01	0.10	−0.06	−0.09	−0.05	1.00	−0.06	−0.05	−0.01	0.00
20	−0.05	−0.08	−0.08	−0.08	−0.09	**0.35**	−0.01	0.03	−0.03	−0.06	−0.14	−0.03	−0.13	0.04	−0.10	0.00	−0.06	1.00	−0.04	−0.05	−0.05
21	−0.06	0.02	−0.09	−0.03	0.00	0.09	0.01	0.04	−0.17	−0.09	−0.05	0.01	−0.10	0.13	−0.17	−0.01	−0.05	−0.04	1.00	−0.03	−0.06
22	−0.09	−0.03	−0.01	**0.22**	−0.03	−0.06	−0.03	0.03	−0.06	−0.10	−0.17	−0.04	−0.10	−0.04	0.02	−0.03	−0.01	−0.05	−0.03	1.00	−0.09
23	−0.04	0.00	0.01	−0.10	−0.03	−0.05	−0.03	−0.07	−0.02	−0.05	−0.06	0.01	−0.10	−0.04	0.02	−0.06	−0.07	−0.03	−0.03	−0.09	−0.04

The first requirement, unidimensionality, is a fundamental assumption of the three IRT models present here. As we discussed in preceding sections, there is no guarantee that an instrument such as the SPCI will probe only a single construct. However, there are methods by which one can test whether or not unidimensionality holds.

Sample-independence reflects the idea that what we measure should not depend on what we use to measure it. If a person measures the length of a table, then her result should not depend on which meter stick she used. Likewise, the fidelity of the meter stick should not depend on which table is measured. In the context of tests, our measures of students' abilities should not depend on which items we use to measure their abilities and our calibration of our test items should not depend on the population of students to which they are administered (Wright 1997). In the case of IRT, parameter invariance encapsulates this idea.

What about invariant comparisons? This is precisely what Rasch (1960) worried about when he formulated his principle of specific objectivity:

> A person having a greater ability than another should have the greater probability of solving *any* item of the type in question, and similarly, one item being more difficult than another one means that for *any* person the probability of solving the second item correctly is the greater one. (p. 117, italics in original)

In other words, the difference in abilities between two people A and B ($\theta_A - \theta_B$) should be the same regardless of the items we use to measure those abilities. Likewise, the difference in the difficulties of two items m and n ($b_m - b_n$) should be the same regardless of who answers the items. This was the idea we used to motivate the Rasch model in Sec. 3. But look again at the ICCs for items 16 and 17 in Figures 1–3. In the 2PL and 3PL models, the ICCs cross. What does this mean? A student of low ability (say, $\theta_p = -3$) finds item 17 is harder than item 16. A student of high ability (say, $\theta_p = 3$) finds item 16 to be harder than item 17. Why should the relative difficulties of items change with ability? The 2PL and 3PL models do not say. (Incidentally, the fact that the relative order of items changes as a function of ability is another reason why Wright maps are not constructed for 2PL and 3PL parameter estimates.) Examining the content of these items offers few clues. We struggle to invent a story that explains why the lifetime-mass relationship for stars (the subject of item 17) should be harder than determining the luminosity of stars (the subject of item 16) for students of low ability, while exactly the opposite is true for students of high ability. This perplexing state of affairs is a general feature of the 2PL and 3PL models: ICCs regularly cross, thus changing the order of item difficulties as a function of test-taker ability (Wright 1997). Crossing ICCs violate specific objectivity. The Rasch model, with its constant discrimination parameter and its lack of a guessing term, is the only IRT model consistent with specific objectivity and thus the only IRT model that allows for invariant comparisons (Rasch 1960; Wright 1997). In other words, the Rasch model is the only IRT model out of the three we have considered that satisfies Wright's third criterion for measurement.

The final requirement, additivity, is the key to interval scales. Additivity reflects our capacity to combine objects end-to-end to form a larger object (Narens and Luce 1986; Borsboom 2005). For example, one can combine two or more meter sticks to measure the height of various objects. These additive combinations allow us to express how much of a property an object possesses (i.e., a measure of the height of a building can be formulated as the number of meter sticks we need to put end-to-end in order to equal the building's height). In the physical sciences, we work with measures that either permit such additive operations or are formed by combining measures that permit additive operations. But when we attempt to measure a latent psychological attribute, such as a student's ability, we run into a problem: What is the additive operation for this measure? Since such latent traits are not directly observable, psychometricians struggled for years to invent additive measures.

Such was the situation until Luce and Tukey's (1964) landmark paper introducing additive conjoint measurement. See also Borsboom (2005), Narens and Luce (1986), and Kyngdon (2008) for reviews of this idea. Conjoint measurement concerns two independent objects, X and Y, which combine to form a dependent variable Z. Additive conjoint measurement says X and Y can be simultaneously quantified such that changes in X can be expressed in terms of the changes in Y needed to maintain a constant value of Z, provided the elements of Z obey certain axioms (Luce and Tukey 1964; Narens and Luce 1986; Borsboom 2005). Measurement is thus conceptualized as a tradeoff: Given a change in Y, what change in X would have changed Z by the same amount (and vice versa)? Since differences in levels of X are matched to differences in levels of Y, X and Y are necessarily placed on an interval scale (Borsboom 2005). If one wants an interval scale for a measure, then one must show that measure obeys the axioms of conjoint measurement.

The Rasch model has many similarities to additive conjoint measurement (Perline, Wright, and Wainer 1979; Wright 1997; Embretson and Reise 2000; Kyngdon 2008). Specifically, a dependent variable (the log odds D_{pi} of person p giving a correct response to item i) is the difference between the person's ability θ_p and the item's difficulty b_i. Changing the difficulty of the item requires an equivalent change in the person's ability in order to maintain the same log odds of a correct response. The Rasch model appears to be consistent with conjoint measurement, even though Rasch formulated this model several years before conjoint measurement was explicated. Of course, such similarities are no guarantee that the Rasch model always achieves conjoint measurement. Whether or not one has additive conjoint measurement depends on whether or not its axioms are satisfied (Kyngdon 2008; Borsboom and Scholten 2008).

In the years since Luce and Tukey published their work, people have debated whether or not the Rasch model is an instance of conjoint measurement (Perline, Wright, and Wainer 1979; Wright 1997; Embretson and Reise 2000; Borsboom 2005; Kyngdon 2008; Michell 2008; Borsboom and Scholten 2008). Unfortunately, only a small number of studies have actually attempt to show that data described by the Rasch model also obey conjoint measurement's axioms (Karabastos 2001; Perline, Wright, and Wainer 1979; Embretson and Reise 2000). Discussing these axioms and investigating whether our SPCI data follows them are beyond the scope of this paper. We simply want to point out that specific objectivity is a necessary but not sufficient condition for conjoint measurement—meaning that the 2PL and 3PL models are not instances of conjoint measurement and, therefore, cannot provide interval scales.

What about the fact that the 2PL and 3PL models fit the data better than the Rasch model? This is not necessarily a virtue. The 2PL and 3PL models are practically guaranteed to fit the data better than the Rasch model since they have more adjustable parameters. This means the 2PL and 3PL models can fit a wider variety of data—but is this a good quality? Masters (1988) described how discriminations that change from item to item may actually be due to testing populations of students who differ in 1) their opportunities to learn the material probed by the item, 2) their test wiseness, or 3) the speed at which they work through the test. These three forms of bias are confounds that may hinder our efforts to measure the construct in which we are interested. What about the guessing parameter? Given that concept inventories, such as the SPCI, are multiple choice tests, should we not include such a parameter? At first, we might suspect that a guessing parameter is needed since even examinees of very low ability could potentially select a correct answer by merely guessing. But at another level, this seems anathema to the very idea of a concept inventory. As we mentioned earlier, the items and the answer choices that appear on concept inventories are based on research into common student difficulties. A good item should therefore have one or more plausible distractors that students of low ability are drawn to. From this perspective, and as noted in Sec. 4, students of low ability should not have a significantly nonzero probability of giving the correct answer to any item (Sadler *et al.* 2010). The discrimination and guessing parameters may actually obscure potential problems. Data that do not fit the Rasch model might be a symptom of problems with the test or the testing procedures.

The Rasch, 2PL, and 3PL models do not only differ in the forms of their equations; they also differ with respect to the underlying goals of measurement (Andrich 2004; Michell 2008; Wright 1997). Proponents of the 2PL and 3PL models argue that the Rasch model is often too simple to describe real data. They use the 2PL or 3PL models because they do a better job of describing the data (Andrich 2004). Proponents of the Rasch model counter that the 2PL and 3PL models may hide problems that the researcher should be aware of (Andrich 2004; Masters 1988; Wright 1997). The 2PL and 3PL models also cannot be used to conduct conjoint measurement and construct interval scales (Wright 1997). Rasch model advocates argue that we need to find data that fits the Rasch model—otherwise, despite all of our mathematical manipulations, we are not actually *measuring* anything (Andrich 2004; Michell 2008; Wright 1997). Wright (1997) summarized this position:

> There is a vast difference between gerrymandering whatever kind of model might seem to give a locally good description of some transient set of data and searching, instead, for the kind of data that can yield inferentially stable—that is, generalizable—meaning to the parameter estimates of interest. The 3P model is data driven: The model must fit, or another model must be found. The 3P model seldom objects to an item, no matter how badly it functions. The Rasch model is theory driven: The data must fit, or else better data must be found. Indeed, it is the search for better data that sets the stage for discovery. The only way discovery can occur is as an unexpected discrepancy from an otherwise stable frame of reference. When we study data misfit to the Rasch model, we discover new things about the nature of what we are measuring and the way that people are able to tell us about it in their responses. These discoveries are important events that strengthen and clarify our construct as well as our ability to measure it. (p. 43)

Researchers who design and/or use concept inventories should carefully consider these issues since they get to the heart of the nature of measurement. If we are inarticulate about what we want to measure and the criteria by which something is considered a measurement, then how can we have any confidence in our capacity to measure students' abilities?

9. SUMMARY AND CONCLUSIONS

What information should readers take away from this paper? We emphasize three separate "take home" messages.

First, this paper is meant to provide astronomy education researchers an introduction to the theory and methods of IRT. We motivated the need for IRT by highlighting some weaknesses with CTT and traditional learning gain calculations in Sec. 2. We introduced the basics of IRT along with the Rasch, 2PL, and 3PL models in Sec. 3. We devoted the majority of this paper (Secs. 4–7) to demonstrating how one can apply IRT by using the Rasch, 2PL, and 3PL models to analyze the SPCI. Our discussion emphasized the necessity of checking whether a given IRT model fits the data and whether the assumptions of IRT hold. *IRT is not a panacea that can be applied to any set of data to cure all imperfections; if one wants to leverage the strengths of IRT over other models, such as CTT, then one must ensure that its assumptions hold true and the model fits.* We acknowledge that our discussion of IRT is necessarily brief. We advise readers searching for more detail to consult the foundational works of IRT (Lord and Novick 1968; Rasch 1960) as well as subsequent pedagogical treatments (Embretson and Reise 2000; Hambleton and Jones 1993; Harris 1989; Ding and Beichner 2009).

Second, we used IRT to investigate the SPCI. Our investigation highlights a number of important points. First, the SPCI contains some problematic items (e.g., items 3 and 13). Item 3 has already been removed from the current version of the SPCI and, as a result of our analysis, item 13 is now a candidate for revision. Second, the difficulties of many SPCI items appear mismatched with the majority of students' abilities, both pre- and post-instruction. This is best demonstrated by looking at the Wright map in Figure 11, which shows there are many items with logit values significantly higher than the logit values of students' post-instruction abilities. Future iterations of the SPCI should include more items with lower difficulty values so that we can more accurately measure the abilities of students at the lower end of the distribution. Third, the efficacy of many of the distractors on the SPCI should be re-evaluated since our 3PL analysis revealed that even students of low abilities have significant (20–25%) probabilities of guessing the correct answer. Fourth, we must carefully re-examine all items that do not fit each IRT model. Such item misfit investigations have the potential to uncover previously unsuspected problems with individual items. For example, Planinic (2006) and Planinic, Ivanjek, and Susac (2010) discussed the role misfit investigations play in detecting problems with the *Force Concept Inventory* (Hestenes, Wells, and Swackhamer 1992) and the *Conceptual Survey of Electricity and Magnetism* (Maloney *et al.* 2001). Finally, the SPCI's apparent departures from unidimensionality and violations of local independence should be further investigated in order to improve future versions of the test. Alternatively, we may reanalyze the SPCI with one or more multidimensional IRT models (Ackerman, Gierl, and Walker 2003; Briggs and Wilson 2003). In general, we do not consider the problems we uncovered with the SPCI as constituting any sort of failure. After all, developing an instrument to measure students' mental processes is necessarily a complex and time-intensive endeavor; why should we expect perfection early on? We think the issues we uncovered using IRT will help us learn more about this important instrument and offer guidance toward ongoing revisions and refinements.

Finally, we hope we can jump-start a debate in the AER community about the nature and goals of measurement. Concept inventories are typically used to measure students' learning gains—but what does measuring entail? Is simply assigning numbers according to rules *à la* Stevens (1946) a sufficient condition for measurement? At the very least, Stevens (1946) warns us to carefully consider the type of scale we generate by our number-assignment rules. Depending on the scale, one cannot perform some mathematical operations and make certain inferences. However, if we want measures that share properties with measures in the physical sciences, then we must move beyond Stevens's conceptualization of measurement. We must worry about properties such as specific objectivity (Rasch 1960) and satisfying the axioms of conjoint measurement (Luce and Tukey 1964). These requirements restrict us to the Rasch model (Wright 1997). Yet such a restriction is required if we want interval scales and the inferential and comparative capabilities they bring.

What are some future steps the AER community can take? Certainly, other concept inventories should be investigated using IRT models. These investigations will help our community develop better instruments for measuring students' abilities. If researchers decide that they need instruments that yield interval scales, then more ambitious studies should test whether or not the axioms of conjoint measurement hold. Testing conjoint measurement would be a fascinating and badly needed study in the field of psychometrics in general, since only a few studies have applied its axioms to real data (Embretson and Reise 2000; Perline, Wright, and Wainer 1979).

Of course, applying IRT may be excessive for some studies. Depending on the goals and nature of the study, simpler procedures, such as those from CTT, may provide sufficient information. However, these procedures must be explicitly defended within the context of the goals and nature of the study. If we do not carefully consider what we are trying to measure, then we may select methods that are either too simplistic to provide us with the information we desire or are more complex than is necessary. If we make such errors, especially the former, then critics may justifiably question what, if anything, we are measuring.

Acknowledgments

We owe a special thanks to Derek Briggs, from whom we learned IRT and who drew our attention to many of the fundamental issues in theories of measurement. This article would not have been possible without his support and guidance. Ed Prather and Doug Duncan also contributed helpful comments and feedback. This material is based in part upon work supported by the National Science Foundation under Grant Nos. 0833364 and 0715517, a CCLI Phase III Grant for the Collaboration of Astronomy Teaching Scholars (CATS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Ackerman, T. A., Gierl, M. J., and Walker, C. M. 2003, “Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests,” *Educ. Meas.: Issues & Pract.*, 22, 37–51.

Allen, K. 2007, “Getting More from Your Data: Application of Item Response Theory to the Statistics Concept Inventory,” in *2007 ASEE Annual Conference and Exposition Proceedings*, Honolulu, HI.

Andersen, E. B. 1977, “Sufficient Statistics and Latent Trait Models,” *Psychometrika*, 42, 69–81.

Andrich, D. 2004, “Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms?,” *Med. Care*, 42, I7–16.

Bailey, J. M. 2007, “Development of a Concept Inventory to Assess Students’ Understanding and Reasoning Difficulties about the Properties and Formation of Stars,” *Astron. Educ. Rev.*, 6, 133–9.

Bailey, J. M. 2009, “Concept Inventories for ASTR0 101,” *Phys. Teach.*, 47, 439–41.

Baker, F. B. and Kim, S. 2004, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., New York: Dekker.

Bao, L. 2006, “Theoretical Comparisons of Average Normalized Gain Calculations,” *Am. J. Phys.*, 74, 917–22.

Bardar, E. M., Prather, E. E., Brecher, K., and Slater, T. F. 2007, “Development and Validation of the Light and Spectroscopy Concept Inventory,” *Astron. Educ. Rev.*, 5, 103–13.

Bejar, I. I. 1980, “A Procedure for Investigating the Unidimensionality of Achievement Tests Based on Item Parameter Estimates,” *J. Educ. Meas.*, 17, 283–96.

Bereiter, C. 1963, “Some Persisting Dilemmas in the Measurement of Change,” in *Problems in Measuring Change*, ed. C. W. Harris, Madison, WI: The University of Wisconsin Press, pp. 3–20.

Borsboom, D. 2005, *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*, New York: Cambridge University Press.

Borsboom, D. and Scholten, A. Z. 2008, “The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics,” *Theory & Psych.*, 18, 111–7.

Briggs, D. C. and Wilson, M. 2003, “An Introduction to Multidimensional Measurement Using Rasch

Models," *J. App. Meas.*, 4, 87–100.

Brogt, E., Sabers, D., Prather, E. E., Deming, G. L., Hufnagel, B., and Slater, T. F. 2007, "Analysis of the Astronomy Diagnostic Test," *Astron. Educ. Rev.*, 6, 25–42.

Crocker, L. and Algina, J. 1986, *Introduction to Classical and Modern Test Theory*, Orlando, FL: Harcourt Brace Jovanovitch.

Cronbach, L. J. and Furby, L. 1970, "How We Should Measure 'Change'—Or Should We?," *Psych. Bull.*, 74, 68–80.

Ding, L. and Beichner, R. 2009, "Approaches to Data Analysis of Multiple-Choice Questions," *Phys. Rev. ST: Phys. Educ. Res.*, 5, 020103.

Ding, L., Chabay, R., Sherwood, B., and Beichner, R. 2006, "Evaluating an Electricity and Magnetism Assessment Tool: Brief Electricity and Magnetism Assessment," *Phys. Rev. ST: Phys. Educ. Res.*, 2, 010105.

Embretson, S. E. and Reise, S. P. 2000, *Item Response Theory for Psychologists*, Mahwah, NJ: Erlbaum.

Fischer, G. H. 1995, "Derivations of the Rasch Model," in *Rasch Models: Foundations, Recent Developments, and Applications*, eds. G. H. Fischer and I. W. Molenaar, New York, NY: Springer-Verlag, pp. 15–38.

George, D. and Mallery, P. 2009, *SPSS for Windows Step by Step: A Simple Guide and Reference*, Boston, MA: Pearson Education.

Hake, R. R. 1998, "Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *Am. J. Phys.*, 66, 64–74.

Hambleton, R. K. and Jones, R. J. 1993, "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development," *Educ. Meas.: Issues & Pract.*, 12, 253–62.

Harris, D. 1989, "Comparison of 1-, 2-, and 3-Parameter IRT Models," *Educ. Meas.: Issues & Pract.*, 8, 157–63.

Herrmann-Abell, C. F., DeBoer, G. E., and Roseman, J. E. 2009, "Using Rasch Modeling to Analyze Standards-Based Assessment Items Aligned to Middle School Chemistry Ideas," Poster presented at the DR-K12 PI Meeting (Washington, D.C.).

Hestenes, D., Wells, M., and Swackhamer, G. 1992, "Force Concept Inventory," *Phys. Teach.*, 30, 141–58.

Holland, P. W. 1990, "On the Sampling Theory Foundations of Item Response Theory Models," *Psychometrika*, 55, 577–601.

Karabastos, G. 2001, "The Rasch Model Additive Conjoint Measurement and New Models of Probabilistic Measurement Theory," *J. Appl. Meas.*, 2, 389–423.

Keller, J. M. 2006, "Development of a Concept Inventory Addressing Students' Beliefs and Reasoning Difficulties Regarding the Greenhouse Effect," Ph.D. thesis, University of Arizona.

Kyngdon, A. 2008, "The Rasch Model from the Perspective of the Representational Theory of Measurement," *Theory & Psych.*, 18, 89–109.

Lee, Y., Palazzo, D. J., Warnakulasooriya, R., and Pritchard, D. E. 2008, "Measuring Student Learning with Item Response Theory," *Phys. Rev. ST: Phys. Educ. Res.*, 4, 010102.

Libarkin, J. C. and Anderson, S. W. 2005, "Assessment of Learning in Entry-level Geoscience Courses: Results from the Geoscience Concept Inventory," *J. Geo. Educ.*, 53, 394–401.

Linacre, J. M. 2005, "Measurement, Meaning and Morality," MESA Research Memorandum No. 71. Available at: www.rasch.org/memo71.pdf.

Lindell, R. S. 2001, “Enhancing College Students’ Understanding of Lunar Phases,” Ph.D. thesis, University of Nebraska.

Lord, F. M. 1980, *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Erlbaum.

Lord, F. M. and Novick, M. R. 1968, *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Luce, R. D. and Tukey, J. W. 1964, “Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement,” *J. Math. Psych.*, 1, 1–27.

Maloney, D. P., O’Kuma, T. L., Hieggelke, C. J., and van Heuvelen, A. 2001, “Surveying Students’ Conceptual Knowledge of Electricity and Magnetism,” *Am. J. Phys.*, 69, S12–23.

Marshall, J. A., Hagedorn, E. A., and O’Connor, J. 2009, “Anatomy of a Physics Test: Validation of the Physics Items on the Texas Assessment of Knowledge and Skills,” *Phys. Rev. ST: Phys. Educ. Rev.*, 5, 010104.

Marx, J. D. and Cummings, K. 2007, “Normalized Change,” *Am. J. Phys.*, 75, 87–91.

Masters, G. N. 1988, “Item Discrimination: When More Is Worse,” *J. Educ. Meas.*, 25, 15–29.

Masters, G. N. 2001, “The Key to Objective Measurement,” MESA Research Memorandum No. 70. Available at: www.rasch.org/memo70.pdf.

Michell, J. 2008, “Conjoint Measurement and the Rasch Paradox: A Response to Kyngdon,” *Theory & Psych.*, 18, 119–24.

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T., and McCauley, V. 2006, “Testing the Test: Item Response Curves and Test Quality,” *Am. J. Phys.*, 74, 449–53.

Narens, L. and Luce, R. D. 1986, “Measurement: The Theory of Numerical Assignments,” *Psych. Bull.*, 99, 166–80.

Orlando, M. and Thissen, D. 2000, “Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models,” *App. Psych. Meas.*, 24, 50–64.

Pek, P. and Poh, K. 2000, “Framework of a Decision-Theoretic Tutoring System for Learning of Mechanics,” *J. Sci. Educ. & Tech.*, 9, 343–56.

Perline, R., Wright, B. D., and Wainer, H. 1979, “The Rasch Model as Additive Conjoint Measurement,” *Appl. Psych. Meas.*, 3, 237–55.

Planinic, M. 2006, “The Rasch Model-Based Analysis of the Conceptual Survey of Electricity and Magnetism,” in *Proceedings of GIREP Conference 2006: Modeling in Physics and Physics Education*, eds. D. van den Berg and T. Ellermeijer, Amsterdam, NL: University of Amsterdam, pp. 133–134.

Planinic, M., Ivanjek, L., and Susac, A. 2010, “Rasch Model Based Analysis of the Force Concept Inventory,” *Phys. Rev. ST: Phys. Educ. Res.*, 6, 010103.

Prather, E. E., Rudolph, A. L., Brissenden, G., and Schlingman, W. 2009, “A National Study Assessing the Teaching and Learning of Introductory Astronomy. Part I. The Effect of Interactive Instruction,” *Am. J. Phys.*, 77, 320–30.

Rasch, G. 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago, IL: University of Chicago Press.

Rogosa, D. R. and Willett, J. B. 1983, “Demonstrating the Reliability of the Different Score in the Measurement of Change,” *J. Educ. Meas.*, 20, 335–43.

Rupp, A. A. and Zumbo, B. D. 2006, “Understanding Parameter Invariance in Unidimensional IRT Models,” *Educ. & Psych. Meas.*, 66, 63–84.

Sadler, P. M. 1998, "Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-Driven Assessment Instruments," *J. Res. Sci. Teach.*, 35, 265–96.

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., and Gould, R. R. 2010, "The Astronomy and Space Science Concept Inventory: Development and Validation of Assessment Instruments Aligned with the K-12 National Science Standards," *Astron. Educ. Rev.*, 8, 010111.

Stevens, S. S. 1946, "On the Theory of Scales of Measurement," *Science*, 103, 677–80.

Thompson, B. 2003, "Understanding Reliability and Coefficient alpha, Really," in *Score Reliability*, ed. B. Thompson, Thousand Oaks, CA: SAGE, pp. 3–30.

Vogt, W. P. 2007, *Quantitative Research Methods for Professionals*, Boston, MA: Pearson Education.

Wang, J. and Bao, L. 2010, "Analyzing Force Concept Inventory with Item Response Theory," *Am. J. Phys.* 78, 1064–1070.

Whitely, S. E. and Dawis, R. V. 1974, "The Nature of Objectivity with the Rasch Model," *J. Educ. Meas.*, 11, 163–78.

Wilson, M. 2005, *Constructing Measures: An Item Response Modeling Approach*, Mahwah, NJ: Erlbaum.

Wright, B. D. 1997, "A History of Social Science Measurement," *Educ. Meas.: Issues & Pract.*, 16, 33–45.

Wright, B. D. and Linacre, J. M. 1989, "Observations are Always Ordinal; Measurements, However, Must Be Interval," *Archiv. Phys. Med. & Rehab.*, 70, 857–60.

Wu, M. and Adams, R. J. 2010, "Properties of Rasch Residual Fit Statistics," *J. Appl. Meas.*, (in press).

Yen, W. M. 1984, "Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model," *Appl. Psych. Meas.*, 8, 125–45.

Yen, W. M. and Fitzpatrick, A. R. 2006, "Item Response Theory," in *Educational Measurement*, 4th Ed., ed. R. Brennan, Westport, CT: American Council on Education/Praeger, pp. 111–153.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. 1996, *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*, Chicago, IL: Scientific Software.