International Journal of Teaching and Learning in Higher Education 2011, Volume 23, Number 3, 350-362
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Modern Measurement Information Graphics for Understanding Student Performance Differences

Kent Rittschof and Wendy Chambers
Georgia Southern University

We present an example analysis and corresponding information graphics of data from a cognitive ability assessment as a means to illustrate the use of a Rasch measurement approach and advantages inherent in such an approach for a wide variety of teaching and learning investigations. The importance of placing measurements of student performances and measurements of assessment item difficulties on the same scale is demonstrated through the use of the information graphics. The possibilities for teacher-scholars to begin including basic Rasch analysis and graphics within studies of students are highlighted. Improved understanding of the relationships between student performances and the validity of instruments used to assess those performances is emphasized. The importance of key measurement principles, as illustrated with an ability assessment, is discussed in relation to potential application with classroom assessment of learning and survey assessment.

Introduction

Studies of teaching and learning in higher education or professional settings frequently make use of assessments that quantify learners' knowledge levels, abilities, motivations, and perspectives. The conclusions regarding students that teacher-scholars draw from such studies are often affected by the diversity among participants examined. Conclusions can also be affected by the validity of the instruments used to collect data and the analytic approach used to determine measures that allow meaning to be drawn from data. With these fundamental influences in mind we present an example study of college student abilities that illustrates the use of a modern measurement approach known as the Rasch model (Rasch, 1980), as well as the benefits inherent in such an approach for visualizing data across a wide variety of teaching and learning studies. Through this example we emphasize the need to scrutinize the functioning of the instruments as a means to improved understanding of the implications from investigations. Furthermore, we hope to encourage teacher-scholars who are unfamiliar with the procedures described here by suggesting appropriate software tools and resources.

An Example Study of Student Differences

The context of the illustrative investigation will be briefly characterized here to clarify purpose and help stimulate thinking about relevance to instructional and learning studies generally. First, some of the theoretical issues that interested us centered on how students' differences in visuospatial ability were distributed for a particular group of our students. We also wanted to determine whether our selected instrument was an appropriate tool for our population of students. That is, we first asked whether our cohort of students was primarily similar to one another or primarily different from one another in visuospatial ability, and second, whether the quality and difficulty level of the assessment was a good match for most of our students in this program. This study was considered part of an analysis of student characteristics to inform instructional and curriculum design for this type of student cohort.

To address these issues we selected a widely used instrument to assess cognitive visuospatial ability. Although many assessment instruments could be used we selected a particular instrument for its appropriateness in illustrating key fundamental issues and for its long history of use in research within higher education and other settings (Rittschof, 2010; Witkin, Oltman, Raskin, & Karp, 1971). Although we focused on an ability instrument, many of the general principles described are relevant to classroom assessment of learning and survey assessment. For example, by using this instrument we can focus on' (a) the issues of item difficulty, which are relevant to classroom assessments of specific content areas; and (b) the issues of student differences, which are relevant to deeper understanding of students as learners.

The purpose of using a Rasch model approach as part of the data analysis was to consider the findings relative to individual students, the sample of students, the instrument's individual items, and the instrument as a whole. Further explanation of this rationale for using a Rasch approach will follow.

Visuospatial Ability

Current psychological research on the architecture of the human mind often involves the components of working memory such as those dealing with the visuospatial processes (Baddeley, 1999). Investigations into the working memory's visuospatial processes are important for improved understanding of human perception and learning, as imagery-based information

is used increasingly within contemporary instructional contexts. One approach to examining visuospatial processes involves the administration of tests that require perceptual disembedding, i.e., visual locating, of simple shapes from within more complex shapes (Miyake, Witzki, & Emerson, 2001). A frequently used test of perceptual disembedding, commonly referred to as field dependence-independence (FDI), is the Group Embedded Figures Test (GEFT; Witkin, Oltman, Raskin, & Karp, 1971). The use of tests such as GEFT in psychological and instructional studies has been common since the 1960's and has continued to the present (Zhang, 2004).

Examples of recent applied studies dealing with learning, training, and visuospatial ability include those that have focused on problem solving with text and visual instruction (Angeli, & Valanides, 2004), web-based learning (Chen, & Macredie, 2004), and training needs among astronauts for improvement of 3-dimensional spatial orientation skills (Richards, Oman, Shebilske, Beall, Liu, & Natapoff, 2002). Over several decades of studies assessing students with the GEFT instrument, higher GEFT performance has repeatedly been associated with cognitive and learning advantages in a variety of content domains and instructional settings (Rittschof, 2010). Furthermore the GEFT was shown to have a reliability of r = 0.89 over 3 years for males and females (Witkin et al., 1971). It is worth noting that while numerous studies over the years have mislabeled the GEFT instrument as a test of cognitive style, here we build upon empirical investigations (e.g., MacLeod, Jackson, & Palmer, 1986; Miyake, Witzki, & Emerson, 2001; Zhang, 2004) that have confirmed GEFT to be more accurately classified as a test of cognitive ability and not a test of style.

Modern Measurement

The Rasch model is actually a collection, or family, of contemporary measurement models (Wright & Mok, 2004) for determining properties of instruments and data in human research. Appropriate use of Rasch measures and diagnostic tools represents application of a modern paradigm and can lead to substantive differences in the interpretations of investigation outcomes when compared with classical test theory methods (Andrich, 2004). For example, when only raw scores and corresponding percentages are used, scores do not reflect the differences in difficulty among test items. Use of raw scores rather than constructed measures can lead to the inaccurate assumption that point or percent differences among students are of the same magnitude at the low end or mid range of performances as they are at the high end of performances, for instance. In contrast, scaled measures provide advantages for comparing scores of people and assessment items because the student performance measurement values and the item difficulty measurement values are placed on a common scale. Two commonly used example members of this family of Rasch models are referred to as the dichotomous model and the rating scale model. These models allow measurement scaling of student differences by using raw ordinal scores from assessments to construct the scaled scores as interval level measures.

When assessing student abilities, the Rasch model allows student ability and an assessment item difficulty parameter to exist on the same measurement scale, thus allowing them to be directly comparable. Using the Rasch model the probability of a correct response can be determined as a function of the difference between the measured ability of the student and a difficulty parameter of the item in question. For instance, when an item and an examinee both have the same Rasch measure, this will mean that the person has a 50% probability of scoring correctly on that item.

The Rasch model can be used with the relatively modest participant sample sizes (e.g., 50 to 200 students) that are of interest in many studies of teaching and learning where measurement of individual student performances and item characteristics is desired. In addition, the Rasch approach can allow examination of validity for both student groups and individual students, even when some data are missing. For mathematical descriptions of Rasch analyses and comparisons to different analytic models, Smith and Smith (2004) provide a comprehensive and readable resource.

Method

Participants

University students (N = 114) at the sophomore level attending a medium-size university in the Southeastern United States volunteered as an optional activity within a teacher education prerequisite course. Participants were primarily female (approximately 85%) between the approximate ages of 19 to 22 years.

Instruments

The Group Embedded Figures Test (GEFT; Witkin et al., 1971) assesses visuospatial ability using 18 items that each require visually locating, or disembedding, specific simple shapes from within larger complex shapes, then correctly tracing the outline of the embedded simple shapes. Simple shapes include the outlines of a hexagon, a rectangular prism, and a cross, as well as the outlines of shapes resembling a simple house, a necktie, a letter *t*, and the lower right half of a picture frame.

Procedure

Experimenters administered the GEFT in classroom settings. The procedure included approximately 5 minutes for the instruction and practice section, then 10 additional minutes for completion of the two sections of the test.

Analyses

Rasch dichotomous model procedures were used on examinee scores in order to determine measurement properties of the GEFT instrument and of the student performances. The dichotomous model is appropriate for assessments such as GEFT in which items are scored as correct or incorrect. Measurement properties of interest that are addressed by the Rasch procedures include *additivity, unidimensionality,* and *invariance*. Additivity refers to a measure that approximates an interval scale so values can be added meaningfully, for instance. Unidimensionality refers to the single construct the instrument is measuring. The construct should closely approximate a single identifiable dimension or domain rather than many dimensions or domains. This construct dimension is often referred to as a *latent trait* whereby this trait directly influences examinee responses to the items designed to measure that trait (Reise, Ainsworth, & Haviland, 2005). Invariance refers to the need for measurement scales to not differ excessively on the construct with different situations or groups. That is, the scale should be a reliable metric for various categories of people on the construct of interest.

The Winsteps computer program (Linacre, 2006b) was used for the Rasch dichotomous model analysis. Winsteps was selected for its functionality, its compatibility with other data formats, its comparative low cost, and its worldwide availability. Microsoft Excel was used in conjunction with Winsteps to generate some of the graphics.

Student results were reported on the following measures: scaled ability measures, standard errors, and a measure that indicates how well each student fits with overall expected responding when compared the other students. Assessment instrument item outcomes were also indicated by scaled difficulty measures, standard error, and fit with the other items. By placing both student outcomes and item outcomes on an identical scale, students and items were directly and meaningfully compared for greater understanding of group, student, instrument, and item performance. Graphic illustrations will be used to support connections between students and items.

Results

A Scale of Performance: Person Ability Measures

The range and distribution characteristics of scores were of interest as we began to understand how individuals performed. The measurement scaling of those raw scores allowed for the examination of interval measures, as opposed to ordered quantities that are frequently used in traditional test score analyses. For example, comparisons among individuals can account for the fact that a one point raw score difference among high scorers can mean a larger measured difference than a 1 point raw score difference among average scorers due to typical variations in item difficulties. Thus, after constructing measures, differences among groupings of scores along the distribution were more meaningfully compared than were differences from raw scores or corresponding percentages.

Four participants had extreme scores, suggesting they were not a suitable match with the test. Of these, three scored the maximum of 18 correct. Thus, abilities of the three high scorers could not be estimated specifically because the test was too easy for them, not unlike many testing situations. This finding has implications for possible revision of the test. At the other extreme one student scored the minimum of 0 correct on the GEFT. It should be noted that manual scoring of each participant's test revealed that low scores were not simply due to participants leaving all or most items blank. That is, all participants attempted items throughout the test.

Although eliminating such outliers from further analysis is often appropriate depending upon one's purpose, we retained these outliers for our primary analysis as the responses appeared valid and useful for this illustration. Table 1 shows selected examples of person statistics ranging from the highest scorers to the lowest. The mean raw score was 10.5 out of 18. These raw scores were scaled using a Rasch procedure that yields log odds units known as logits. Scaled ability measures varied from -4.66 to 4.74 logits with the mean score set at zero. By comparison, removing the four outliers led to a range of measures from -3.36 to 3.42 logits. The fourth column of Table 1 shows examples of the student measures. Examinees who scored 9 out of 18, which is near the midpoint, were 0.29 logits lower on the scale than those who scored 10 out of 18. On the other hand, those who scored 17 out of 18 were 1.32 logits lower on the scale than those who scored 18 out of 18. These logit differences illustrate the distinction between raw scores versus constructed measures discussed previously.

A reliability estimate of 0.85 was also calculated using Cronbach's alpha procedure with the students' scores. This reliability level supported the favorable internal consistency of the assessment with this sample of students.

Patterns of Expected Performance: Person Pathway Plots

One important reason for placing student scores and items on the same scale is to examine the relationship

Table 1
GEFT Instrument Rasch Person Statistics in Descending Measure Order

Entry Number	Total Score	Count	Measure	Model S.E.	Infit ZSTD
18	18.0	18	4.74	1.86	Maximum Estimated Measure
67	18.0	18	4.74	1.86	Maximum Estimated Measure
77	18.0	18	4.74	1.86	Maximum Estimated Measure
13	17.0	18	3.42	1.08	0.2
31	17.0	18	3.42	1.08	0.2
. . .	. . .	. . .	. . .	. . .	. . .
11	12.0	18	0.87	0.57	0.9
. . .	. . .	. . .	. . .	. . .	. . .
10	10.0	18	0.27	0.54	-1.3
. . .	. . .	. . .	. . .	. . .	. . .
3	9.0	18	-0.02	0.53	2.3
. . .	. . .	. . .	. . .	. . .	. . .
8	4.0	18	-1.58	0.62	-0.4
. . .	. . .	. . .	. . .	. . .	. . .
62	2.0	18	-2.54	0.79	-0.5
. . .	. . .	. . .	. . .	. . .	. . .
97	1.0	18	-3.36	1.06	0.4
25	0.0	18	-4.66	1.86	Minimum Estimated Error
Mean	10.5	18	0.55	0.68	0.0
S.D.	4.3	0	1.69	0.27	1.0

Note. Values shown represent only an illustrative sample of statistics from across the distribution of 114 students.

between ability and item difficulty. Two types of analyses are useful for this purpose: error and accuracy. The amount of error associated with each student is important for placing student scores in proper context. Error is influenced by a student's score relative to the number of items that have measured difficulty levels near that student's score. That is, more items with difficulty levels near the student's ability level typically decrease error. This reduction in error occurs because each assessment point at or near a student's ability level can add some reliability to the overall measure.

Standard error was calculated for each student, as shown in column 5 of Table 1. Error is also reported in logits and corresponds with each different student measure. Thus, error can be added and subtracted from each measure to yield a range in which each student's measure falls. For example, examinee #13's measure would fall between 2.34 and 4.50. Notice that extreme measures such as those of examinee #18 and examinee #25 have the greatest error due to the smaller number of items represented at the extremes.

In addition to error, accuracy can be examined with respect to the likelihood that a student's responses tend to fit with expectations. These expectations are based on the difficulty levels among items and the patterns of responses by students at the various item difficulty levels. For instance, we expect the high scoring students to usually perform well on the easiest items. Similarly, we expect the low scoring students to typically perform less well than high scoring students on the most difficult items.

Accuracy of each measure is reported according to how well the measure fits the overall pattern of expected scores. It is this pattern of expected scores that characterizes the Rasch model. Accuracy is reported as *infit*, one type of weighted fit index that is sensitive to systematic misfitting student responses (see column 6 of Table 1). In general, scores that exceed standardized infit of 2.0 may be problematic in that they are beyond the accepted range of fit to be considered unidimensional with other items of the instrument. A misfitting item does not appear to represent well the construct being measured, judging from the pattern of responses to that item. Another type of useful and important fit statistic is *outfit*, which is an unweighted fit index that provides another helpful perspective on fit, particularly at the extreme values. Outfit analysis has similarities to the infit analysis so will not be illustrated in this article.

An information graphic known as a person pathway plot (Bond & Fox, 2007), shown in Figure 1, illustrates the fit (circle location) and error (relative circle size) for each student measure. The plot shows that most examinees had productive fit with the Rasch model. Four examinees were shown to underfit the Rasch model predictions (Examinees 3, 7, 26, and 48) with infit standardized

statistics above 2.0. The pattern of responses suggests that some of the performances of these examinees conflicted dramatically with scores that would be expected on the basis of item difficulties and examinee abilities. For example, for Examinee 7, the most extreme underfit to the model, the measure was near average at -.02 logits, but the examinee responded correctly to item 9 and 8, which were two of the most difficult items, while responding incorrectly to item 10, the easiest item. Examinees 3, 26, and 48 showed similarly unexpected response patterns, but to a lesser degree than that of Examinee 7. In sum, only 3.5% of examinees showed unexpected patterns of performance on the 18 items. Three examinees were shown to overfit the Rasch model (46, 58, 86) with infit statistics below -2.0. Overfit of person measures indicates very highly predictable patterns of responses. This means that the examinee performances met expectations by matching the pattern of responding better than expected using the relative difficulties, which is not usually as problematic as failing to meet expectations. Excessive overfitting can, however, potentially mislead by inflating reliability values, so it should not be ignored. Excessive overfitting can also indicate redundancy among particular items which on some assessments may be undesired.

Item Difficulty Measures

When placing focus on an instrument's individual items, as opposed to the student performances, the item difficulty is an important component of meaningful measurement, as indicated by the discussion above. The difficulty of each item is based upon the sample of student performances being examined. This fact is crucial as different groups or samples of students are considered. Hence the larger and more diverse the sample of students, the more accurate and invariant the measurement of difficulty will tend to be.

Measures of difficulty for the 18 items ranged from -2.23 to 2.47 logits as shown in Table 2. No item values were extreme outliers. Difficulty of items ranged from 21% correct for item #10 to 88% correct for item #9. Thus no item was shown as too difficult or too easy for this sample overall. Standard error averaged .25 across the 18 items, ranging from .22 to .33. The item reliability was .95, supporting a wide range of item difficulties and a sufficient sample for this analysis. Difficulty data show that that almost all (8 out of 9) of the most difficult items were among the initial items presented (#2 through #9), which does not seem ideal to us from a test design perspective.

Quality Control For The Assessment: Item Pathway Plots

As with the analysis of student performances, an examination of each item's fit relative to all other items allowed us to understand whether the items were performing in a coherent, unidimensional way. That is, by fit we mean that we examined whether each item appeared to reflect the construct of interest, field dependence-independence (FDI), which the GEFT test is designed to assess.

Item fit statistics (Table 2) were reported as standardized scores which allow the level of 2.0 to serve as a quality control line. All of the 18 items were below or at the infit level of 2.0, indicating acceptable fit and correspondence to a unidimensional FDI construct from all items. An item pathway plot (Figure 2) illustrates that all 18 items showed productive fit with the Rasch model. However, items #4 and #6 were below the -2.0 standardized infit level, overfitting the Rasch model. This means that the items met expectations by matching the pattern of responding better than predicted using relative abilities, as noted above with person measures.

Item #5 was close to under-fitting the Rasch model just at the 2.0 level of the standardized infit statistic indicating some unexplained noise, but at an acceptable level. Future examination of this item is warranted by the near underfit and this possible concern.

Comparing The Assessment With The Students: Item-Person Map

By placing scores and difficulty levels on the same scale, another useful visualization tool known as an item-person map (left side of Figure 3) can be created. Along with the pathway plots described above, the item-person map allows efficient examination and interpretation of large amounts of data that even a modestly sized group can yield. Item-person maps can also be generated using solid bars rather than individual symbols representing each person or item.

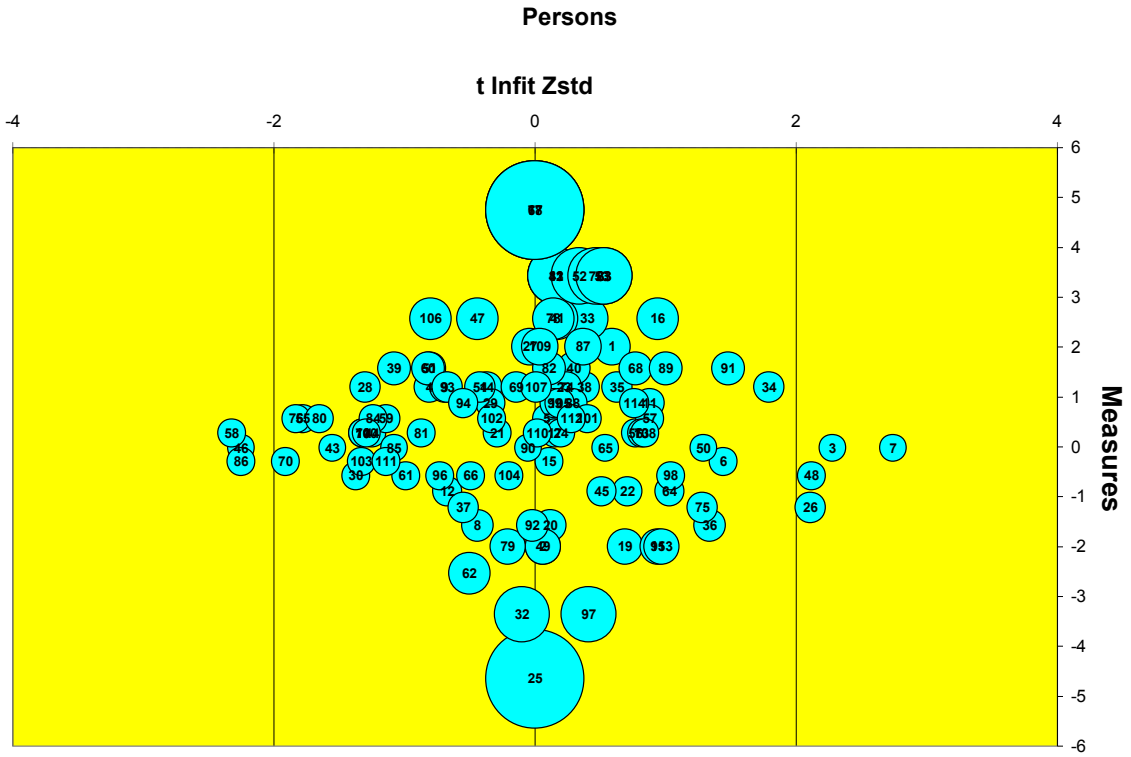
The item-person map shows a varied range in item difficulty with a small amount of duplication in items having similar difficulty levels (items #2 and #3; items #4 and #7) slightly above the middle range of all 18 items. Examinee ability levels are spread across the levels of difficulty with most examinees in the middle range and their abilities corresponding well with the distribution of item difficulties.

Still, 17 examinees (15%) were measured at ability levels above the difficulty level of item #9, the most difficult item. In other words, the probability was high that these 17 examinees would perform well on any of the 18 items, despite their imperfect scores. In addition, as noted earlier, three examinees earned perfect scores of 18 correct. These observations suggest the need for at least one item with greater difficulty than item #9 to help improve accuracy. On the other side of the measures, three examinees performed relatively lower than the difficulty level of item 10, the easiest item.

Table 2
GEFT Instrument Rasch Item Statistics in Ascending Measure Order for the 18 Items

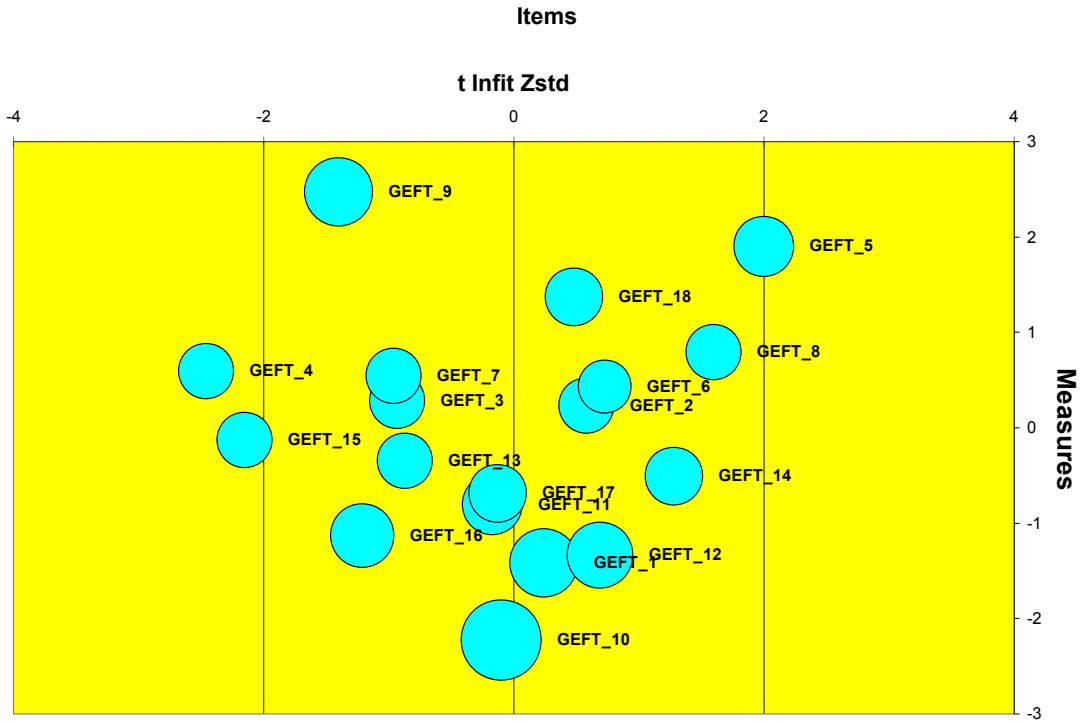
Entry Number	Total Score	Count	Measure	Model S.E.	Infit ZSTD
9	24.0	114	2.47	0.28	-1.4
5	32.0	114	1.90	0.25	2.0
18	41.0	114	1.37	0.24	0.5
8	52.0	114	0.79	0.23	1.6
4	56.0	114	0.59	0.23	-2.5
7	57.0	114	0.54	0.23	-1.0
6	59.0	114	0.43	0.22	0.7
3	62.0	114	0.28	0.23	-0.9
2	63.0	114	0.23	0.23	0.6
15	70.0	114	-0.13	0.23	-2.1
13	74.0	114	-0.35	0.23	-0.9
14	77.0	114	-0.51	0.24	1.3
17	80.0	114	-0.69	0.24	-0.1
11	82.0	114	-0.81	0.25	-0.2
16	87.0	114	-1.13	0.26	-1.2
12	90.0	114	-1.34	0.27	0.7
1	91.0	114	-1.42	0.28	0.2
10	100.0	114	-2.23	0.33	-0.1
Mean	63.5	110	0.0	0.25	-0.2
S.D.	20.3	0	1.17	0.03	1.2

Figure 1
Person Pathway for 114 Participants on the Group Embedded Figures Test (GEFT)



Note. Rasch measures (logits), standardized fit, and standard error (circle size) are plotted. A standardized quality control line of +2 is used to highlight those persons who underfit the Rasch model.

Figure 2
Item Pathway for 18 Items of the Group Embedded Figures Test (GEFT)



Note. Rasch measures (logits), standardized fit, and standard error (circle size) are plotted. A standardized quality control line of +2 is used to highlight those items that underfit the Rasch model.

This suggests some potential benefit to also including an item that is easier than item 10, depending upon the purpose of the FDI assessment.

As shown in Figure 3, the measurement scale accounts for the meaningful distinctions among performances toward the ends of the distribution while the raw score distribution fails to reveal these crucial performance distinctions. Specifically, on the right hand side of Figure 3 is a distribution of the raw scores that are not Rasch scaled. The red lines connecting this raw score distribution back to parts of the measurement scale distribution of those same scores (the item-person map on the left hand side) highlight the differences between the two distributions. On the left hand item-person map, distances between scores increasingly expand beyond one standard deviation above and below the mean, while on the raw score distribution distances between subsequent scores appear as the same amount.

A Test Item Diagnosis Tool: Item Characteristic Curves

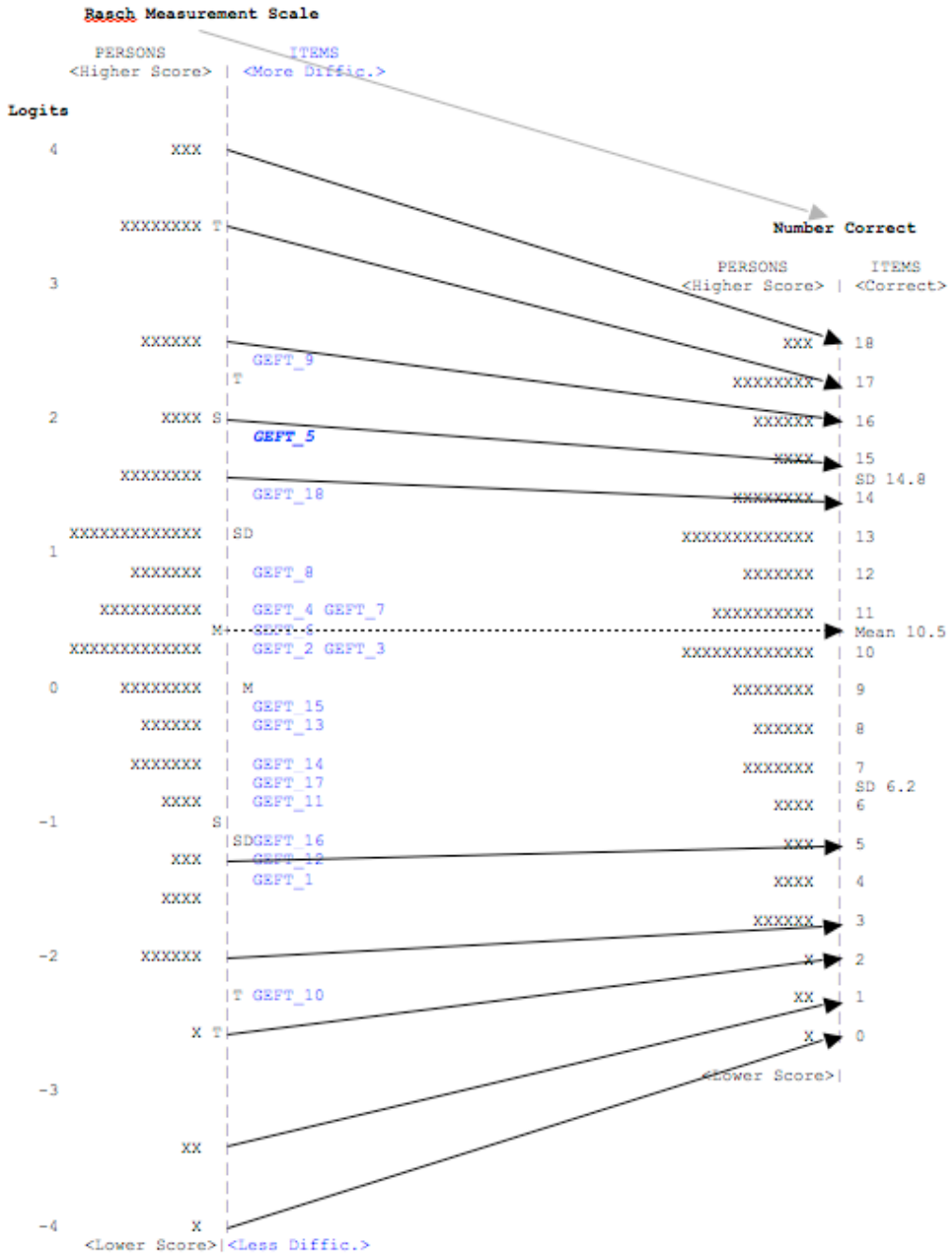
For visualization of specific item performances, line plots of actual scores on each item can be created alongside the Rasch model's expectation of item-person performance on GEFT (see Figure 4). The 95% confidence interval lines assist with the visualization of departures from the Rasch modeled expectations and their relative locations to lower, middle, or higher ability levels, from left to right on the item characteristic curve, respectively.

The example item characteristic curve of item #5 shows deviations from the 95% confidence interval lines. Item #5 was indicated previously for closer scrutiny due to near underfit of the model. The deviation from the empirical curve above the upper confidence interval line for lower ability levels illustrates the possible fit problem with item #5. This type of plot can be compared with the fit statistics overall, as well as with additional analyses and comparisons when alternative sample scores become available.

Measurement Quality: Item and Person Invariance

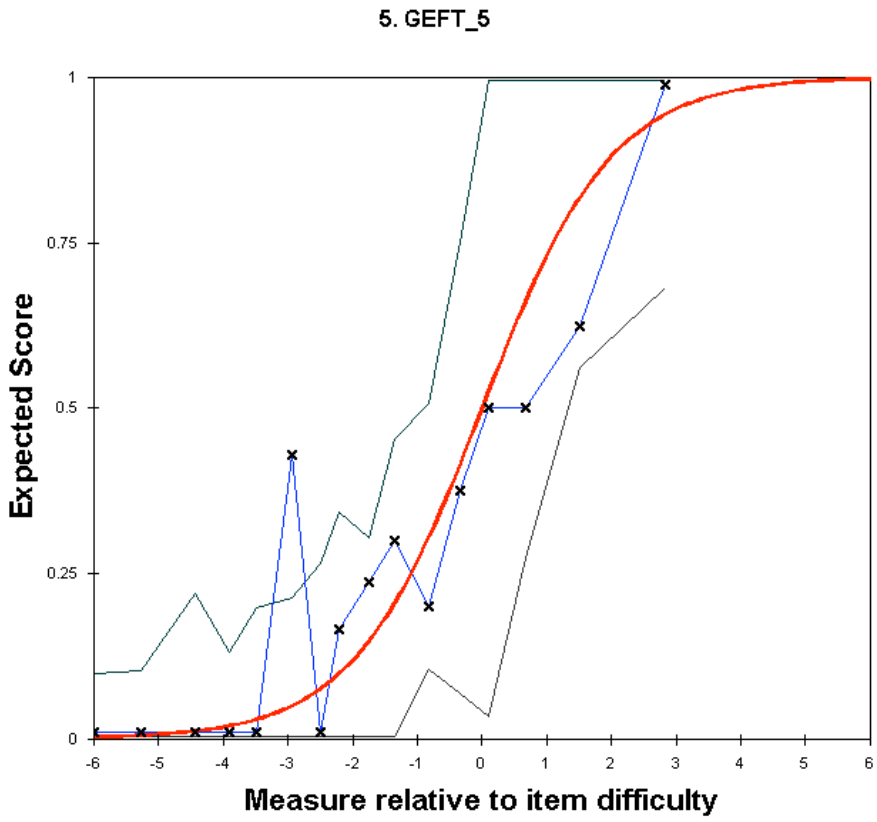
The idea behind the concept of invariance of estimates is that measures of items and students should not vary or differ excessively when either the items or people are divided up into groups of interest. To examine this crucial measurement requirement, two procedures were used that follow from the work of Wright and Stone

Figure 3
Item-person Map (left) and Number Correct Distribution (right) for 114 Examinees
on the Group Embedded Figures Test (GEFT)



Note. Arrow lines highlight the scale differences between the Rasch measurement scale constructed from scores versus those same scores on a traditional 'number correct' distribution.

Figure 4
Expected (smooth) and Empirical (jagged) Score Item Characteristic Curves with 95% Confidence Interval Lines for Item 5



(1979). In keeping with the need to focus on both persons and items, the first procedure allowed for the analysis of item difficulty invariance while the second procedure allowed for analysis of person ability invariance. The following discussion and information graphic should clarify this concept further.

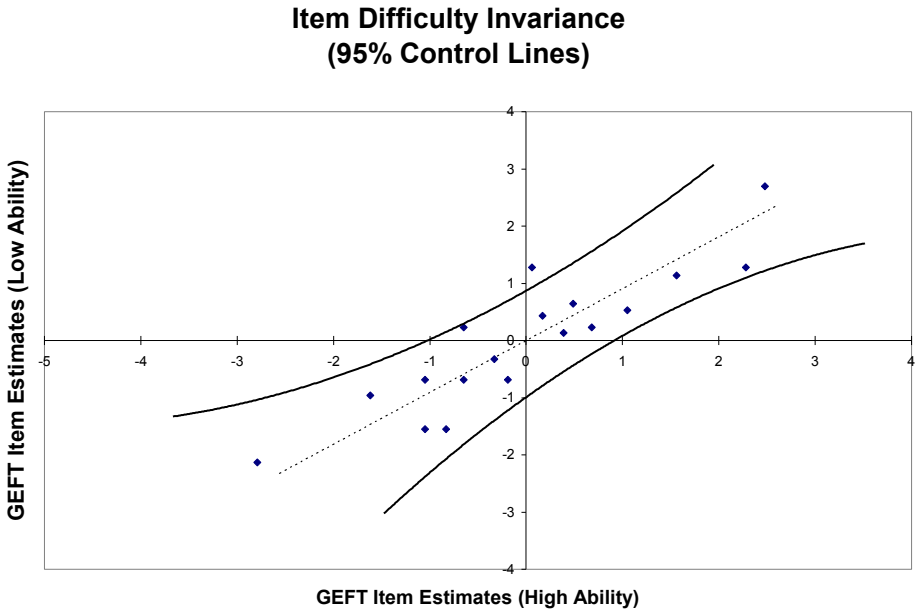
First, the examinee sample was divided into two groups according to ability. Item estimates for the high versus the low ability groups were then plotted along with 95% control lines (Figure 5). These control lines were based upon the standard errors and used to determine whether the plotted points were sufficiently invariant. Figure 5 shows that only one point lies outside the control lines, supporting the invariance, on the whole, for items on GEFT. That is, item points are near (allowing for error) the Rasch modeled dotted center line, representing invariance. Precision was reduced from the original analysis as reflected in the mean error rates for items equaling .31 for low ability and .41 for high ability versus .25 for all original students together.

Second, the GEFT items were divided into two groups according to item difficulty. Person/case estimates for difficult versus easy items were plotted with corresponding 95% control lines. Figure 6 shows that 46 of 52 points plotted were within the control lines. Of the 6 points plotted that were outside control lines, 5 were above the upper line and 1 was below the lower line. Although the precision of this comparison is relatively lower when 9 difficult versus 9 easy items were examined, person invariance is generally supported by the preponderance of items (88%) near the middle dotted line allowing for error. Again, precision was reduced from the original analysis as reflected in the mean error rates for persons equaling 1.12 for easy items and .95 for difficult items versus .68 for all original items together.

The reduction in the number of analyzed cases and items between these two invariance analyses also yielded lower person reliability estimates, as expected. Cronbach Alpha for the high ability analysis was .53, while that of the low ability analysis was .68. For the

Figure 5

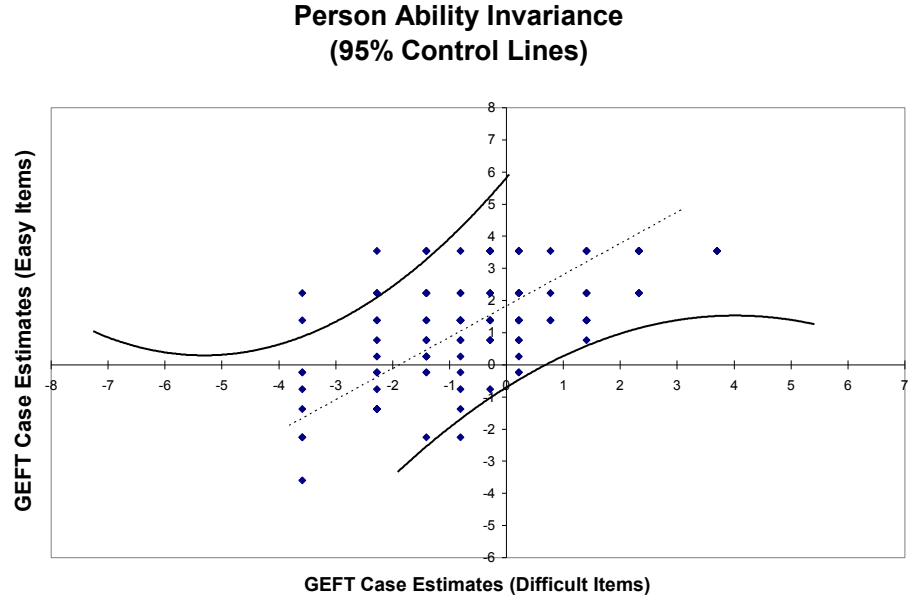
Low Ability Examinee Versus High Ability Examinee Item Measures on the Group Embedded Figures Test (GEFT)



Note. Data plotted near the central dotted line and within the 95% control lines reflect item difficulty invariance allowing for error.

Figure 6

Less Difficult (easy) Item Versus Difficult Item Person/Case Measures on the Group Embedded Figures Test (GEFT)



Note. Data plotted near the central dotted line and within the 95% control lines reflect person ability invariance allowing for error.

difficult item analysis, Cronbach alpha was .76, while that of the easy item analysis was .78. These lower reliability levels highlight a limitation of using smaller participant sample sizes and a modest number of assessment items (e.g., 18) when dividing the sample or the items in half for analysis. Dividing the participant sample was particularly problematic for reliability. However, the outcome of these two procedures demonstrated that invariance analyses can yield useful preliminary findings even when participant groups of interest are closer to 50 than 100 in number, for example. Ultimately, these types of error and reliability estimates can help researchers decide whether their samples of participants and items are sufficient for meaningful interpretation of analyses for the context.

Summary of Findings

By examining a few of the general conclusions that follow from this analysis, application of these Rasch procedures to other data sets can be further considered. Overall, in our example analysis college sophomores who were seeking entrance to teacher certification programs were shown to represent a broad range of visuospatial abilities. Rasch analyses allowed us to identify with greater accuracy and confidence the relative differences among our students in field dependence-independence. This type of identification can, for example, lead to improved understanding among instructional faculty of potential challenges for particular students on certain pedagogical approaches such as those involving complex problem solving and complex spatial information (Angeli, & Valanides, 2004; Chen, & Macredie, 2004; Richards et al., 2002) . The accurately measured differences among our students can be used to help anticipate the amount of support or the amount of challenge that might be necessary to facilitate learning growth among all our students.

The demonstrated diversity of this sample's visuospatial abilities also allowed for a useful examination of the 18 items that make up the GEFT instrument. We found that the addition of one or more items of greater difficulty is suggested by the 15% of examinees whose ability measure exceeded the measure of the most difficult item. Findings supported item and person invariance, and thus the potential for productive use of GEFT with this type of adult sample for working memory investigations and studies of learning, training, and instruction. All 18 items fit the Rasch model, though item #5 was close to underfitting the model, likely due to deviation from expected scores at the lower ability range. This general finding of fit indicates that the items are useful together as parts of this measure of field dependence-independence. However, further examination of item sequencing was suggested by the imbalance in difficulty levels among items across the instrument.

For future comparisons using this instrument toward a continuing process of validation (Messick, 1995), individual statistics for both items and examinees were provided on logit measures, standard errors, point biserial correlations, and fit. Overall reliability indices were also generated. Information graphics that included pathway plots, an item map, item characteristic curves, and invariance plots allowed visualization of patterns within the statistics. These Rasch statistics and graphic can be useful for further examination of individual student performances and the efficacy of items. This analysis can also be used for comparisons with future Rasch analyzed performance data using different embedded figures tests such as HFT (Ekstrom, French, Harman, & Dermen, 1976) and similar instruments.

Furthermore, with the continued progress in understandings of perceptual disembedding, working memory functions (Miyake et al., 2001), and associated brain region analysis (Walter & Dassonville, 2007), the benefits of using, redesigning, and refining instruments such as GEFT were supported by this analysis.

Implications and Discussion

This illustrative study is one example of how powerful Rasch analytic tools can be meaningfully used with a relatively modest participant sample of interest. The ability to generate useful measures and other related statistics from samples of students is essential for many studies of teaching and learning where groups of interest are not extremely large. Although the 114 students used in this study may represent a larger sample than many single class sizes, it is also much smaller than the hundreds or thousands (Jones, Smith, & Talley, 2006) often needed for other types of contemporary latent trait analyses. The common measurement scale used in the Rasch approach provides interpretation advantages for instructors or researchers, particularly when compared with the many classical test approaches that lack error estimates and a common additive scale for both persons and items. The invariance analysis illustrated the deleterious effect on error and reliability when a group of 54 was used instead of the original 114 participants. Where possible and appropriate, combining student data from several classes who take a common assessment can be used to improve reliability and accuracy of measures.

As noted above, the Winsteps computer program is an inexpensive tool that works well with Microsoft Excel, and it also imports data from common statistical programs such as SPSS, SAS, R, and STATA. Winsteps also has a demo version called

Ministeps as well as a training version called Bond&FoxSteps (Linacre, 2006a), which complements a widely recommended Rasch measurement text (Bond & Fox, 2007) and features simple functions for creating invariance of estimates graphics such as those in Figures 5 and 6. Other Rasch computer programs worth investigating include Conquest, Facets, RASCAL, and RUMM. These programs can be very useful for exploring the possible analytic approaches described in the current assessment and measurement literature.

Rasch analytic approaches and visualization tools can be a beneficial means to improving instruments and ultimately the validity of measurements (Wolfe & Smith, 2007) that help lead to more precise understandings of issues involving student learning and the associated teaching applications. Classroom assessments of learning outcomes, rating scales, and surveys can also be analyzed using Rasch approaches. Note, however, that the different purposes among ability tests, surveys, and classroom assessments call for different models, assumptions, and uses of the measurement scales. For example, when analyzing surveys on student differences, instead of ability and difficulty one might focus on a person's agreeability and the test item's endorsability (e.g., agreement or disagreement with an attitude statement). In such an example, distributions of Likert responses or rating scales could be usefully examined with Rasch rating scale model that is sensitive to the inherent differences with these types of assessments (see Bond & Fox, 2007).

Other important applications can include pretest and posttest differences on classroom assessments of learning that allow for sensitivity to scale distinctions in change scores among low pretest performers versus those of high pretest performers (Dimitrov & Rumrill, 2003; Wright, 2003). By estimating measures of student performances as opposed to merely quantifying performances with ordered data that lacks legitimate additivity, both large-scale and small-scale studies of students can yield more comparable and thus meaningful information toward improved decision making and inquiry about teaching and learning.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp.143-166). Maple Grove, MN: JAM Press.

Angeli, C., & Valanides, N. (2004). Examining effects of text-only and text-and-visual instructional materials on the achievement of field-dependent and field-dependent learners during problem-solving with modeling software. *Educational Technology Research and Development, 52*(4), 23-36.

Baddeley, A. D. (1999). *Essentials of human memory*. Hove, England: Psychology Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chen, S., & Macredie, R. D. (2004). Cognitive modeling of student learning in web-based instructional programs. *International Journal of Human-Computer Interaction, 17*(3), 375-402.

Dimitrov, D., & Rumrill, P. (2003). Pretest-posttest designs and the measurement of change. *Work, 20*, 159-165.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Hidden figures test: CF-1, revised: Kit of referenced tests for cognitive factors*. Princeton: Educational Testing Services.

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 487-525). Mahwah, NJ: Lawrence Erlbaum.

Linacre, J. M. (2006a). *Bond&FoxSteps Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2006b). *WINSTEPS Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 74-149.

Miyake, A., Witzki, A. H., & Emerson, M. J. (2001). Field dependence-independence from a working memory perspective: A dual-task investigation of the hidden figures test. *Memory, 9*(4), 445-457.

MacLeod, C. M, Jackson, R. A., & Palmer, J. (1986). On the relation between spatial ability and field independence. *Intelligence, 10*(2), 141-151.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published 1960, Copenhagen: Danish Institute for Educational Research).

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101.

Richards, J. T., Oman, C. M., Shebilske, W. L., Beall, A. C., Liu, A., & Natapoff, A. (2002). Training, transfer, and retention of three-dimensional spatial memory in virtual environments. *Journal of Vestibular Research, 12*, 223-238.

Rittschof, K. A. (2010). Field dependence-independence as visuospatial and executive

functioning in working memory: Implications for instructional systems design and research. *Educational Technology Research and Development*, *58*(1), 99-114. doi: 10.1007/s11423-008-9093-6

Smith, E. V., & Smith, R. M. (2004). *Introduction to Rasch measurement: Theory, models, and applications*. Maple Grove, MN: JAM Press.

Thompson, B., & Melancon, J. G. (1987). Measurement characteristics of the Group Embedded Figures Test. *Educational and Psychological Measurement, 47*(3), 765-772.

Walter, E., & Dassonville, P. (2007). In search of the hidden: Contextual processing in parietal cortex. *Journal of Vision, 7*(9). doi: 10.1167/7.9.1061

Witkin, H. A., Oltman, P., Raskin, E., & Karp, S. (1971). *A manual for the embedded figures test*. Palo Alto, CA: Consulting Psychologists Press.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I – instrument development tools. *Journal of Applied Measurement, 8*(1), 97-123.

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2. *Rasch Measurement Transaction, 17*(1), 905-906.

Wright, B. W., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E.V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 1-24). Maple Grove, MN: JAM Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Zhang, L. (2004). Field dependence/independence: Cognitive style or perceptual ability – validating thinking styles and academic achievement. *Personality and Individual Differences, 37*, 1295-1311.

KENT A. RITTSCHOF, Ph.D., is a Professor of Educational Psychology in the Department of Curriculum, Foundations & Reading at Georgia Southern University. He earned his advanced degrees from Arizona State University, specializing in Learning and Instructional Technology. His research has primarily involved spatial cognition, learning and assessment technologies, and psychological measurement. His teaching centers on applying learning and motivational theories to practice in all areas of education. He is currently working with funded projects involving the advancement of math and science learning.

WENDY L. CHAMBERS, Ph.D., is an Associate Professor of Developmental Psychology in the Department of Curriculum, Foundations & Reading at Georgia Southern University, where she has worked since receiving her doctorate in Developmental Psychology from the University of Florida in 1993. Her research interests include various aspects of cognitive development in early childhood, such as development of pretense and mental state understanding, as well as memory development. In addition, she has been involved with several research projects examining and implementing effective pedagogical practices for college students.