Chapman University
Chapman University Digital Commons

Physical Therapy Faculty Articles and Research | Physical Therapy

2005

Comparing Rasch Analyses Probability Estimates to Sensitivity, Specificity and Likelihood Ratios when Examining the Utility of Medical Diagnostic Tests

Daniel Cipriani
Chapman University, cipriani@chapman.edu

Christine Fox
University of Toledo

Sadik Khuder
Medical College of Ohio

Nancy Boudreau
Bowling Green State University

Follow this and additional works at: http://digitalcommons.chapman.edu/pt_articles

Recommended Citation
Cipriani, D., Fox, C., Khuder, S., & Boudreau, N. (2005). Comparing Rasch analyses probability estimates to sensitivity, specificity and likelihood ratios when examining the utility of medical diagnostic tests. *J Appl Meas, 6*(2), 180-201.

This Article is brought to you for free and open access by the Physical Therapy at Chapman University Digital Commons. It has been accepted for inclusion in Physical Therapy Faculty Articles and Research by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

Comparing Rasch Analyses Probability Estimates to Sensitivity, Specificity and Likelihood Ratios when Examining the Utility of Medical Diagnostic Tests

Comments
This article was originally published in *Journal of Applied Measurement*, volume 6, issue 2, in 2005.

Copyright
JAM Press

This article is available at Chapman University Digital Commons: http://digitalcommons.chapman.edu/pt_articles/39

JOURNAL OF APPLIED MEASUREMENT, *6*(2),
Copyright© 2005

Comparing Rasch Analyses Probability Estimates to Sensitivity, Specificity and Likelihood Ratios when Examining the Utility of Medical Diagnostic Tests

Daniel Cipriani
The Medical College of Ohio

Christine Fox
The University of Toledo

Sadik Khuder
The Medical College of Ohio

Nancy Boudreau
Bowling Green State University

Introduction: Medical diagnostic tests are evaluated based on measures of sensitivity (Sn), specificity (Sp), and likelihood ratios (LR). These procedures are limited in the event of a biased gold standard or missing data. Interpretations of these measures are frequently inappropriate. Purpose: The Rasch Measurement Model (RMM) was examined as a method to provide evidence of diagnostic test utility in order to overcome the limitations of Sn, Sp, and LR. Methods: Patients suspected of a knee ligament tear (n = 825) were studied, by evaluating four diagnostic tests. The RMM probability estimates for each test were compared to estimates of Sn, Sp, and LR. Results: The RMM provided probability estimates for the diagnosis that were comparable to likelihood ratios. These probability estimates correlated with the estimates of Sn, Sp, and LR. The RMM estimates were not affected by missing data. Discussion: The RMM may provide an alternative means to study the utility of medical diagnostic tests to estimate the probability of disease presence/absence.

Requests for reprints should be sent to Daniel Cipriani, Department of Physical Therapy, The Medical College of Ohio, 3015 Arlington Ave, Toledo OH 43614-5803, e-mail: dcipriani@mco.edu.

Diagnostic Test Utility: Reliability and Validity Evidence

Medical professionals rely on useful diagnostic testing procedures in order to base decisions regarding interventions and treatment planning for individuals with health related conditions. The utility of a diagnostic test is governed by the reliability and validity of the data for making accurate inferences, that is, the psychometric soundness of the test must be demonstrated before it is applied clinically. A useful diagnostic test provides the clinician with the confidence needed to make life-altering decisions on the part of the patient.

Examining the utility of a diagnostic test can involve a variety of different statistical and measurement approaches (Bohannon, 1997). For instance, in order to establish reliability of a score or outcome from a diagnostic test, researchers typically apply classical test theory models (e.g., test-retest reliability, internal consistency measures, etc.), as well as generalizability theory models (e.g. intraclass correlation coefficient) to measure intra and inter rater reliability (Crocker and Algina, 1986; Marcoulides, 1999; Shavelson and Webb, 1991). Researchers examine reliability of diagnostic tests in terms of the scores obtained from a single rater and/or multiple raters (i.e., inter and intra rater) when using a particular test or instrument (Portney and Watkins, 2000).

Common validity evidence for medical diagnostic tests include test sensitivity, specificity, and likelihood ratios (Indrayan and Sarmukaddam, 2001; Lilienfeld and Stolley, 1994; Rothman and Greenland, 1998; Woodward, 1999). These validity indices are used for diagnostic test outcomes, and are based on a comparison with a criterion test, often referred to as a gold standard. This gold standard provides a reference point for establishing validity, and relies upon a presumably definitive known outcome (Fritz and Wainner, 2001; Guggenmoos-Holzmann and van Houwelingen, 2000; Irwig, Glasziou, Chock, et al., 1994; Reid, Lachs, and Feinstein, 1995). Measures of sensitivity, specificity, and likelihood ratios are estimated using this gold standard; further, all contribute to the clinician's appreciation for the utility of a diagnostic test (Boyko, 1994; Hawkins, Garrett, and Stephenson, 2001; Phelps and Huston, 1995; Sackett, 1992; Sox, 1996).

The ability of a diagnostic test to correctly classify a person as diseased or not diseased, healthy or sick, etc. is paramount to a useful diagnostic test. Diagnostic tests are designed to discriminate between persons of different levels of health or disease. In the special case in which the diagnostic test produces the binary outcome of positive/negative, diseased/healthy, etc., these properties are referred to as sensitivity and specificity of the test. Sensitivity is the probability of a positive test result when the disease is present, a true positive; specificity is the probability of a negative test result when the disease is not present, a true negative (Begg and Greenes, 1983). Tests that accurately identify a person with a disease/dysfunction possess sensitivity. A test that is 100% sensitive will correctly identify all persons with the disease. Tests that accurately identify a person as healthy (i.e., disease/dysfunction free) possess specificity. A test that is 100% specific will correctly identify all persons without the disease.

Interpretation of sensitivity (Sn) and specificity (Sp) can be problematic. These measures are an indication of how well the diagnostic test works, and are not an indication of whether a disease is in fact present or absent. A highly sensitive test, for example 95%, is a test that will likely yield a positive outcome in the presence of a disease or condition. This high Sn value would at first appear desirable. However, if this same test has low specificity, for instance 35%, then this test results in a large number of false positives. Diagnostic tests for shoulder impingement syndrome, a common painful condition of the shoulder, are problematic for these reasons. For example, the Hawkins test is highly sensitive (> 90%) but has low specificity (< 25%). Yet, clinicians rely on this test to determine whether a diagnosis of impingement is present. This test results in a high likelihood of a false

positive, leading to over-diagnosis of impingement syndrome (Cahs, Akgun, Birtane, et al, 2000).

In order to assist clinicians with the interpretation of Sn and Sp, likelihood ratios were developed to explain the diagnostic utility of a test. Likelihood ratios consider both the Sn and Sp of a test in order to provide the clinician with a decision making process. Given known values of Sn and Sp, the clinician can calculate the likelihood that a disease is present or absent. Calculation of likelihood ratios, expressed as either a positive likelihood ratio (LR+) or a negative likelihood ratio (LR-) are based on Sn and Sp as follows:

Positive Likelihood Ratio = sensitivity / 1– specificity

Negative Likelihood Ratio = 1 - sensitivity / specificity

Based on this relationship, a positive likelihood ratio of 1.0 indicates that the test result does nothing to change the odds of a person actually having the disease of interest. Likelihood ratios greater than 1.0 increase the odds favoring the condition. Negative likelihood ratios of 1.0 also do not provide any useful information. However, negative likelihood ratios less than one decrease the odds favoring the condition. Small negative likelihood ratios correspond to high sensitivity values, yielding a measure that is useful for ruling out a condition. Likewise, large positive likelihood ratios correspond with high specificity, indicating a measure that is useful for ruling in a condition (Fritz and Wainner, 2001).

In order to apply likelihood ratios, clinicians must first estimate the pre-test probability of a particular disease or condition. Once a pre-test probability has been established, this probability is converted to an odds. Clinicians can then apply the likelihood ratio to the pre-test probability in order to obtain an estimation of a post-test probability for the disease. Many clinicians apply the nomogram proposed by Fagan (1975) as an application of Bayes' theorem for interpretation of likelihood ratios. A more accurate approach is the method proposed by Sacket, Haynes, Guyatt, and Tugwell (1992). This process involves three simple calculations. The first calculation is to convert the pre-test probability to an odds using the formula

Pre-test odds = pre-test probability / (1 – pre-test probability).

The pre-test odds are then multiplied by the likelihood ratio, to estimate the post-test odds:

Post-test odds = (pre-test odds) x (likelihood ratio).

Finally, the post-test odds are converted to a post-test probability using

Post-test probability = post-test odds / (post-test odds + 1)

For example a clinician might assign a pre-test probability of 75% for a patient presenting with symptoms (i.e., a medical and symptom history) consistent with other patients who have a given diagnosis. The clinician then administers diagnostic tests and applies the likelihood ratio values to this pre-test probability. The pretest probability of 75% has a pre-test odds of 3.0. Assuming a diagnostic test has a LR+ of 5.46, the post-test odds becomes 16.38 and a post-test probability that the diagnosis is present of 94.25%. Thus, the positive diagnostic test improved the clinicians confidence from 75% to 94%.

Weakness of Current Methods to Examine Test Utility

Sensitivity and specificity pose interpretation problems for the clinician, as described earlier. In addition, Sn and Sp are highly dependent on a perfect gold standard—a standard that is equitably applied to all individuals in the population of interest. Unfortunately, the gold standard is likely to be biased because of missing observations, selection bias or because only a subset of the population of interest are exposed to the gold standard (Bates, Margolis, and Evans, 1993; Green, Black, and Johnson, 1998; Hlatky, Pryor, Harrell, et al., 1984; Phelps and Hutson, 1995; Valenstein, 1990). A biased gold standard results in biased estimates of the performance of the criterion test (Begg and Greenes, 1983; Bates, Margolis, and Evans, 1993). For example, sur-

gery is a common gold standard and generally only those patients in which the surgeon is confident needs surgery will actually be subjected to this gold standard. Those who the surgeon may not be as confident will likely not be subjected to the gold standard. Thus, these cases can not be included in the analysis of sensitivity and specificity, resulting in missing and biased data for the diagnostic tests (de Bock, Houwing-Duistermaat, Springer, et al., 1994; Joseph, Gyorkos, and Coupal, 1995). Finally, as noted by Guggenmoos-Holzmann and van Houwelingen (2000) values of sensitivity and specificity relate more to the performance of the test rather than to the likelihood that a disease is present or absent—that is, these values are test centered rather than patient centered.

The use of likelihood ratios of diagnostic tests require that clinicians and researchers provide an estimate of the pre-test probability of a disease. This estimate is susceptible to subjective bias on the part of the clinician or researcher (Fox, Landrum-McNiff, Zhong, et al., 1999; Reid, Lane, and Feinstein, 1998; Timmermans, 1994).

The ability to determine the utility of diagnostic tests is limited by the problems presented using standard sensitivity, specificity, and likelihood ratios (e.g., sample/item dependence, gold standard dependence, interpretation problems). In summary, the problems of these procedures include interpretation concerns, reliance on a gold standard, an inability to analyze missing observations, an inability to make direct comparisons between tests as to the possibility of redundancy, and the reliance on a subjective pre-test probability estimate of disease presence

The Rasch Regression Model

A potential solution to these problems is the Rasch Measurement Model (Rasch, 1980). Recently, the Rasch Measurement Model (RMM) has gained popularity as an alternative means to examine the reliability, validity, and utility of measures in medicine and health care (Campbell, Kolobe, Osten, Lenke, and Girolami, 1995; Chang and Chan, 1995; Creel, Light, and Thigpen, 2001; Fisher, 1993; Haley, McHorney, and Ware, 1994; Heinemann, Harvey, McGuire, et al., 1997; Lai, Fisher, Magalhaes, and Bundy; 1996; MacKnight and Rockwood, 2000; Morris, Morris, and Iansek, 2001; Silverstein, Fisher, Kilgore, Harley, and Harvey, 1992; Velozo, Kielhofner, and Lai, 1998). Originally, educational and psychology professionals used the RMM to examine tests; health care professionals now use the RMM to examine health and medical tests (Bond and Fox, 2001; Chang, Slaughter, Cartwright, and Chan, 1997; Fisher, 1993; Fisher, Harvey, Taylor, Kilgore, and Kelly, 1995; Fox, 1999; Fox and Jones, 1998; Harada, Chiu, Damron-Rodriguez, et al., 1995; Heinemann, Linacre, Wright, Hamilton, and Granger, 1993; Karabatsos, 1997; Prieto, Roset, and Badia, 2001; Rheault and Coulson, 1991; Tesio, Granger, and Fiedler, 1997; Velozo, Magalhaes, Pan, and Leiter, 1995).

Traditionally investigators used the RMM to examine one test at a time (i.e., a test that contains numerous items such as a math test or a survey questionnaire), focusing on the interaction between the persons/patients and each item of a test. Recently investigators have proposed applying the RMM to examine multiple tests simultaneously, as a means to examine the predictive ability of a combination of tests (Wright, Perkins, and Dorsey, 2000; Beltykova, Cipriani, Yan, Ughrin, and Fox, 2000). In fact, Wright, Perkins, and Dorsey (2000) refer to this approach as multiple regression through measurement. In this case, the various diagnostic tests are treated as individual test items and the patients are the persons who are measured by these items. A person with "more" of a disease would be more likely to respond positively to a diagnostic test just as a person with "more" of an ability/construct would be more likely to respond positively to an easier item.

Wright, Perkins, and Dorsey (2000) entered multiple predictor variables into the RMM as a means to predict the outcome of the disease gout. By aligning the predictor variables on a single calibrated ruler, they were able to identify at which points each predictor variable would likely result in a positive outcome of gout. The investigators created transformed scores representing

the predictor variables of urea nitrogen, uric acid, and creatinine, all three being blood values. Using the RMM, they were able to identify points of each blood value that would give the most likely outcome of gout in this sample of patients.

Beltyukova, Cipriani, Yan, et al. (2001) performed a similar analysis to predict driver capability in older adults. Based on two predictor variables (a visual test and a clock performance test), they were able to examine how well these two tests could predict the dichotomous outcome of capable or incapable of driving. While the predictor variables were able to create some separation in the sample, they were not able to predict with any degree of confidence the desired outcome. They determined that the two tests would not make good predictors of driving ability, as nearly 65% of the participants could not be diagnosed with greater than 75% accuracy.

A potential advantage of using the RMM over traditional approaches to examine diagnostic test utility is that the RMM provides information not readily available by the other procedures previously mentioned. These features include individual estimates for person and test/item fit (a measure of validity and unidimensionality), individual estimates of person and test/item reliability, an index of how well the tests separate persons into different categories (a measure of validity), and an indication of order validity (i.e., items/persons span a continuum from least to most of a latent trait) which is often referred to as hierarchical structure (Fox, 1999; Andrich, 1988; Karabatsos, 2001; McNamara, 1996; Klauer, 1995; Smith, 2001; Smith, 2000).

The RMM may be able to provide information similar to the sensitivity/specificity of dichotomous tests as well as the probability of disease/health in an individual (e.g., likelihood ratios). Because the RMM allows individuals and items to be fit along a common ruler, it should be possible to estimate the probability of a disease/condition based on any individual's position in regards to each item of the battery of tests. In addition, because all items are fitted along a common ruler, this feature will allow for direct comparisons between two or more tests. This comparison will allow the investigator to examine if certain tests are in fact redundant, or if each test provides additional information regarding the likelihood of a given diagnosis.

The purpose of this study was to compare the outcomes/interpretations of the RMM with standard procedures (i.e., sensitivity, specificity, likelihood ratios) on the criteria of reliability, and validity of diagnostic tests. The outcomes we examined were 1) the ability of the Rasch Measurement Model to provide useful interpretations of the outcomes of diagnostic tests, 2) the ability of the Rasch Measurement Model to identify redundant testing procedures in diagnostic testing—by making direct comparisons between each diagnostic test on their ability to separate persons into distinct health categories, and 3) the ability of the Rasch Measurement Model to overcome the difficulty of missing observations in the gold standard data.

Specifically, this study examined the ability of the Rasch Measurement Model to overcome the limitations of current methods to examine the utility of diagnostic tests such as sensitivity, specificity, and likelihood ratios.

Methods

Procedure

This investigation used a retrospective chart review to generate a case-control design. Patient files were obtained from an outpatient orthopaedic surgery clinic. The diagnosis for cases was a tear of the anterior cruciate ligament of the knee, a very common diagnosis in this particular setting. Cases included individuals with a confirmed tear of the anterior cruciate ligament of the knee. Controls included individuals with a suspected knee injury and the absence of a tear to the anterior cruciate ligament. The "gold standard" test was the outcome following arthroscopic investigation and/or surgery of the knee (Kirkley, 1997; DeHaven, 1983). The diagnostic tests included the Lachman's test, Pivot Shift Test, history of an audible pop ("pop"), a Magnetic Resonance Imagery (MRI), and a mechanical ligamentous examination (Mech.Exam). These

tests are useful for the screening of a knee with a potential tear to the ACL (Solomon, Simel, Bates, Katz, and Schaffer, 2001; Neeb, Aufdemkampe, Wagener, and Mastenbroek, 1997).

Sample

The file review included 825 patient charts from individuals who underwent at least the Lachman's test as part of a normal knee examination as well as arthroscopic exploration and/or surgery of the knee. In this way, even patients without a tear to the ACL were included in the sample. Thus, a true diagnosis was available for these patients. A subset of patients' records were examined for analysis. This subset consisted of patients without a known diagnosis for their knee injury (n = 52). These patients underwent the standard physical examination, however, they did not undergo exploratory or reconstructive arthroscopic surgery. These cases were used to demonstrate the application of the RMM for missing data.

Age, gender, and time since injury (i.e., the time in weeks between the actual injury and the time of the examination) were recorded as demographic data. These data were used to compare the sample of this study with the literature.

Variables

The following diagnostic tests served as the predictor variables: Lachman's test, Pivot shift test, report of a "pop" sensation at time of injury, Magnetic Resonance Imaging (MRI) result**,** mechanical knee ligament examination (e.g., the knee signature system, the KT-1000/2000) result. Positive and negative outcomes of a test were based on the clinician's interpretation of each test, based on the definitions provided by the literature (Losee, 1983; Jakob, Staubli, and Deland, 1987; Neeb, Aufdemkampe, Wagener, and Mastenbroek, 1997; Malcom, Daniel, Stone, and Sachs, 1985; Forster, Warren-Smith, and Tew, 1989; Tomberlin and Saunders, 1999)**.** The mechanical knee exam (Mech.Exam) variable was treated as a continuous variable, measured in millimeters of difference between the healthy and involved knee. This measure was based on the amount of anterior tibial translation relative to a fixed femur while an anterior directed force was applied to the tibia, mechanically.

The gold standard for this investigation was arthroscopic examination/surgery. The post surgery diagnosis served to confirm the diagnosis of ACL tear or no ACL tear. Those individuals who did not undergo arthroscopic surgery were treated as either missing data (i.e., in terms of the gold standard) or treated on a continuum between healthy and torn (i.e., unsure).

Data Analysis

Likelihood ratios were based on the values of sensitivity and specificity, thus representing the ability of a diagnostic test to identify persons with a tear or a healthy knee. In addition, the mechanical knee examination was treated as a continuous variable, measured in millimeters of difference between the healthy and involved knee. This predictor variable was entered in a simple logistic regression model to obtain individual estimates for the 2x2 table.

The RMM analysis used the Partial Credit Model (Wright and Masters, 1982) with WINSTEPS software (Linacre and Wright, 1991). The person ability estimate represented the health status (i.e., ability) of a person. Persons of poor knee health (i.e. ACL tear) were expected to respond positively to a given diagnostic test (i.e., a positive test result). Persons of good knee health were expected to respond negatively to a given diagnostic test (i.e., a negative test result). The item difficulty estimate represented the difficulty (or ability to discriminate between healthy and injured) of a diagnostic test to elicit a positive or negative response from a person. Thus, a person's overall performance on the diagnostic tests was a function of that person's level of health (ability) and the discrimination ability of the test. For clarification, diagnostic tests were referred to as "items" to represent the items of the battery of diagnostic tests entered into the model. Just as a traditional questionnaire or intelligence test possesses multiple "items," the battery of diagnostic tests consisted of multiple items (i.e., Lachman's test, pivot shift test, MRI, etc).

For the continuous variable (KSS or KT1000), the partial credit RMM treats each millimeter increment as a discrete category. Thus, the mechanical knee examination was coded on a scale of 0 – 9 mm of instability of the knee, based on the millimeter difference in displacement of the tibia relative to the fixed femur between the healthy knee and the injured knee; a 0 indicated no displacement and each subsequent one mm represented additional anterior displacement of the tibia of the injured knee compared to the uninjured knee. The greater the displacement, the more likely the ACL is torn/injured (Neeb, Aufdemkampe, Wagener, and Mastenbroek, 1997; Forster, Warren-Smith, and Tew, 1989).

Based on the requirements of the RMM, the data were modeled such that the probability of a given response (i.e., responding positively to a given item) was conditional on the diagnostic status of the patient (i.e., the level of diagnosis as either positive or negative) and the discrimination ability of the item (i.e., how well it functions to distinguish between a positive or negative patient). This is synonymous with the probability of successfully responding to a test question given a person's ability level and the difficulty level of the test item. Thus, persons with definite positive diagnosis for ACL tear should respond positive to a test that discriminates between those with healthy knees and those with injured knees. Similarly, persons with healthy knees should not respond positive to a test that requires a tear for positive response.

All diagnostic items were entered into WINSTEPS for analysis on the first step. This allowed for anchoring of the person and item values based on the responses of persons to each diagnostic item. The actual diagnosis variable (i.e., outcome of surgical exploration) was then be entered into the analysis as a separate dependent variable (Wright, Perkins, and Dorsey, 2002).

The average category measures for each diagnostic test item represented the sample average measure for each category of each item. Thus, it was possible to obtain the average measure of a positive test as well as the average measure of a negative test for each diagnostic test item. The measure for the outcome variable (i.e., diagnosis) was then used to estimate the increase/decrease in probability of a tear to the ACL, given a positive/negative result of any diagnostic test procedure. The difference in logit values between the diagnosis measure and a measure of a positive or negative response on a test, was used to estimate the probability of a tear or no tear. Thus, if a positive measure on a particular test was equal to the diagnosis measure (i.e., a difference of zero logit between the two measures), a person would have a 50% probability of a tear to the ACL given a positive response on that test, based on the logit transformation

$$\text{Probability} = \exp(\text{logit}) / 1 + \exp(\text{logit}),$$

where $\exp(0) = 1$ and $1 / (1 + 1) = 0.50$. Thus, subtracting the diagnosis logit value from the average positive logit value of a diagnostic test provided the probability of a tear of the ACL given a positive test result for that particular diagnostic procedure. Similarly, obtaining the difference in the diagnosis logit and the average negative logit value of a diagnostic test provided the probability of a tear of the ACL given a negative result of that test.

Finally, the RMM was used to generate reliability estimates of the measures for persons and items. In addition, the separation index was generated and examined in order to assess the ability of the diagnostic tests to discriminate between persons with healthy and unhealthy ACLs.

The RMM and Missing Data in the Response Variable

In order to test the RMM as a method to analyze diagnostic tests in the absence of a perfect gold standard (i.e., missing observations), data from patients who did not undergo knee surgery were included in this analysis. This analysis consisted of comparing the average measures for a positive and negative outcome on the diagnostic tests, between a set of data with all known diagnoses ($n = 825$) and a set of data with missing diagnoses ($n = 877$). In the data set of missing diagnoses, 52 patients did not undergo knee surgery and were therefore without a known diagnosis.

Results

Description of File Review

This investigator examined 825 patient files of individuals who had undergone arthroscopic surgery between the dates of January 2001 and January 2003. Of these files, 52.8% consisted of patients with no tear to the ACL (n = 436 controls) and 47.2% consisted of patients with a confirmed tear to the ACL (n = 389 cases). The gender distribution of these patients was 59.0% males (n = 487) and 41.0% females (n = 338). The proportion of males and females with a confirmed tear of the ACL was 48.3% and 45.6% respectively, and these proportions were not significantly different (c^2 [1] = 0.580, $p > .05$). Table 1 contains these frequencies. The data of these 825 patients were then used to compare the outcomes of the RMM with standard tests of sensitivity, specificity, and likelihood ratios.

A second file review was completed to produce cases of patients who had undergone knee examination for suspected injury to the ACL, but no surgical report was available. These cases were used to include missing cases. In all cases, patients had chosen not to undergo surgical exploration and/or repair. Fifty-two (52) cases were reviewed. The gender distribution was 61.5% males (n = 32) and 38.5% females (n = 20).

Sensitivity, Specificity, Likelihood Ratios for all Data

In order to compare the interpretations of standard approaches to examine the utility of diagnostic tests with the RMM, this investigation first compared test sensitivity, specificity, and likelihood ratios to the logit (and probability) estimates of the RMM, using the entire set of available data.

Table 2 provides sensitivity, specificity, and likelihood ratio values for each of the diagnostic tests. The mechanical knee examination was the most useful for making a positive or negative diagnosis (Sn = 99.3%, Sp = 99.6%, LR+ = 228.4, LR- = 139.9, respectively), whereas the history of a pop sensation at the time of injury proved to be the least useful of the tests (Sn = 72.5%, Sp = 76.1%, LR+ = 3.0, LR- = 2.8, respectively). Table 3 provides the Rasch Model logit estimates and probability estimates for each of the diagnostic tests. Once again, the mechanical knee examination proved to be the most useful (logit positive = 3.2, logit negative = -3.5, probability of positive = 96%, probability of negative = 97%) of the diagnostic tests; the history of a pop proved to be the least useful (logit positive = 1.7, logit negative = -1.4, probability of positive = 85%, probability of negative = 81%) of the diagnostic tests.

Table 1

Frequency of patients in each category of diagnosis and gender

	% of All Data Sets Combined(n = 825)
Diagnosed with tear	47.2%
Diagnosed no tear	52.8%
Male	59.0%
Female	41.0%
Male with tear	48.3%
Female with tear	45.6%

Table 2

Standard Diagnostic Test Values for the Diagnosis of an ACL Tear of the Knee

Test	Sn (*SE*)	Sp (*SE*)	LR+ (*SE*)	LR- (*SE*)
Hx of Pop[a]	72.49 (.02)	76.15 (.02)	3.04 (1.10)	2.77 (1.09)
Lachman	92.03 (.01)	91.97 (.01)	11.46 (1.18)	11.54 (1.19)
Pivot Shift	76.72 (.02)	99.63 (.01)	208.68 (2.71)	4.28 (1.11)
Mech.Exam[b]	99.29 (.01)	99.57 (.01)	228.36 (2.71)	139.89 (2.02)
MRI	92.55 (.01)	83.05 (.02)	5.46 (1.14)	11.15 (1.24)

Note: Sn (Sensitivity), Sp (Specificity), LR+ (Positive Likelihood Ratio), LR- (Negative Likelihood Ratio).
[a]History of "pop" sensation at time of injury
[b]Mechanical Knee Examination

The interpretations derived from the standard procedures and the RMM appear to be similar. Both procedures identify the mechanical knee exam as the test yielding the highest probability of a tear or no tear, given a positive or negative result, compared with the other diagnostic tests. Both procedures identify the history of "pop" as the least useful test, with lower likelihood values and probability values compared with the other tests. Further, the pivot shift test, which has a high specificity but low sensitivity, is interpreted as more appropriate to correctly identify a tear to the ACL compared with correctly identifying a healthy knee. In other words, given a positive pivot shift, the likelihood of a tear (LR+ = 208.7) is much greater than the likelihood of no tear (LR- = 4.3) given a negative pivot shift. Similarly, the probability of a tear (92.5%), based on the RMM estimate, is higher than the probability of no tear (79.9%) based on the RMM estimate.

Testing for a formal correlation between the standard procedures and the RMM estimates demonstrated significant correlations between standard approaches and RMM estimates ($p < .05$). A significant correlation was found between sensitivity values and RMM estimates of probabilities of a negative diagnosis ($r = .89$, $p < .05$) and a significant correlation was found between specificity values and RMM estimates of probabilities of a positive diagnosis ($r = .75$, $p < .05$). Table 4 provides this correlation matrix.

Further, positive likelihood ratios were significantly correlated with the probability values of a positive diagnosis from the RMM ($r = .78$, $p < .05$). Negative likelihood ratios were correlated with the probability values of a negative ($r = .88$, $p < .05$) and positive ($r = .94$, $p < .05$) diagnosis from the RMM. Table 5 contains this correlation matrix.

RMM Analysis in the Presence of Missing Data in the Response Variable

The estimated measures of a positive and negative outcome for each of the predictor variables (i.e., Lachman's test, pivot shift, mech.exam, and MRI) were calculated for two sets of related data. These data sets included a set ($n = 825$) of all patients with a known diagnosis and a set ($n = 877$) of the original 825 patients including an additional 52 patients with no known diagnosis (missing gold standard). The estimated category averages, in logit values, are presented in Table 6 for each diagnostic test, for each data set. There was no difference between the three sets of estimated measures for the data sets ($p > .05$). Thus, regardless of the data set,

Table 3

Estimated Logit Values and Probabilities Associated with Definite Positive and Definite Negative Decisions of Diagnostic Tests using the Rasch Model

Test	Logit Value for Positive (*SE*)	Probability for Positive (*SE*)	Logit Value for Negative (*SE*)	Probability for Negative (*SE*)
Hx of Pop	1.74 (.08)	85.07 (.02)	-1.44 (.08)	80.84 (.02)
Lachman	2.32 (.05)	91.05 (.02)	-1.77 (.07)	85.44 (.02)
Pivot Shift	2.51 (.07)	92.48 (.01)	-1.38 (.09)	79.89 (.03)
Mech. Exam	3.17 (.12)	95.97 (.02)	-3.46 (.17)	96.95 (.01)
MRI	2.21 (.08)	90.11 (.02)	-2.07 (.09)	88.79 (.02)

Table 4

Correlation Matrix of Sensitivity and Specificity with the Rasch Estimated Probabilities of a Positive or Negative Torn ACL

	Rasch Probability of Positive	Rasch Probability of Negative
Sensitivity	0.71	0.89[a]
Specificity	0.75[a]	0.33

[a]significant at p < .05

the RMM provided similar estimates for the category averages of a positive tear or a negative tear, for each of the diagnostic tests and the actual diagnosis.

Validity and Reliability Estimates of the RMM

The function of the diagnostic tests to correctly identify persons with a tear or no tear to the ACL was supported using the RMM separation index (G_p). The analysis of 825 observations and the five predictor variables resulted in a separation index of $G_p = 1.82$. This index resulted in at least two statistically distinct strata based on the strata estimation of $H_p = (4G_p + 1) / 3 = 2.76$. Thus, the four diagnostic tests sufficiently separated the sample into at least two distinct categories, a positive diagnosis of a tear and a negative diagnosis of no tear to the ACL.

The reliability of the patient measure estimates and the diagnostic test measure estimates were both high. The patient reliability $R_p = .77$ and the diagnostic test item reliability $R_i = .98$. The higher test reliability compared with the patient reliability is a reflection of the large sample size ($n = 825$) to estimate the tests compared with the small number of tests ($n = 4$) to estimate the persons.

Discussion

Utility of the Diagnostic Tests: Comparing the RMM with Standard Procedures

The RMM probability estimates function in a similar fashion to sensitivity and specificity. The RMM produces estimates of a person's health and estimates of a test's function along a common linear scale. In other words, a person's health

Table 5

Correlation Matrix of Likelihood Ratios and Odds Ratios with the Rasch Estimated Probabilities of a Positive or Negative Torn ACL

	Probability Yes[a]	Probability No[b]
LR+[c]	.78[f]	.33
LR-[d]	.94[e]	.88[e]

[a]Rasch Estimate for the probability of a positive diagnosis, given a positive test
[b]Rasch Estimate for the probability of a negative diagnosis, given a negative test
[c]Positive likelihood ratio
[d]Negative likelihood ratio
[e]significant correlation at $p < .05$

Table 6

Estimated measures comparing a RMM with a known diagnosis (n = 825), a missing missing gold standard (n = 877), and an ordinal outcome gold standard (n = 877).

	Estimated Logits (*SE*)	
	Known Diagnosis	Missing Gold Standard
+ Lachman	2.09 (.11)	2.03 (.08)
- Lachman	-3.78 (.11)	-3.65 (.10)
+ Pivotshift	2.43 (.11)	2.38 (.11)
- Pivotshift	-3.16 (.15)	-3.02 (.13)
+ MRI	2.09 (.12)	2.04 (.11)
- MRI	-4.26 (.15)	-4.09 (.14)
+ Mech.Exam	4.62 (.21)	4.63 (.20)
- Mech.Exam	-7.48 (.10)	-7.01 (.09)
+ Diagnosis	1.74 (.07)	1.71 (.07)
- Diagnosis	-3.87 (.10)	-3.73 (.09)

Note: "+" indicates a positive outcome and "-" indicates a negative outcome on each diagnosistic test and the actual diagnosis.

is estimated in logit values as is the function of the diagnostic test. In this way, a person with a high likelihood of having a particular diagnosis will be assigned a relatively high logit value. A diagnostic test score that is most difficult to attain (i.e., a positive response vs. a negative response) will also be assigned a relatively high logit value. Positive and negative responses to a diagnostic test are aligned at the two ends of the logit scale, with positive responses on the high end of the scale (e.g., positive logit values) and negative response on the low end of the scale (e.g., negative logit values). In this way, the function of the test can be analyzed in terms of its ability to correctly classify persons as either healthy or injured (i.e., diagnosed with the condition). The larger the absolute logit value for a test category, the greater the probability of the outcome (i.e., healthy or unhealthy).

To demonstrate the relationship between sensitivity, specificity, and RMM probability estimates, consider the pivot shift test. The sensitivity of the pivot shift test was relatively low (76.72%), making it a poor test to rule out a diagnosis. This test likely results in a large number of false negative classifications because it incorrectly identifies nearly 25% of the persons as negative, when in fact these persons are actually positive. The RMM probability estimate of a negative diagnosis based on the pivot shift test was 75.00%. The interpretation of this value is that there is a 75% probability that the ACL is not torn, given a negative pivot shift test. Thus, there remains a 25% chance that the ACL is in fact torn, even in the presence of a negative pivot shift test. This test does not function well to rule out a diagnosis. On the other hand, the specificity of the pivot shift test was 99.63%, which is very high. This test should function well to rule in a diagnosis. It does not appear to yield many false positives; it correctly classifies persons without the condition as negative. According to the RMM probability estimate, a positive pivot shift results in a 96.62% probability that the ACL is torn. Thus, this test is useful to rule in a diagnosis. Based on the RMM, there is less than a 5% chance that the person is incorrectly diagnosed with a tear to the ACL.

Positive/Negative Likelihood Ratios, and RMM Probability Estimates

Positive likelihood ratios significantly correlated with RMM probabilities of a positive tear to the ACL ($r = .78$, $p < .05$; $r = .97$, $p < .05$, respectively). Similarly, negative likelihood ratios significantly correlated with the RMM probabilities of a negative diagnosis ($r = .88$, $p < .05$). These relationships were expected given the similarity of interpretations of these values. Positive likelihood ratios reflect the expected probability that a person has a diagnosis, given a positive test result. Thus, a high positive likelihood ratio suggests that a person has a high probability of a positive diagnosis. For example, the mechanical knee exam yielded the highest positive likelihood ratios of all the tests (LR+ = 228). Similarly, the RMM estimates for the probability of a positive diagnosis were also highest with the mechanical knee exam (95.97%) compared with the other diagnostic tests. Thus, a high value on the mechanical knee exam results in the highest likelihood/probability of a positive diagnosis. Similarly, a high measure on the mechanical exam based on the RMM yields a high probability of a positive diagnosis.

Negative likelihood ratios also correlated with RMM estimates, namely the probability of no tear to the ACL. For instance, the report of an absence of a pop sensation at the time of the injury had the lowest negative likelihood ratio (2.77) of all the diagnostic tests. It also yielded the lowest RMM probability estimate that the ACL was not torn, in the event of a negative response on this test. In other words, even though an individual reported no pop sensation, the RMM probability estimate that the ACL was not torn was only 80.84%. Thus, there was still nearly a 20% chance of a tear to the ACL, even in the absence of a pop sensation. The negative likelihood ratio of 2.77 offers little change in the pretest probability that the knee is healthy. In other words, before administering the test, the clinician forms a pre-test probability estimate as to the presence/absence of the diagnosis. Following the test, the probability will either increase or decrease, depending on the result of the test.

In the case of a history of a pop, a negative response only decreases the pre-test probability minimally, with a likelihood ratio of 2.77. This is in stark contrast to a negative likelihood ratio of 139.89 based on the mechanical knee examination.

Strengths of the RMM Application to Examine the Utility of Diagnostic Tests

Interpretation Advantages. One of the primary problems with sensitivity and specificity involve the interpretations of these values. As noted previously by Guggenmoos-Holzmann and van Houwelingen (2000), sensitivity and specificity are test centered, which focuses the interpretation on the function of the test. While this is not necessarily a problem, it does not lend well to interpretation regarding the patient's actual status. In other words, sensitivity and specificity provide evidence of how well a test might accurately diagnose a person as healthy or not healthy, but it does not provide information regarding the actual status of that patient. Yet, clinicians and textbooks rely almost entirely on sensitivity and specificity values when recommending diagnostic tests (Gross, Fetto, and Rosen, 2002; Tomberlin and Saunders, 1999; Starkey and Ryan, 1996).

The values of sensitivity and specificity alone do not function for interpretation. As a guideline, it is recommended that a test with high sensitivity should be used to "rule out" a disorder because this test is so sensitive to a symptom that the absence of symptoms with the test should be viewed with confidence that a disorder does not exist (Boyko, 1994; Fritz and Waimer, 2001). Similarly, a test that has high specificity is most useful to rule in a disorder. However, an informal review of physical therapy textbooks found that tests selected for evaluation of the shoulder for instance were based entirely on sensitivity values of diagnostic tests (Gross, Fetto, and Rosen, 2002; Tomberlin and Saunders, 1999; Starkey and Ryan, 1996; Andrews and Wilk, 1994). The tests were recommended as a means to rule in (i.e., confirm the presence of the condition) shoulder impingement. For example, tests such as the Hawkins Test and Neer Test have high sensitivity values (92% and 89% respectively), but very poor specificity values, 25% and 31% respectively (Calis, Birante, et al., 2000). Thus, while these tests are recommended to rule in a diagnosis, they are actually more effective to rule out a disorder. The Hawkins test, in fact, is so sensitive that it is frequently painful even in healthy shoulders; it results in a high number of false positives (Fritz and Waimer, 2001).

The RMM probability estimates appear to provide a possible solution for this interpretation confusion. The RMM probability estimates for each diagnostic test are based on the transformation of logit values to probabilities. These logit values provide a linear alignment of the diagnostic tests along a common metric. High positive logit values indicate that the test is specific to a disorder and high negative logit values indicate the test is sensitive to a disorder. For example, in Figure 1, the diagnostic tests are aligned for comparison, based on the average logit values of the categories for each test. Mech.Exam values represent millimeters of difference between the injured and healthy knee. The top horizontal line represents the logit values for each category; the bottom horizontal line is the probability estimates based on logit values greater than and lesser than the logit value of diagnosis (-0.85 logit).

The history of a pop has the smallest positive logit value for a category of "tear" compared with the other tests. Thus, it has the lowest specificity and therefore is not as useful to rule in a diagnosis of ACL tear. In other words, the history of a pop at the time of injury provides less than a 90% probability of a tear. Looking at the probabilities associated with the category of "tear" for the other tests shows a higher probability of a tear (> 90%) given a positive outcome on these tests, compared with a history of a pop. In fact, the test with the highest specificity of all the tests, the mechanical exam, has the potential for the highest probability of a tear. A difference of nine millimeters results in a probability near 95% of a tear to the ACL. Thus, to have the greatest confidence that a patient has sustained a tear to the ACL, the clinician would wish to obtain a mechanical knee examination

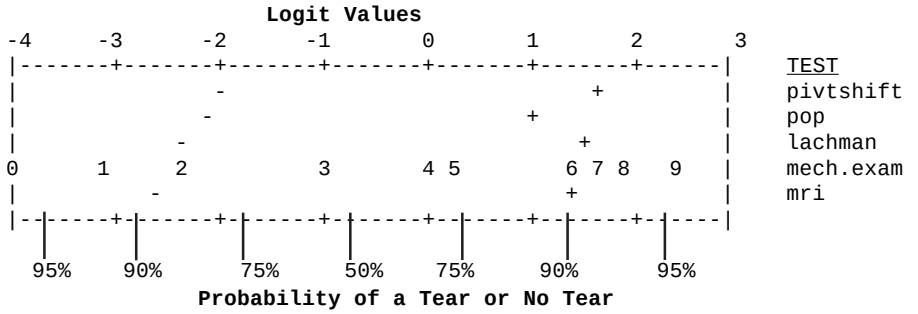

Figure 1. Alignment of Diagnostic Test Measures with the Diagnosis (50% probability represents the point at which a person with a measure at this point would have a 50% probability of a tear to the ACL). Note a "+" indicates a positive test and a "-" indicates a negative test.

with values exceeding six millimeters of difference (Figure 1).

Sensitivity of tests are interpreted with the RMM as those tests with the largest negative logit value. Once again, the mechanical knee exam has the highest sensitivity, based on the potential for the largest negative logit value, resulting in a probability of "no tear" exceeding 95%. The MRI is the next most useful test with a probability of "no tear" approximately 90% in the event of a negative MRI. The pivot shift (Figure 1), with the lowest sensitivity (i.e., it is the most difficult to obtain a positive outcome), provides the least confidence that the knee is healthy in the event of a negative outcome (80% probability).

Because of the inherent difficulty with interpretation of Sn and Sp, likelihood ratios are often used by clinicians to assist with decision making. As noted by Fritz and Wainner (2001) and Boyko (1994), likelihood ratios are the best values to illustrate the usefulness of a diagnostic test. Likelihood ratios are interpreted as the change in pre-test probability that a condition exists, given a positive or negative outcome on a diagnostic test. Thus, the pre-test probability (i.e., the probability of the diagnosis before the test is administered) changes depending on the magnitude of the positive or negative likelihood ratio, resulting in a post-test probability. Positive likelihood ratios increase the probability that a disease/diagnosis is present. The larger the value, the greater the increase in the likelihood that the disease is present. Similarly, negative likelihood ratios increase the probability that the disease/diagnosis is not present. Clinicians then use the post-test probability as the guide for clinical decision making (Frizt and Wainner, 2001).

The major difficulty with using likelihood ratios is the fact that the pre-test probability must first be estimated by the clinician. This estimated value is influenced by the clinician's experience, clinician's research background, and the accuracy of the patient's history information (Fox, Landrum-McNiff, Zhong, et al., 1999; Reid, Lane, and Feinstein, 1998; Timmermans, 1994). Error or variability in the pre-test probability directly influences the resultant post-test probability. Moreover, given that the post-test probability is the driving force behind clinical decision-making, this potential for bias can affect the final clinical decision for a patient.

The RMM provides a solution to this problem. The RMM probability estimates are based on the raw data obtained from the clinical studies. Thus, clinicians are not required to make a pre-test probability estimate that a condition exists or is absent. The probability estimates generated by the RMM provide all the information needed to make a clinical decision. These estimates represent the post-test probability that a condition is present/absent, without the need of a potentially biased pre-test probability estimate. For example, from Figure 1, it is apparent that a positive Lachman's test results in a greater post-test probability of a tear to the ACL than a positive MRI. It is also obvious that in order to have the greatest

post-test probability, a clinician may wish to use the mechanical examination, looking for values greater than six or seven millimeters of difference between the healthy and involved knee.

In addition, the RMM provides a simple method to determine the post-test probability that the knee is healthy, by examining the probabilities associated with negative tests from Figure 1. In this case, the mechanical examination again provides the greatest degree of confidence, provided the difference between the healthy and involved knee are no greater than 1 millimeter. The MRI provides the next best post-test probability relative to the other tests. The pivot shift provides the least amount of confidence that the knee is healthy, in the event of a negative test on the pivot shift, compared with the other tests. In all cases, pre-test probability estimates are not required with this procedure.

The RMM estimates, which are based on patterns of responses, allow for individual probability estimates for each test. Figure 1 provides an illustration of this ability of the RMM. Consider that a clinician wishes to be 90% sure that an event (i.e., a tear of the ACL) has occurred. The clinician can obtain this level of confidence with a positive response with the MRI, Lachman's, pivot shift, or by obtaining a difference on the mechanical exam in excess of six millimeters. To obtain 95% confidence, the only variable is the mechanical examination that can provide this level of confidence. These point estimates are readily obtained based on the linear arrangement (i.e. logit values) of the diagnostic test categories with the diagnosis outcome.

Ability to Make Direct Linear Comparisons of Diagnostic Tests. Related to the interpretation advantage of RMM is the fact that the RMM estimates are created based on a linear arrangement of the categories for each test. In other words, the category estimates for each test are aligned along a continuous scale, based on logit values. This logit arrangement allows for direct comparisons between the different diagnostic tests and the categories of the diagnostic tests.

Consider Figure 1. The estimation process used by the RMM, converting categorical data into ratio level data allows for the alignment of the diagnostic tests along a common ruler for comparison. This orientation makes the interpretation obvious that the information gained with a positive Lachman's test is quite similar to the information gained with a positive MRI. A positive pivot shift is more informative than either those two tests and a high difference on the mechanical examination is the most informative.

This feature is especially useful to avoid unnecessary medical testing to possibly minimize the costs of diagnostic testing. For example, a study by Lee, Hooker, and Harpstrite (2001) found that 62% of imaging (i.e., MRI procedures) were unjustified for knee examination. They based this judgment on the fact that the Lachman's test appeared to provide similar outcomes. Rose and Gold (1996) also found that the physical examination was equally accurate compared with the MRI to diagnose the presence of an ACL tear. They concluded that except under extreme situations, the MRI is not needed in the presence of the Lachman's and/or pivot shift examination findings. Considering the fact that MRI is nearly 200 times the expense of a manual office examination and/or mechanical examination, it is essential that clinicians have a confident and clear means to compare diagnostic test procedures.

Examining Person Diagnosis and Test Measures on a Common Ruler. Because the person abilities (i.e., health status) and diagnostic test measures (i.e., ability to make a diagnosis) are estimated as logit values, it is possible to align and compare person health measures and diagnostic test utility along a common ruler. In other words, the additive nature of the estimates allows persons and tests to be positioned along the same measurement ruler. Figure 2 is an example of such a ruler. The ruler itself is calibrated in logit values. Persons and tests are arranged along this ruler with persons positioned to the left of the vertical ruler and tests to the right of the vertical ruler.

The arrangement in Figure 2 allows for direct comparisons between a test's performance and those persons likely to obtain a positive or negative response on that test. Consider the pa-

tients positioned near the 2.0 logit value. These patients have a high probability of being diagnosed with a tear to the ACL because they are positioned far above the diagnosis measure (-.85 logits). Thus, patients with a measure near -.85 logits have a 50% probability of a tear to the ACL. The patients positioned near 2.0 logits (2.85 logits

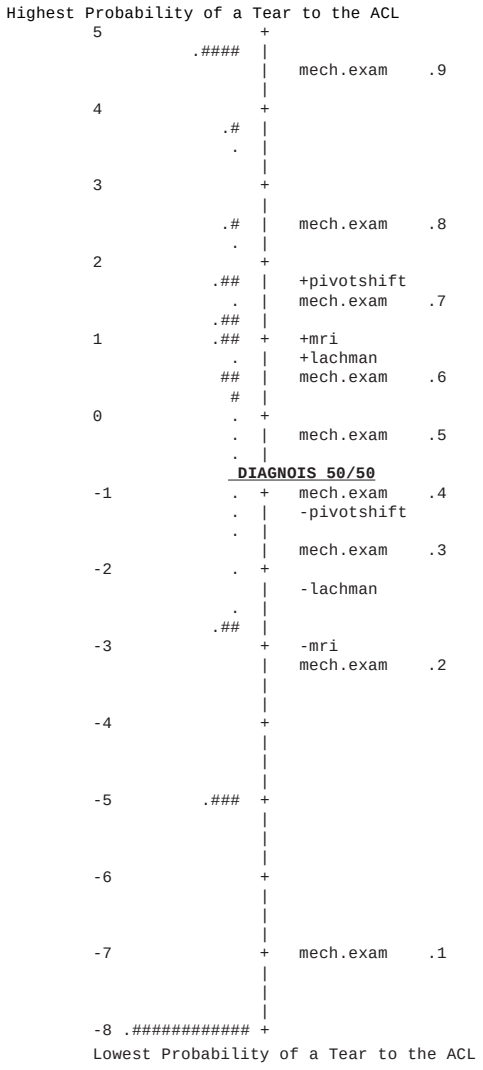

Figure 2. Arrangement of patients ("." and #) and diagnostic tests along a linear ruler calibrated in logit values, –8.0 to 5.0 (# represents 20 patients; "." represents 5 patients; mech.exam values represent mm. of difference between healthy and involved knee).

from the diagnosis measure) have a 94% probability of a tear to the ACL ([P{tear = yes} = ln(2.85) / 1 + ln(2.85)]). In addition, these persons also have a high probability (i.e., > 50%) of having been diagnosed as positive, based on the results of the Lachman's test and the MRI. However, they are less likely to obtain a positive pivot shift test; this test is more difficult that than the previous tests to obtain a positive result. Patients with a measure of approximately 2.0 logits have approximately a 50% chance of obtaining a positive result on the pivot shift test, even though they likely will have passed the Lachman and/or MRI test (i.e., diagnosed as positive on the Lachman and/or MRI).

For example, a particular patient had a measure of 1.51 logit (*SE* = .72). This patient had a positive Lachman's test and a negative MRI, but a high mechanical exam difference of 8 millimeters. Data were not available on the pivot shift test, which was not performed. In light of this data, in which the physical examination was positive yet the MRI was negative, and there was no pivot shift data available, it is possible to estimate the probability of a tear for this patient to be approximately 91.37% (CI95 = 84.25, 98.50). Based on this patient's position in Figure 2, this patient would be expected to obtain a positive Lachman's test but it would be more difficult for this patient to be diagnosed with the pivot shift test. This patient did in fact have a confirmed tear to the ACL.

The RMM and a Missing Gold Standard. An extremely important benefit of the use of the RMM to estimate diagnostic test utility is the ability to overcome the problem of missing data. Missing data in medical diagnostic tests results in several problems, including lost information from some patients and biased estimates of test performance (Bates, Margolis, and Evans, 1993; Green, Black and Johnson, 1998). Biased estimates are especially troublesome in the event that the gold standard is biased. This bias results when only a subset of patients is tested with the gold standard. This is common problem given that the gold standard is often an invasive (e.g. surgery) and/or expensive procedure. Thus, only

patients most in need of surgery or in need of an absolute diagnosis are subjected to the gold standard. The resulting estimates of sensitivity, specificity, and likelihood ratios may result in an over or under reflection of the effectiveness of the diagnostic test (Begg and Greenes, 1983; Bates, Margolis and Evans, 1993). For example, in this study, a subset of patients did not undergo arthroscopic surgery because the outcomes of the diagnostic tests (i.e. Lachman's, pivot shift, MRI) may not have been sufficiently conclusive to warrant surgery, or they simply chose to not undergo surgical repair of the ACL. A true diagnosis was not available on these patients and their clinical exam data could not have been used with traditional methods of examining the diagnostic tests. Those patients with a definitive diagnosis based on surgery likely had sufficient evidence from the diagnostic tests to warrant surgery. Thus, the sensitivity of the tests may be biased (i.e., inflated).

The RMM procedure allows for the inclusion of missing data in the analysis of patient and diagnostic test measures. This is accomplished by estimating probabilities based on response patterns rather than obtaining a count of the raw data. In the case of this study, there was no difference in the diagnostic test measures, whether the outcome variable (i.e., the actual diagnosis) was known from a reliable gold standard (surgery) or was occasionally unknown and treated as missing. Thus, the RMM was able to overcome the problem of missing data. For instance, consider Figure 3, which illustrates the positions of five patients with an unknown diagnosis. Patient numbered u838 most likely has a tear of the ACL whereas patient u851 most likely does not have a tear of the ACL. However, it is patient numbered u859 who poses a dilemma. Based on the physical exam, this patient still has a 50% chance of a tear. Thus, in this patient's case, a more definitive test would be warranted. This patient had not been tested with the mechanical knee examination or an MRI. The mechanical knee examination would be the test of choice because it is a better predictor of knee health and it is much less expensive than the MRI.

This advantage of the RMM will allow diagnosticians to examine the utility of diagnostic tests in the absence of a perfect gold standard. No longer will estimates of test sensitivity, specificity, and likelihood ratios require a perfect gold standard in order to examine the utility of a diagnostic test or tests. Instead, probability estimates from the RMM can be produced to provide evidence of the utility of diagnostic tests.

Reliability Estimates using the RMM. The last feature of the RMM that provides an additional advantage over standard procedures to examine the utility of diagnostic tests is the reliability estimates produced by the RMM. Standard procedures typically use agreement indices

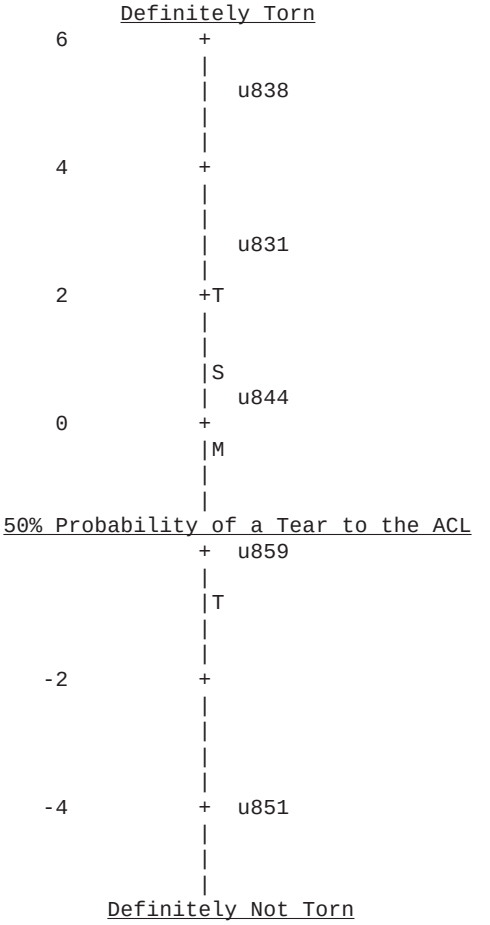

Figure 3. Linear Arrangement of Person Measures for five patients with an unverified diagnosis.

such as the kappa statistic, as a means to estimate reliability (Fritz and Wainner, 2001; Portney and Watkins, 2000). This statistic provides the measure of agreement between a two observation sessions. For instance, researchers might compare the observations from the Lachman's test on two separate occasions, using the same investigator. This approach requires repeat testing of the patients for reliability estimation. In addition, for continuous data, such as that generated by a mechanical knee examination, the reliability estimate is in the form of the intraclass correlation coefficient (Portney and Watkins, 2000). Both procedures provide a reliability of the outcome estimate (i.e., a function of the test), but neither provide a reliability estimate of the person measures/observations. In other words, while the reliability of the outcome decisions can be estimated, these procedures do not allow for an estimate of the reliability that a person's response pattern will be stable.

The RMM provides two important measures of reliability, a reliability index and a separation index. The person separation index, G_p, is an indication as to how well the persons are sufficiently separated into different levels of ability by the diagnostic tests. Just as a math test should be able to identify persons of more and less ability, separating the persons into ability levels, diagnostic tests should be able to identify those persons most likely to have the diagnosis and those least likely to have the diagnosis. This separation index should identify at least two distinct strata. If the sample of persons are not separable into different levels of health (i.e., ability), the diagnostic tests failed to identify persons with a positive or negative diagnosis. Essentially, this separation index is actually a measure of validity, that the tests are in fact measuring the persons based on some diagnosis.

Wright and Masters (1982) provide a simple extension of the person separation index to identify statistically significant strata (H_p). This estimate, $H_p = (4G_p + 1) / 3$, provides the statistically significant number of distinct strata in a sample of persons, as identified by the diagnostic tests. In the case of the sample of 825 patients of this study, the separation index was $G_p = 1.82$; the statistically distinct number of strata, H_p, was 2.76, indicating that the diagnostic tests were able to identify two fully distinct levels of the patients (i.e., those with a tear and those without a tear). Figure 4 represents this distinct separation. In Figure 4, the distribution of patients is presented along the bottom of the horizontal line, representing the distribution of patients with a tear to the ACL and patients without a tear to the ACL.

Closely related to the separation index is the reliability index, R_p, which is a measure of the stability of the person measures and the ability of the diagnostic tests to separate persons into distinct strata of health. This index, loosely estimated as $G^2_p / 1 + G^2_p$ provides the reliability estimate that repeat testing would yield similar outcomes. For instance the person reliability of this investigation (n = 825) was $R_p = .77$. In addition, the RMM provides an estimate of the test reliability (i.e., reliability that the test measures will remain stable over repeat applications). In

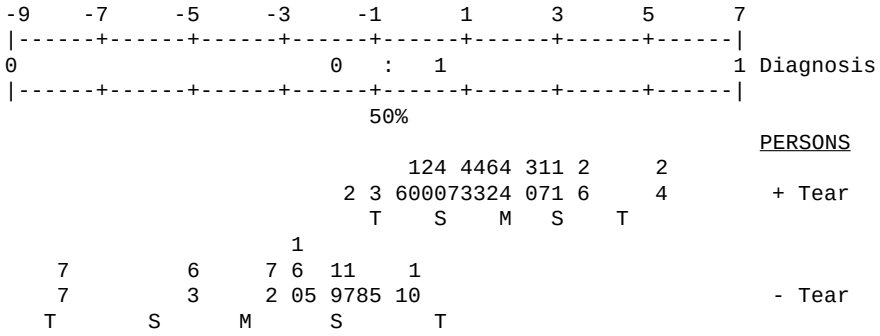

Figure 4. Distribution of patients ("+" Tear and "–" Tear), demonstrating the separation of diagnoses identified by the diagnostic tests.

this investigation, the test was very high, R_i = .98, which is a reflection of the large sample size.

Thus, the RMM provides estimates that provide validity and reliability estimates for both the sample of persons as well as the tests (items) used to measure the persons. The separation index is an indication of validity of the inferences, that the tests were able to diagnose differences in patients; the reliability index is an indication of the stability of the measures. Standard procedures do not provide this level of information regarding diagnostic test utility.

Conclusion

The interpretations of diagnostic test outcomes, based on probability estimates, sensitivity, specificity, and likelihood ratios were comparable between the different approaches to examine diagnostic test utility. The RMM provided estimates of a test's utility that were comparable to estimates of sensitivity, specificity, and likelihood ratios. In addition, the RMM estimates did not change whether the outcome variable (i.e., actual diagnosis) was known, ordinal, or contained missing values.

The RMM was demonstrated to provide simple interpretations of a diagnostic test's meaning regarding a patient's status. These interpretations, in the form of probabilities, did not require subjective estimation of pre-test probabilities, and provided a direct estimate of the probability of the presence/absence of the diagnosis. Further, because the estimates were based on a linear alignment of the tests, based on logit values, direct comparisons could be made between diagnostic tests to determine which test provided the most/least information for a patient's diagnosis.

It was shown that the RMM has the potential to overcome the limitations of a biased (i.e., missing) gold standard and missing data in general. The RMM was able to generate estimates for patients and diagnostic tests with missing data in the data set.

Finally, it was shown that the RMM provided unique measures of validity and reliability evidence to support that utility of diagnostic tests. The separation index in particular was shown to be a useful measure that the diagnostic tests functioned to identify different strata of health in the sample—that is, the tests sufficiently separated persons into a diagnosis of a tear to the ACL or no tear to the ACL.

Future research is warranted to examine the use of the RMM to examine person and item fit measures in terms of diagnostic testing. Person fit in particular, as estimated by the RMM, could be potentially useful in order to identify and eliminate patient data from person's not providing a valid test response—a condition referred to as symptom magnification, in which a patient essentially magnifies, fakes, or distorts symptoms for some secondary gain. Data from these individuals bias the estimated utility of a diagnostic test.

References

Andrews, J. R., and Wilk, K. E. (1994). *The athlete's shoulder*. New York: Churchill Livingstone.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Bates, A. S., Margolis, P. A., and Evans, A. T. (1993). Verification bias in pediatric studies evaluating diagnostic tests. *Journal of Pediatrics, 122*, 585-590.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics, 39*, 207-215.

Beltyukova, S., Cipriani, D., Yan, S., Ughrin, T., and Fox, C. (2000). Rasch regression predicts driving capability. *Rasch Measurement Transactions, 15*, 789-790.

Bode, C. (2001). Understanding Rasch measurement: partial credit model and pivot anchoring. *Journal of Applied Measurement, 2*, 78-95.

Bohannon, R. W. (1987). Simple clinical measures. *Physical Therapy, 67*(12), 1845-1850.

Bond, T. G., and Fox, C. M. (2001) *Applying the Rasch model: Fundamental measurements in the human sciences*. Mahwah, NJ: Erlbaum.

Boyko, E. J. (1994). Ruling out or ruling in disease with the most sensitive or specific diagnostic test: Short cut or wrong turn*? Medical Decision Making, 14*, 175-179.

Campbell, S. K., Kolobe, T. H., Osten, E. T., Lenke, M., and Girolami, G. L. (1995). Construct validity of the Test of Infant Motor Performance. *Physical Therapy, 75*, 585-596.

Cahs, M., Akgun, K., Birtane, M., Karacan, I., Cahs, H., and Tuzun, F. (2000). Diagnostic values of clinical diagnostic tests in subacromial impingement syndrome. *Annals Rheum Dis, 59*, 44-47.

Chang, W. C., and Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*, 934-939.

Chang, W. C., Slaughter, S., Cartwright, D., and Chan, C. (1997). Evaluating the FONE FIM: Part I. Construct validity. *Journal of Outcome Measurement, 1*, 192-218.

Creel, G. L., Light, K. E., and Thigpen, M. T. (2001). Concurrent and construct validity of scores on the Timed Movement Battery. *Physical Therapy, 81*, 789-798.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

de Bock, G. H., Houwing-Duistermaat, J. J., Springer, M.P., Kievit, J., and van Houwelingen, J. C. (1994). Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. *Journal of Clinical Epidemiology, 47*, 1343-1352.

DeHaven, K. E. (1983). Arthoscopy in the diagnosis and management of the anterior cruciate ligament deficient knee. *Clin Orthop Rel Res, 172*, 52-56.

Fagan, T. J. (1975). Nomogram for Bayes's theorem. *New England Journal of Medicine, 293*, 257.

Fisher, A. G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *AJOT, 47*, 319-329.

Fisher, W. P., Jr. (1993). Measurement-related problems in functional assessment. *AJOT, 47*, 331-338.

Fisher, W. P., Harvey, R. F., Taylor, P., Kilgore, K. M., and Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Arch of Physical Medicine and Rehabilitation, 76*, 113-122.

Forster, I. W., Warren-Smith, M. T., and Tew, M. (1989). Is the KT1000 knee ligament arthrometer reliable? *J Bone Joint Surgery (Br), 71B*, 843-847.

Fox, C. (1999). An introduction to the partial credit model for developing nursing assessments. *Journal of Nursing Education, 38*, 340-346.

Fox, C. M., and Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*, 30-45.

Fox, E., Landrum-McNiff, K., Zhong, Z., Dawson, N. V., Wu, A.W., and Lynn, J. (1999). Evaluation of prognostic criteria for determining hospice eligibility in patients with advanced lung, heart, or liver disease. *JAMA, 282*, 1638-1645.

Fritz, J. M., and Wainner, R. S. (2001). Examining diagnostic tests: An evidence-based perspective. *Physical Therapy, 81*, 1546-1564.

Green, T. A., Black, C. M., and Johnson, R. E. (1998). Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *J of Clin Microbiology,36*, 375-381.

Gross, J., Fetto, J., and Rosen, E. (2002). *Musculoskeletal examination* (2nd ed.). Malden, MS: Blackwell Publishing.

Guggenmoos-Holzmann, I, and van Houwelingen, H. C. (2000). The (In)validity of sensitivity and specificity. *Statistics in Medicine, 19*, 1783-1792.

Haley, S. M., McHorney, C. A., and Ware, J. E., Jr. (1994). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *J of Clinical Epidemiology, 47*, 671-684.

Harada, N., Chiu, V., Damron-Rodriguez, J., Fowler, E., Siu, A., and Reuben, D. B. (1995). Screening for balance and mobility impairment in elderly individuals living in residential care facilities. *Phys Ther, 75*, 462-469.

Hawkins, D. M., Garrett, J. A. and Stephenson, B. (2001). Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine, 20*, 1987-2001.

Heinemann, A. W., Harvey, R. L., McGuire, J. R., Ingberman, D., Lovell, L., Semik, P, et al. (1997). Measurement properties of the NIH Stroke Scale during acute rehabilitation. *Stroke, 28*, 1174-1180.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., and Granger, C. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74*, 566-573.

Hlatky, M. A., Pryor, D. B., Harrell, F. E., Jr., Califf, R. M., Mark, D. B., and Rosati, R. A. (1984). Factors affecting sensitivity and specificity of exercise electrocardiography. *American Journal of Medicine, 77*, 64-71.

Indrayan, A., and Sarmukaddam, S. B. (2001). *Medical biostatistics*. New York: Marcel Dekker.

Irwig, L., Glasziou, P. P., Berry, G., Chock, C., Mock, P., and Simpson, J. M. (1994). Efficient study designs to assess the accuracy of screening tests. *Am Journal of Epidemiology, 140*, 759-769.

Jakob, R. P., Staubli, H. U., and Deland, J. T. (1987). Grading the pivot shift: objective tests with implications for treatment. *J Bone Joint Surgery, 69B*, 294-299.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology, 141*, 263-272.

Karabatsos, G. (1997). The sexual experiences survey: Interpretation and validity. *J of Outcome Measurement, 1*, 305-328.

Karabatsos, G. (2001). Understanding Rasch measurement: The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*, 389-423.

Kirkley, S. (1997). A comparison of accuracy between clinical examination and magnetic resonance imaging in the diagnosis of meniscal and anterior cruciate ligament tears. *Arthroscopy, 13*, 279-280.

Klauer, K. C. (1995). The assessment of person fit. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97-110). New York: Springer-Verlag.

Lai, J.-S., Fisher, A. G., Magalhães, L. C., and Bundy, A. C. (1996). Construct validity of the Sensory Integration and Praxis Tests. *The Occupational Therapy Journal of Research, 16*, 75-97.

Lee, G., Hooker, M., and Harpstrite, K. (2001). Magnetic resonance imaging in a military setting: A utilization analysis. *Military Medicine, 166*, 126-131.

Lilienfeld, D. R., and Stolley, P. D. (1994). *Foundations of epidemiology* (Rev. ed.). New York: Oxford University Press.

Linacre, J. M., and Wright, B. D. (1991). *A User's Guide to Winsteps: Rasch-Model Computer Programs.* Chicago: MESA Press.

Losse, R. E. (1983). Concepts of the pivot shift. *Clinical Orthopaedics and Related Research, 172*, 45-51.

MacKnight, C., and Rockwood, K. (2000). Rasch analysis of the hierarchical assessment of balance and mobility (HABAM). *Journal of Clinical Epidemiology, 53*, 1242-1247.

Malcom, L. L, Daniel, D. M., Stone, M. L., Sachs, R. (1985). The measurement of anterior knee laxity after ACL reconstructive surgery. *Clinical Orthopedics and Related Research, 196, 35-41.*

Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S.E. Embretson and S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 129-152). Mahwah, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McNamara, T. (1996). Concepts and procedures in Rasch measurement. In *Measuring Second Language Performance* (pp. 149-181). Boston: Addison Wesley.

Morris, S., Morris, M. E., and Iansek, R. (2001). Reliability of measurements obtained with the Timed "Up and Go" Test in people with Parkinson disease. *Physical Therapy, 81*, 810-818.

Neeb, T. B., Aufdemkampe, G., Wagener, J. H., and Mastenbroek, L. (1997). Assessing anterior cruciate ligament injuries: the association and differential value of questionnaires, clinical tests, and functional tests. *J Orthop Sports Phys Ther, 26,* 324-331.

Phelps, C. E., and Hutson, A. (1995). Estimating diagnostic test accuracy using a "fuzzy gold standard". *Medical Decision Making, 15*, 44-57.

Portney, L. G. and Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice,* (2nd ed.). Upper Saddle River, NJ: Prentice Hall Health.

Prieto, L., Roset, M., and Badia, X. (2001). Rasch measurement in the assessment of growth hormone deficiency in adult patients. *Journal of Applied Measurement, 2*, 48-67.

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

Reid, M. C., Lachs, M. S., and Feinstein, A. R. (1995). Use of methodological standards in diagnostic test research: Getting better but still not good. *JAMA, 274*, 645-651.

Reid, M. C., Lane, D. A., and Feinstein, A. R. (1998). Academic calculations versus clinical judgements: Practicing physicians' use of quantitative measures of test accuracy. *American Journal of Medicine, 104*, 374-380.

Rheault, W., and Coulson, E. (1991). Use of the Rasch model in the development of a clinical competence scale. *Journal of Physical Therapy Education, 5*, 10-13.

Rose, N. E., and Gold, S. M. (1996). A comparison of accuracy between clinical examination and magnetic resonance imaging in the diagnosis of meniscal and ACL tears. *Arthroscopy, 12,* 398-405.

Rothman, K. J., and Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott Williams and Wilkins.

Sackett, D. L. (1992). A primer on the precision and accuracy of the clinical examination. *JAMA, 267*, 2638-2644.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., and Tugwell, P. (1992). *Clinical Epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little Brown and Co.

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Silverstein, B., Fisher, W. P., Kilgore, K. M., Harley, J. P., and Harvey, R. F. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. *Archives of Physical Medicine and Rehabilitation*, 73, 507-518.

Simel, D. L., Feussner, J. R., Delong, E. R., and Matchar, D. B. (1987). Intermediate, indeterminate, and uninterpretable diagnostic test results. *Medical Decision Making, 7*, 107-114.

Smith, E. V. (2001). Understanding Rasch measurement: evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement prespective. *Journal of Applied Measurement, 2,* 281-311.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1,* 199-218.

Solomon, D. H., Simel, D. L., Bates, D. W., Katz, J. N., and Schaffer, J. L. (2001). Does this patient have a torn meniscus or ligament of the knee? Value of the physical examination. *J Am Medical Assoc, 286,* 1610-1620.

Sox, H. C. (1996). The evaluation of diagnostic tests: Principles, problems, and new developments. *Ann. Rev. Med., 47,* 463-471.

Starkey, C. and Ryan, J. (1996). *Evaluation of musculoskeletal and athletic injuries.* Philadelphia: F.A. Davis.

Tesio, L., Granger, C. V., and Fiedler, R. C. (1997). A unidimensional pain/disability measure for low-back pain syndromes. *Pain, 69,* 269-278.

Timmermans, D. (1994). The roles of experience and domain of expertise in using numerical and verbal probability terms in medical decisions. *Medical Decision Making, 14,* 146-156.

Tomberlin, J. P., and Saunders, H. D. (1999). *Evaluation, treatment and prevention of musculoskeletal disorders: Extremities* (3rd ed.). Chaska, MN: The Saunders Group.

Valenstein, P. N. (1990). Evaluating diagnostic tests with imperfect standards. *Am J of Clinical Pathology, 93,* 252-258.

Velozo, C. A., Kielhofner, G., and Lai, J. S. (1999). The use of Rasch analysis to produce scale-free measurement of functional ability. *AJOT, 53,* 83-90.

Velozo, C. A., Magalhaes, L. C., Pan, A. W., and Leiter, P. (1995). Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation, 76,* 705-712.

Woodward, M. (1999). *Epidemiology: Study design and data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Wright, B. D., and Masters, G. N. (1982*). Rating scale analysis*. Chicago: MESA.

Wright, B. D., Perkins, K., and Dorsey, K. (2000). Multiple regression via measurement. *Rasch Measurement Transactions, 14,* 729-731.