Journal of Applied Measurement Abstracts

Volume 5, Number 1 (2004)

Establishing Mathematical Laws of Genomic Variation

Nathan J. Markward
V. A. Metrics, Inc.

As the biological arm of the Rasch community, genomic measurement is concerned with asserting and testing hypotheses regarding the quantitative status of genomic variables, including alleles, genotypes, gene expression levels, and phenotypes, as well as DNA, RNA, and protein sequence information. The defining goal of this scientific paradigm, in contrast to the sample-dependent model-fitting and deterministic hypothesis testing of classical statistical genetics, is the identification, validation, and maintenance of a common unit of genomic measurement that maintains its magnitude and meaning, within an allowable range of error, regardless of the laboratory technology used to generate outcomes or the particular group of individuals or organisms under investigation. Such an invariant metric, the basis of a standard genometric scale and associated system of genomic metrology, can be identified, validated, and maintained through 1) routine implementation of the Rasch family of measurement models to construct sample- and scale-free measures from different types of genomic data and 2) cross-calibration of genomic measurement instruments between and among researchers, laboratories, universities, corporations, and databases. This manuscript provides an introductory overview of the guiding principles of fundamental measurement theory and the work of Rasch, connects these concepts to well-known tenets of population genetics, and highlights the potential benefits, both theoretical and applied, associated with achieving objectivity in genomic measurement.

Comparing Traditional and Rasch Analyses of the Mississippi PTSD Scale: Revealing Limitations of Reverse-Scored Items

Kendon J. Conrad
Midwest Center for Health Services and Policy Research,
Hines VA Hospital and University of Illinois at Chicago
Benjamin D. Wright
University of Chicago
Patrick McKnight
University of Arizona
Miles McFall
VA Puget Sound Health Care System
Alan Fontana
VA Connecticut Healthcare System
Yale University School of Medicine
Robert Rosenheck
VA Connecticut Healthcare System
Yale University School of Medicine

This study examined whether Rasch analysis could provide more information than true score theory (TST) in determining the usefulness of reverse-scored items in the Mississippi Scale for Posttraumatic Stress Disorder (M-PTSD). Subjects were 803 individuals in inpatient PTSD units at 10 VA sites. TST indicated that the MPTSD performed well and could be improved slightly by deleting one item. Factor analysis using raw scores indicated that the reverse-scored items formed the second factor and had poor relationships with normally scored items. However, since item-total correlations supported their usefulness, they were kept. The subsequent Rasch analysis indicated that five of the seven worst fitting items were reverse-scored items. We concluded that

using reversed items with disturbed patients can cause confusion that reduces reliability. Deleting them improved validity without loss of reliability. The study supports the use of Rasch analysis over TST in health research since it indicated ways to reduce respondent burden while maintaining reliability and improving validity.

Evaluating Judge Performance in Sport

Marilyn A. Looney
Northern Illinois University

Many sports, such as, gymnastics, diving, ski jumping, and figure skating, use judges' scores to determine the winner of a competition. These judges use some type of rating scale when judging performances (e.g., figure skating: 0.0 - 6.0). Sport governing bodies have the responsibility of setting and enforcing quality control parameters for judge performance. Given the judging scandals in figure skating at the 1998 and 2002 Olympics, judge performance in sport is receiving greater scrutiny. The purpose of this article is to illustrate how results from Rasch analyses can be used to provide in-depth feedback to judges about their scoring patterns. Nine judges' scores for 20 pairs of figure skaters who competed at the 2002 Winter Olympics were analyzed using a four-faceted (skater pair ability, skating aspect difficulty, program difficulty, and judge severity) Rasch rating scale model that was not common to all judges. Fit statistics, the logical ordering of skating aspects, skating programs, and separation indices all indicated a good fit of the data to the model. The type of feedback that can be given to judges about their scoring pattern was illustrated for one judge (USA) whose performance was flagged as being unpredictable. Feedback included a detailed description of how the rating scale was used; for example, 10% of all marks given by the American judge were unexpected by the model ($Z > |2|$). Three figures illustrated differences between the judge's observed and expected marks arranged according to the pairs' skating order and final placement in the competition. Scores which may represent "nationalistic bias" or a skating order influence were flagged by looking at these figures. If sport governing bodies wish to improve the performance of their judges, they need to employ methods that monitor the internal consistency of each judge as a many-facet Rasch analysis does.

The Effect of Sample Size for Estimating Rasch/IRT Parameters with Dichotomous Items

Mark Stone
The Adler School of Professional Psychology
Futoshi Yumoto
The Stoelting Company

Thirteen samples were randomly drawn from the normative database for the latest edition of Knox's Cube Test-Revised (KCT-R). Parameter estimates for the Rasch model and two and three parameter logistic models were derived and compared. Sample size influenced these estimates as might be expected. Rasch parameter estimates consistently showed the smallest values by sample size using a goodness of fit index.

Equating Student Satisfaction Measures

Svetlana A. Beltyukova
Gregory E. Stone
Christine M. Fox
The University of Toledo

Colleges and universities conduct student satisfaction studies for many important policy making reasons. However the differences in instrumentation and the use of students' self-reported ratings of satisfaction make such decisions sample-, instrument-, and institution-dependent. A common metric of student satisfaction would assist decision makers by providing a richness of information not typically obtained. The present study investigated the extent to

which two nationally known instruments of student satisfaction could be scaled on the same quantitative metric. Pseudo-common item equating (Fisher, 1997) based on five link items of low and high endorsability enabled comparisons of "similar, but not identical items, from different instruments, calibrated on different samples" (p. 87). Results suggest that both instruments measured similar constructs and could be reasonably used to create a single, common metric. While samples used in the experiment were less than ideal, results clearly demonstrated the usefulness and reasonability of the pseudo-common item equating process.

Treating Test-Item Nonresponse

Hamish Coats
University of Melbourne

This study presents the results of an empirical investigation into the effects of nonresponse on the measurement of student ability in large scale educational achievement studies. The analyses explored the bias in student and national ability estimate distributions relating to particular psychometric treatments of nonresponse. A range of replicated analyses were undertaken using data collected from a cross-national study of reading achievement. The relative merits of these treatments are summarized in conclusion, and implications for normative and methodological research and practice are considered. It is suggested that environmental and psychological antecedents of nonresponse need to be determined and that related variables be included as essential components in item response modelling.

Rasch Model Estimation: Further Topics

John M. Linacre
University of the Sunshine Coast, Australia

Building on Wright and Masters (1982), several Rasch estimation methods are briefly described, including Marginal Maximum Likelihood Estimation (MMLE) and minimum chi-square methods. General attributes of Rasch estimation algorithms are discussed, including the handling of missing data, precision and accuracy, estimate consistency, bias and symmetry. Reasons for, and the implications of, measure misestimation are explained, including the effect of loose convergence criteria, and failure of Newton-Raphson iteration to converge. Alternative parameterizations of rating scales broaden the scope of Rasch measurement methodology.

Volume 5, Number 2

The Impact of Model Misfit on Partial Credit Model Parameter Estimates

Randall D. Penfield
University of Florida

The partial credit model (PCM) is commonly employed to parameterize items and individuals using responses to a set of polytomous items. Because the PCM does not include a discrimination parameter, it may encounter substantial lack of fit to the data in certain situations. To determine the impact of model misfit on the estimation of person and item parameters using the PCM, a simulation study was conducted in which data were generated according to the generalized partial credit model, and the bias and efficiency of the resulting person and item parameter estimates were assessed. The results suggest that small amounts of unsystematic misfit do not lead to dramatic levels of bias or loss of efficiency of the estimators, but large levels of unsystematic misfit and moderate levels of systematic misfit result in substantial loss of efficiency and bias of the estimators.

Calibrating the Genome

Nathan J. Markward
V. A. Metrics, Inc.
William P. Fisher, Jr.
MetaMetrics, Inc.

Purpose: This project demonstrates how to calibrate different samples and scales of genomic information to a common scale of genomic measurement. Materials and Methods: 1,113 persons were genotyped at the 13 Combined DNA Index System (CODIS) short tandem repeat (STR) marker loci used by the Federal Bureau of Investigation (FBI) for human identity testing. A measurement model of form $\ln[(P_{nik})/(1-P_{nik})] = B_n - D_i - L_k$ is used to construct person measures and locus calibrations from information contained in the CODIS database. Winsteps (Wright and Linacre, 2003) is employed to maximize initial estimates and to investigate the necessity and sufficiency of different rating classification schema. Results: Model fit is satisfactory in all analyses. Study outcomes are found in Tables 1-6. Conclusions: Additive, divisible, and interchangeable measures and calibrations can be created from raw genomic information that transcend sample- and scale dependencies associated with racial and ethnic descent, chromosomal location, and locus-specific allele expansion structures.

Comparisons of Mathematics Achievement of Grade 8 Students in the United States and the Russian Federation

Saodat I. Bazarova
George Engelhard, Jr.
Emory University

Using the Mantel-Haenszel (MH) Procedure, we analyzed data for 7,087 American and 4,022 Russian Grade 8 students from the Third International Mathematics and Science Study (TIMSS) to compare mathematics achievement in the two countries on each of the 124 multiple-choice items. The results of the analyses indicate that the performance of the students on individual multiple-choice mathematics items vary by country. The results also suggest that the relationship between country and item performance differ as a function of content area. A total score of a country's achievement does not provide the whole picture of achievement dynamics; it averages out potentially important information on student achievement and the causes of their performance relative to other countries. The dynamics of achievement across countries will not be revealed unless the analyses are done at the item level.

A Rasch Analysis of Three of the Wisconsin Scales of Psychosis Proneness: Measurement of Schizotypy

Roger E. Graves
University of Victoria
Sara Weinstein
University of British Columbia

Rasch analyses were conducted with data from 90 university students on three of the Wisconsin Scales of Psychosis Proneness—the Magical Ideation (Eckblad and Chapman, 1983), Perceptual Aberration (Chapman, Chapman, and Raulin, 1978), and Revised Social Anhedonia Scales (Eckblad, Chapman, Chapman, and Mishlove, 1982). All of the items, for each of the individual scales, plus all of the items from the combined Perceptual Aberration/Magical Ideation (Per-Mag) Scale, showed satisfactory fit to the Rasch model. These results show that personality traits including these psychosis proneness, or schizotypy, traits can be measured on a theoretically sound quantitative interval scale. Rasch scale equivalents for raw scores are provided. Possible improvements to the Magical Ideation, Perceptual Aberration, and Per-Mag scales are suggested by the item analysis. Advantages of Rasch scaling for clinical applications include detection of invalid test protocols, more meaningful interpretations of test scores, and direct comparison of scores from different tests of the same construct.

Evaluation of the 0.3 Logits Screening Criterion in Common Item Equating

G. Edward Miller
Texas Education Agency
Ourania Rotou
Educational Testing Service
Jon S. Twing
Pearson Educational Measurement

A number of state assessment programs that employ Rasch-based common item equating procedures estimate the equating constant with only those common items for which the two tests' Rasch item difficulty parameter estimates differ by less than 0.3 logits. The results of this study presents evidence that this practice results in an inflated probability of incorrectly dropping an item from the common item set if the number of examinees is small (e.g., 500 or less) and the reverse if the number of examinees is large (e.g., 5000 or more). An asymptotic experiment-wise error rate criterion was algebraically derived. This same criterion can also be applied to the Mantel-Haenszel statistic. Bonferroni test statistics were found to provide excellent approximations to the (asymptotically) exact test statistics.

The Effect of Dropping Low Scores on Ability Estimates

Ryan P. Bowles
University of Virginia

Dropping low scores is a common technique used in combining scores from multiple assessments, but no research has addressed the validity of ability estimates when dropping low scores. Using a simulation approach, the expected bias, root mean squared error (RMSE), and benefit to examinees resulting from dropping low scores was estimated for two ability estimation methods, Rasch estimation and proportion correct scoring. The simulation was done for three testing conditions: a normal condition; a bad day condition, when an examinee has a lower ability on one assessment; and a bad test condition, when one assessment is contaminated by an irrelevant factor. Ability estimates based on the complete data were generally preferable to estimates based on data with low scores dropped, suggesting that the use of dropping low scores is not warranted in most assessment situations.

Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II

Carol M. Myford
University of Illinois at Chicago
Edward W. Wolfe
Michigan State University

The purpose of this two-part paper is to introduce researchers to the many-facet Rasch measurement (MFRM) approach for detecting and measuring rater effects. In Part II of the paper, researchers will learn how to use the Facets (Linacre, 2001) computer program to study five effects: leniency/severity, central tendency, randomness, halo, and differential leniency/severity. As we introduce each effect, we operationally define it within the context of a MFRM approach, specify the particular measurement model(s) needed to detect it, identify group- and individual-level statistical indicators of the effect, and show output from a Facets analysis, pinpointing the various indicators and explaining how to interpret each one. At the close of the paper, we describe other statistical procedures that have been used to detect and measure rater effects to help researchers become aware of important and influential literature on the topic and to gain an appreciation for the diversity of psychometric perspectives that researchers bring to bear on their work. Finally, we consider future directions for research in the detection and measurement of rater effects.

Volume 5, Number 3

Equating Rehabilitation Outcome Scales: Developing Common Metrics

Richard M. Smith
Data Recognition Corporation
Patricia A. Taylor
Rehabilitation Institute of Chicago

Transparency with regard to measuring devices is one of the fundamental requirements for progress in science. The ability to derive comparable measures from different measuring devices is the cornerstone of transparency. To this end, progress in measuring and understanding rehabilitation outcomes requires that there is a method of measuring outcomes that is independent of the particular collection of items that is used to assess the outcomes. The purpose of this study is to develop a equivalence between the PECS Motor Skills and Cognition and Communication LifeScales with the FIM Motor Skills and Cognitive items. However, only the results of the Motor Skills Scale are reported here in the interest of brevity. This equating is based on approximately 500 simultaneous evaluations using bout the PECS and FIM scales on admission and discharge. The patients in this study were consecutive admissions to a free-standing rehabilitation hospital in early 1998. Patients from five diagnostic groups were included in this study, Brain Injury, Spinal Cord Injury, Stroke, Neuromuscular, and Musceloskeletal. The results indicate that it is possible to construct a common equal interval translation between the PECS and FIM for the two scales. Measures on the common metric can be based to either scale and are independent of the number of items completed. This use of these anchored scales will allow institutions using either the PECS or FIM to make direct comparisons of clinical outcomes with other institutions, independent of the particular outcome tool used to evaluate patients.

Using Rasch Models to Reveal Contours of Teachers' Knowledge

Constantia Hadjidemetriou
Julian Williams
The University of Manchester

Teachers' knowledge is usually categorized into subject matter (SMK) and pedagogical content knowledge (PCK). Previously, measurement instruments and consequent cognitive scales have been developed to assess students' and teachers' subject knowledge. A number of qualitative studies have explored teachers' pedagogical content knowledge. This study developed a means to investigate one aspect of PCK—teachers' awareness of their students' knowledge—using a combination of measurement and qualitative interpretation. We asked teachers to estimate on a Likert scale (and also describe qualitatively) the difficulty their pupils would have with test items which we had already scaled using data from their pupils. We then constructed, using various models, a "Teacher's collective Perception of Item Difficulty" (TPID) scale and contrasted this with the student's ability scale by comparing the two sets of item-difficulty parameters. The results were triangulated with qualitative data. We suggest the methodology is best supported by an Inverse Partial Credit Model (IPCM) but we compare the results across alternative Rasch models.

Validation of Scores from Self-Learning Scales for Primary Students Using True-Score and Rasch Measurement Methods

Magdalena Mo Ching Mok
The Hong Kong Institute of Education

The validation of scores from the Self-learning Scales for primary pupils is presented in this study. The sample for the study comprised 1253 pupils from 20 Year-3 and 20 Year-5 classes from ten primary schools in Hong Kong. The 10-item Usefulness Scale is designed to measure primary pupils' attitudes toward the usefulness of self-learning strategies situated in ten learning contexts. The 10-item Deployment Scale is designed to measure pupils' frequency in using the self-learning strategies. Both scales use 3-point Likert response scale. Construct validity of scores from the scales for use with primary pupils is supported by confirmatory factor analysis and Rasch measurement. Gender and year level differences were identified on the Rasch person measures. Generalizability of the scores from the two scales across gender and year level needs to be undertaken with caution.

Reporting the Incidence of School Violence across Grade Levels in the U.S. Using the Third International Mathematics and Science Study (TIMSS)

Lei Yu
Educational Testing Service

School violence has increasingly captured public attention due to deadly school shootings. Controversy on school violence is demonstrated by a mixed picture of school safety and the lack of consensus on the definitions of violence, which makes comparison of findings across studies difficult. This study extended the application of the Rasch model to school violence research using TIMSS data. The results show that school violence occurred at a level much lower than the predictions of the measurement model. Across all grade levels the most frequently reported type of violence is intimidation or verbal abuse of students and the least frequently reported physical injury to teachers or staff.

Pre-Equating: A Simulation Study Based on a Large Scale Assessment Model

Husein M. Taherbhai
Michael J. Young
Harcourt Assessment

Although post-equating (PE) has proven to be an acceptable method in the scaling and equating of items and forms, there are times when the turn-around period for equating and converting raw scores to scale scores is so small that PE cannot be undertaken within the prescribed time frame. In such cases, pre-equating (PrE) could be considered as an acceptable alternative. Assessing the feasibility of using item calibrations from the item bank (as in PrE) is conditioned on the equivalency of the calibrations and the errors associated with it vis a vis the results obtained via PE. This paper creates item banks over three periods of item introduction into the banks and uses the Rasch model in examining data with respect to the recovery of item parameters, the measurement error, and the effect cut-points have on examinee placement in both the PrE and PE situations. Results indicate that PrE is a viable solution to PE provided the stability of the item calibrations are enhanced by using large sample sizes (perhaps as large as full-population) in populating the item bank.

The Equivalence of Three Data Collection Methods with Field Test Data: A FACETS Application

Mark Pomplun
Michael Custer
Riverside Publishing Company

The present study demonstrated the utility of the FACETS software for evaluating items in the field test stage of item development for a clinical early childhood instrument. The research focus was the equivalence of the parent/caretaker interview, structured assessment, and observational methods for data collection for a

developmental inventory for children from birth to age seven. Data for this study were from a field test with some missing responses. The Rasch-based software FACETS was used to test the equivalence of the methods of data collection as well as to identify items for which the methods did not provide equivalent information. Thirty-five items were studied from the adaptive domain, 26 items from the communication subdomain, 34 items from the motor domain, and 31 from the personal-social domain. When the methods were unconstrained, the overall test indicated that at least two of the methods were not equivalent. However, FACETS bias analyses with the method measures constrained to equality allowed the identification of a limited number of items and associated methods that were possibly problematic. The use of FACETS allowed test developers to focus on items from a field test event that were inconsistent with the targeted test development model.

Rasch Measurement Using Dichotomous Scoring

Randall E. Schumacker
University of North Texas

The Rasch measurement model using dichotomous scoring of item response data from a newly created Mobility Scale administered to elderly independent living individuals is presented. The dichotomous scoring model, item calibration, person calibration, logit scale, normative scale score, reliability, and validity are explained. Results indicated that additional activity statements need to be written and tested to improve the Mobility Scale instrument.

Volume 5, Number 4

Measuring Higher Education Outcomes with a Multidimensional Rasch Model

Christine E. DeMars
James Madison University

A multidimensional Rasch model was applied to two instruments measuring abilities in two related areas of a university general education curriculum. Grades from related courses were also calibrated using the Rasch model. Thus, course grades, test items, and persons were all placed on the same metric. Incorporating grades within the metric provided additional meaning to the measures; instructors could see which items were matched to students in a particular grade range for a course. This could help not only in interpreting items but also in interpreting grades. Test items and grades fit the model reasonably well, with adequate person separation reliability.

Measurement in Clinical vs. Biological Medicine: The Rasch Model as a Bridge on a Widening Gap

Luigi Tesio
Istituto Auxologico Italiano

In the dominant Bio-medical paradigm Medicine is mostly Biology applied to Man. Measurement in Biology stems from physical sciences and has established validity. This is not the case for whole-person variables such as behaviors and psychic conditions (disability, pain, knowledge). The very existence of these variables can only be inferred by observing samples of representative behaviors. The quantity of the inferred variable may only come from subjective and discrete counts (scores) of events (coming in a questionnaire). Contemporary statistics demonstrated that raw scores intrinsically lack fundamental properties for scientific measurement, whatever their algebraic manipulations. This adds to the stigmatization of Clinical Medicine as “soft science”, compared to Bio-medicine. In the ‘60s Georg Rasch inaugurated a new statistical approach allowing transformation of raw scores into objective linear measures comparable to physical measures. This may help the Biomedical paradigm to redirect resources from laboratory bench back to bedside.

Dimensionality and Construct Validity of an Instrument Designed to Measure the Metacognitive Orientation of Science Classroom Learning Environments

Gregory P. Thomas
The Hong Kong Institute of Education

The purpose of this study was to establish the factorial construct validity and dimensionality of the Metacognitive Orientation Learning Environment Scale—Science (MOLES-S) which was designed to measure the metacognitive orientation of science classroom learning environments. The metacognitive orientation of a science classroom learning environment is the extent to which psychosocial conditions that are known to enhance students' metacognition are evident within that classroom. The development of items comprising this scale was based on a theoretical understanding of metacognition, learning environments and the development of previous learning environments instruments. Four possible hypothesized structure models, each consistent with the literature, were reviewed and their merits were compared on the basis of empirical data drawn from two populations of 1026 and 1223 Hong Kong secondary school students using confirmatory factor analysis procedures. The scale was calibrated using the Rasch rating scale model using data from the 1223 student sample. The results suggest that there is strong evidence to support the factorial construct validity of the MOLES-S but that, on the basis of the Rasch analysis, there are still suggestions for further refinement and improvement of the MOLES-S.

A New Class of Parametric IRT Models for Dichotomous Item Scores

David J. Hessen
University of Amsterdam

A new class of parametric IRT models for dichotomously scored items is presented. The new class of models is a subclass of both the class of models defined by the four-parameter logistic item response function and the nonparametric Double Monotonicity (DM) model. Three special cases of this new class of models are discussed. One of these special cases is shown to be the one-parameter logistic Rasch model. Both specific objectivity at the interval level of measurement and the sufficiency of the total score for the latent trait are shown to be measurement properties of the whole new class of models. For maximum likelihood estimation of the model parameters, both a joint and a conditional likelihood function are proposed.

Comparing Factor Analysis and the Rasch Model for Ordered Response Categories: An Investigation of the Scale of Gambling Choices

Andrew Kyngdon
University of New South Wales

Using both factor analysis (Spearman, 1904) and the Rasch model for ordered response categories (Andrich, 1978), the present study investigated the structure of the Scale of Gambling Choices (SGC, Baron, Dickerson and Blaszczynski, 1995). The scale was administered to a participant sample (n = 210) consisting of 57 first year psychology students, 104 in situ club Electronic Gaming Machine (EGM) players and 49 self-referred problem gamblers. It was hypothesized that the results yielded by factor analysis and Andrich's model would not agree with respect to the behavior of individual items. This hypothesis was supported; supporting the results of previous research (Johnson, et al., 1995; Raju, et al., 2002; Reise, et al., 1993). It was also hypothesized that a relationship would exist between item factor loadings and item expected value curve slope coefficients. This hypothesis was not supported and so hence did not support the findings of Parsons and Hulin (1982) and Roskam (1985). It was concluded that this was perhaps due to the different latent variable conceptions which exist between the Rasch models and factor analysis (Bollen, 2002). The limitations of the research were outlined and suggestions for future research were made.

Assessing the Assumption of Symmetric Proximity Measures in the Context of Multidimensional Scaling

Ken Kelley
University of Notre Dame

Applications of multidimensional scaling often make the assumption of symmetry for the population matrix of proximity measures. Although the likelihood of such an assumption holding true varies from one area of research to another, formal assessment of such an assumption has received little attention. The present article develops a nonparametric procedure that can be used in a confirmatory fashion or in an exploratory fashion in order to probabilistically assess the assumption of population symmetry for proximity measures in a multidimensional scaling context. The proposed procedure makes use of the bootstrap technique and alleviates the assumptions of parametric statistical procedures. Computer code for R and S-Plus is included in an appendix in order to carry out the proposed procedures.

Detecting Item Bias with the Rasch Model

Richard M. Smith
Data Recognition Corporation

The purpose of this article is to introduce the concept of item bias, highlighting the differences between the definition of the term as it is used within Rasch measurement and the definition of the term as it is used in the true-score model, non-model based approaches, or multi-item parameter latent trait models. The discussion continues with a description of alternative methods of assessing item bias within the Rasch measurement framework and discusses the power of these methods to detect the presence of item bias. The discussion concludes with several examples drawn from a number of different mathematics tests. This includes a comparison of the Rasch separate calibration t-test and the Mantel-Haenszel approaches.
