INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Université d'Ottawa · University of Ottawa

A Comparison of the Sample Invariance of Item Statistics from the Classical Test Model, Item Response Model, and Structural Equation Model: A Case Study of Real Response Data

Krista J. Breithaupt

Faculty of Education

Thesis submitted to the School of Graduate Studies and Research

in partial fulfillment of the requirements for the PhD degree in Education

University of Ottawa

October, 2000

Acknowledgement

This thesis would not have been accomplished without the guidance and timely intervention of several individuals. Sincere gratitude is due to: Bruno Zumbo who is always reaching and makes time to help others up along his way; Ian McDowell who inspired and supported this effort from inception to fulfillment; John R. Boulet who has been mentor and sherpa; Zoe I. Breithaupt, who measured the years of this project with her love, humour, and patience; Frances M. Breithaupt, who set the standard and lived 93 years to see her son, and granddaughter, meet it; and to Robert W. Breithaupt and Margaret R. Breithaupt, for expecting nothing less. Consolation may be offered for all their kindnesses; "One thing, at least, is certain: everything, absolutely everything, that I shall say here is entirely and exclusively my own fault" (Dali, 1942, p.6).

Abstract

The sample dependency of item statistics and the practical importance of alternative models for test scores are evaluated in this case study using real data. The hypothesized superiority of the item response model (IRM) attributed to the sample-invariance of item statistics is tested against a classical test theory (CTT) model and a structural equation model (SEM) for responses to the Center for Epidemiologic Studies-Depression (CES-D) scale. Sample-invariance of item statistics is tested in 10 random samples of 500 people, and across gender, age, and different health groups. Practical implications are considered in a comparison of differential score reliability, and individual rankings, using each test model. Item estimates from a 2-parameter logistic IRM were compared with classical item difficulty and discrimination estimates, and with item regression path estimates from a uni-factor SEM. An intraclass correlation coefficient (ICC) was calculated to evaluate the level of absolute agreement between item statistics in each condition.

Item statistics from all test models were very similar across random samples, indicating a high level of invariance. However, IRM threshold parameters were least sensitive to sampling compared with other models. Greater variance was found among item statistics based on all test models across groups that differed in age, health, or gender. IRM discrimination estimates were most stable across contrasting groups, compared with those from other test models. Rankings assigned to individuals were most similar when CTT scores and linear transformed IRM scores were compared. The largest

variation in individual rankings was obtained when SEM factor scores were compared with CTT scores in the higher score ranges. The reliability estimate for factor scores based on the SEM was highest overall. However, IRM optimal scores and the modified reliability estimate based on these, provide a more accurate estimate of average measurement error.

This evidence supports the hypothesis of improved score precision when tests are constructed and scored using IRM techniques. However, rankings based on individual CES-D scores were very similar when the CTT, IRM, and SEM techniques were compared. Therefore, CTT or SEM scoring are reasonable alternatives to the IRM when norm-referenced score interpretations are based on CES-D scores.

Table of Contents

List of Appendices

List of Tables

List of Figures

Chapter 1: Introduction

Theory and methods in educational measurement have evolved over the past fifty years to meet the increasing demand for defensible and valid educational tests. Interpretations made from individual scores must be justifiable based on a review of all steps taken in the test development and scoring process. These procedures are often subject to rigorous examination using the most appropriate psychometric methods available. Theoretical and technological developments have been most apparent in standardized testing situations, where individual and political stakes are high (Linn, 1993). One important accomplishment has been the development of item response models (IRM). The primary reason IRM techniques have received wide acceptance in education is the value of a central model feature, namely the theoretical invariance of test and item statistics and individual scores (Engelhard, 1992; Rudner, 1983). The IRM is compared in this case study with a structural equation model (SEM) and a classical test theory model (CTT) via an examination of item parameter invariance and scoring properties.

Mathematical models are developed in the social sciences to describe or predict human characteristics that cannot be observed directly. The utility of such models depends partly on their accuracy and appropriateness across different situations. The IRM is one example of a mathematical model that provides accurate information about test items and individual abilities in a variety of applications, when used appropriately. The IRM belongs to the class of non-linear regression models and, theoretically, will provide invariant item estimates (also termed sample-invariant test statistics). This means properties of tests and items derived from the IRM (e.g., item and test statistics) are not

theoretically sensitive to examinee characteristics unrelated to ability (such as gender, or average group performance). In addition, IRM individual scores are not dependent on the items selected for inclusion in the test (test-free individual scores). A correctly specified IRM will result in invariant item and test statistics when there is good fit of the model to the data (Hambleton, Swaminathan, & Rogers, 1991).

The practical advantages of invariance include the possibility to generate optimal individual scores, to tailor tests, and to examine test validity via a wider range of item and score statistics. IRM scoring reduces score bias related to group composition and allows comparison of individuals across different tests. Individual scores are based on predictions from IRM item parameter estimates and the pattern of responses given to test items. The calculation is similar to a multiple linear regression where one predicts an outcome variable (ability) using a set of model parameters (e.g., item properties) and the responses to independent variables (the item responses). When IRM estimates of item statistics are used in test development, item selection is unbiased by the composition of the pilot sample who provide data for calibration. Due to the invariance property of the IRM, item difficulty and discrimination estimates based on a separate subpopulation will be equivalent up to a linear transformation of scale (Rudner, 1983). In addition, it is possible to assign more precise scores to individuals (with smaller errors of measurement). IRM analyses allow us to test empirically some aspects of score validity that cannot be made explicit using a classical test theory (CTT) model (Hambleton, 1984).

Another popular method of examining the relationship between responses to items and the latent trait represented by a measure is structural equation modeling (SEM). This

technique makes use of confirmatory factor analysis (CFA) for modeling measurement properties and allows method effects related to item wording to be specified. In addition, CFA belongs to the larger class of SEM (Stevens, 1996) and provides a statistical significance test of possible differences in response data obtained from contrasting groups of respondents (e.g., Benson, 1987; Byrne, 1994).

Researchers suggested the use of SEM for educational measurement early in the literature; however, it has not been fully exploited. Muthen (1985) demonstrated a form of SEM useful for the study of item homogeneity and possible score bias. In that study of educational test responses, Muthen incorporated into the SEM relationships among items (possible method effects) that are not consistent with a basic assumption of IRM (namely, the conditional independence of item responses). Muthen concluded that the SEM "...makes it possible to investigate invariance hypotheses regarding both measurement and structural parameters, which would be valuable, for example, in studies of test item bias" (Muthen, 1985, p.132).

Method effects may occur over and above the shared construct measured by the test, when pairs or small groups of items are similar in content or wording. The response patterns contain correlations among these small sets of items. IRM- or CTT -based scores fail to account for this source of covariance and may have larger measurement errors. For example, a positively worded item set may elicit a different pattern of responses compared to a negatively worded item set. Researchers have suggested that we must account for these effects when examining test and item psychometric properties and when calculating individual scores (Andrews, 1984; Carmines & Zeller, 1979; Tomas & Oliver, 1999). Empirical evidence for the properties of items, including score invariance based on

SEM, is often used to establish the validity of score interpretations. It is also possible to obtain optimal scores (e.g., factor scores) based on appropriate SEM of response data.

The purpose of this study is to compare the invariance of item parameters and total scores across samples drawn from a large set of real data using three mathematical models for test and item scores (CTT, IRM, and SEM). Theorists in modern educational measurement have stated that item and person estimates from an appropriate IRM will exhibit invariance, whereas CTT item and person estimates will be less similar across random and contrast group samples. In addition, it is reasonable to expect that items on a short scale that assess a broad abstract domain will not have equal precision of measurement. Therefore, both IRM and SEM measurement models should result in less varied estimates for items and more accurate individual scores when compared with item estimates and scores based on a simple linear CTT model.

The invariance of item statistics will be compared in this study using empirical evidence from a case of real data. Item parameter estimates based on each scoring model will be cross-validated in random replication samples. In addition, the models will be tested across groups where possible group-related item and total score bias is an important issue (gender, age and health groups). Finally, the practical impact of each measurement model will be assessed by identifying differences in score ranks, distributions of obtained scores, and estimates of average score reliability.

This thesis contains a literature review that frames the research questions, and introduces the case study and test models (Chapter 2). There are five sections in the literature review: a description of related measurement theory; a rationale and introduction for the case study, a review of the test models to be applied in the study; a

description of item statistics from these test models and research findings related to sample invariance of these statistics, and a description of the Center for Epidemiologic Studies Depression (CES-D) scale examined in this study. The chapter ends with a summary of gaps identified in the literature and a presentation of three research questions.

Chapter 3 includes seven sections that describe the study methods. These sections begin with a description of the source of data, the levels of data analysis, and the analysis steps. Then, the estimation methods applied for each test model are described, the method used to summarize invariance of item statistics, and procedures for obtaining individual scores are described. Specifically, a short form of the CES-D is proposed and the fit of IRM and SEM are described for this test form. Also, the sample is truncated to represent more closely the response characteristics that would be obtained in an optimal performance test.

The results of the case study are reported in Chapter 4, in a sequence that reflects the order of steps taken in the analysis; a description of the sample, an examination of test model assumptions for the data, model selection and estimation of item statistics, evidence of model fit, and findings related to invariance of item statistics. The distribution of scores from each test model are presented next, followed by the closeness of ranks for individuals (compared across test models), and a summary of reliability estimates for individual scores. The final section of this chapter contains the results of three post-hoc analyses that may aid generalization of these results to educational tests.

Chapter 5 is a summary and discussion of the study results. In view of the length of the thesis, a review of study results describing the invariance of item statistics is provided for ease of comprehension in the discussion section. This review revisits each

research question, and provides a response based on the study results. The discussion section follows. This is broken into topics: the invariance of item statistics; the importance of test models for individual scores; and a section intended to address some limitations of this case study in terms of internal and external validity. Next, implications for research and practice are considered, including decisions taken based on individual scores, method effects related to item wording, and the dimensionality of the CES-D. This final section focuses on clinical interpretations important for researchers with a primary interest in the measurement instrument examined. The thesis concludes with comments on the contribution of this study to educational measurement.

It is important to note that this study does not provide a direct test of bias in test scores or item statistics, or of differential item functioning (DIF). It would be necessary to obtain a criterion measure for the true latent variable for the CES-D (e.g., medical diagnosis of depression) to determine if true bias in item estimates or individual scores were present. A separate literature describes methods appropriate for the identification of test and item bias; this goes beyond the purpose and limitations of this case study.

A related issue concerns the use of the term invariance in this study. The term invariance has not been reserved to conditions where average group ability is expected to have an impact on item parameter estimates. In practice, an examination of the invariance of item statistics may be conducted to ensure that item statistics from samples assumed to be equivalent is frequently used to study pools of test items for computerized testing (e.g., in studies of item parameter drift). The groups are not expected to differ, and a lack of invariance in item parameter estimates may have a variety of explanations (e.g., changes in curriculum, or the presence of some items that cue the correct answer to other items on

a test form). A broader interpretation of invariance useful for this study may include sampling conditions where a lack of invariance is related to poor model fit, or to sampling error. In this case study, conditions representing sampling error (random samples), and possible item bias (contrast samples) are both examined under the broader interpretation of invariance expected of IRM. This thesis was constructed with a view to testing theory-based sample invariance properties of item statistics in a case of applied data.

Chapter 2: Review of Essential Literature

The literature review chapter begins with a brief introduction to three foundational topics that guide and situate the research questions to be addressed in this study. These topics consist of an overview of validity theory from educational measurement, a discussion of measurement techniques applied to educational and health measures, and an introduction to the three models for test scores that will be applied in this study. The next section of this chapter provides an introduction to the case study that forms the application for these research questions. This section also presents a rationale for the selection of this case of real data and a description of the measure. The similarities between the CES-D and short educational screening tools are then considered.

The final sections of the chapter describe in detail the IRM and SEM test models that will be applied in the study. A description of an appropriate interpretation of the IRM for CES-D responses is provided also, as the language of IRM is typically for educational measurement. In the next section, item statistics for the IRM and SEM are defined, and the relationship between these formulae is considered. Some results from published studies that examined the sample-invariance of item statistics from CTT, IRM or SEM are reviewed next.

Previous research that describes the development and properties of the CES-D is then described. This section includes a description of some alternative scoring methods and empirical evidence for measurement properties and scale structure. This chapter closes with a statement of the three research questions for the study.

Foundations

Measurement validity has been described in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) as a unitary idea, termed construct validity. This umbrella concept includes many forms of evidence given for test score validity. Validity issues in educational measurement have been discussed in detail by others (Cronbach, 1988; Messick, 1993; Moss, 1995). Historically, forms of validity evidence were considered as distinct. Examples of evidence for score validity include predictive or criterion validity, contextual validity, face validity, and discriminant or convergent construct validity. Recently, the consequences of appropriate test use have also been included as an important aspect of score validity (Messick, 1989). Any of these sources of evidence, or a combination of them, may be examined to determine whether scores from a measure lead to valid inferences. This evidence is usually appraised with respect to a given case of applied measurement.

Theorists have emphasized that inferences based on scores from the measure are validated, not the test itself (Messick, 1993; Moss 1995). Inferences will depend on the purpose of the measure in each application. The inferences are simply interpretations regarding the construct the measure represents. Scores are derived from test items, and their properties are the subject of measurement validity studies. The IRM has been used for the selection of items, the design of tailored tests, computer adaptive testing, investigations of item and test bias, test form equating, and in determining standards or cutting scores (Hambleton & Jones, 1993). Some studies in educational assessment have

shown that IRM techniques provide improved validity of score interpretations due to the property of invariance (Hambleton et al., 1991; Hulin, Drasgow, & Parsons, 1983). Linn (1990) made the important observation that the validity of any IRM for test construction and scoring must be demonstrated empirically.

Analytic techniques based on an IRM hold promise for developing and evaluating scales used to measure individual and public health (Gumpel & Wilson, 1996; McDowell & Newell, 1996; McHorney, Haley, & Ware; 1997; McHorney, 1997; Streiner & Norman, 1995). Studies of population health are based directly or indirectly on scores from screening and evaluative health measures. These scores are used to determine program effectiveness, develop clinical interventions, allocate resources, and set policy in health care. Decisions about individual clients may also be made using scores obtained from screening scales. This kind of health screening in a normal population is analogous to affect or achievement screening tests used in education (some examples of these are provided in the Case Study section of this chapter). Interpretations made from health scale scores have economic and personal impact comparable in many respects to the use of test scores in education.

IRM methods have been used in some health measurement studies (e.g., Kirisci, Moss, & Tarter, 1996; Teresi, Golden, Cross, Gurland, Klienman, & Wilder, 1995). However, a review of published studies using IRM methods with health scale data reveals only a slowly increasing trend. The scarcity of such examples, when compared with assessment in education, suggests many health educators and researchers do not know of the possible advantages of IRM techniques for scale development and scoring, or their potential appropriateness for health data. This may be partly due to challenges inherent in

sharing methodology across disciplines. Education and health each have a specialized vocabulary and unique scientific history.

IRM analyses in the validation of health scales have been limited, incomplete, and sometimes incorrect. One form of the IRM dominates (the Rasch model), and has been used primarily to examine item validity. Important issues such as IRM fit and appropriateness have been largely ignored. Many useful aspects of IRM analyses have not been explored by health educators and researchers; these include the examination of sample invariance of item statistics and the possible benefit of optimal scoring.

The choice of an analytic model for test scores may have an impact on individual score interpretations. CTT, SEM, and IRM allow the estimation of item statistics or parameters, and each model also provides a scoring method for individuals. These three models can be described as simple linear, weighted or linear FA, and non-linear FA, respectively. Only one example exists in the current literature where the CES-D has been examined using IRM analysis methods. That study involved a non-parametric IRM analysis (Santor, Zurloff, Ramsay, Cervantes, & Palacios, 1995). The IRM computer program applied by Santor et al. (1995) provided a graphic display of item and test properties, and was not useful for identifying complex structure. The authors do not provide item parameter estimates for the CES-D and interpret the IRM only peripherally. Many implications related to the application of the IRM, such as sample-invariance of item statistics, and potential benefits related to optimal scoring, are not explored in the health measurement literature.

Any improvement in the validity of individual scores derived from a superior scoring model would be associated with more precise total scores for individuals.

Potential gains in precision can be examined via techniques that estimate the reliability of individual scores (Crocker & Algina, 1986). A scoring system that allows for different item discrimination values will weight responses with smaller errors of measurement more heavily in the calculation of total scores. When an IRM analysis is used to score individuals, the item discrimination values provide this differential weighting (Hambleton et al., 1991). Item regressions on the latent variable determine the impact of item responses on total factor scores when SEM is applied (Byrne, 1997). Simple linear CTT scoring is based on a simple sum of item responses, where items are equally weighted and presumed to have equal errors of measurement. Therefore, when items differ in their precision of measurement (discrimination), a weighted scoring method will lead to more precise total scores and higher reliability estimates.

For example, when an internal consistency estimate is used to represent the reliability of individual scores (e.g., Cronbach's alpha), weighted scoring will lead to higher estimates than simple linear scoring given particular conditions of the data (Carmines & Zeller, 1979). This is the case when items have unequal errors of measurement or when there is complex dimensionality in the latent construct (e.g., more than one underlying trait).

An internal consistency reliability estimate based on factor scores has been described as a "maximized" alpha (Carmines & Zeller, 1979, p. 61). Whereas the traditional alpha applied to CTT item scores is a lower bound estimate of reliability, the reliability estimate from weighted item responses is the "....*closest estimate* to the true reliability of the measure" (Carmines & Zeller, 1979, p. 62). A comparison of weighted and unweighted reliability estimates is one test of the usefulness or impact of the more

complex scoring model. Any practical gain of the IRM optimal scoring system or the SEM factor scoring system over the simple linear CTT will be evident in higher reliability estimates for individual scores. In this way, improved precision derived from a complex model for test scores may be tested empirically in a case of real data.

Empirical evidence for the usefulness of each measurement model with real response data has a range of important implications. Tests of item parameter invariance for each measurement model (via multiple replications) are relevant for current practice in test development and scoring. In this study, the empirical evidence for possible superiority of the IRM due to optimal scoring and item parameter invariance is examined in circumstances of real response data. This information may be used by educators who must select test models to develop and score short assessment tests in an applied setting. The practical impact of a complex model (IRM) is examined with direct comparison to the simpler classical test model (CTT). In addition, the potential benefit of optimal scoring that incorporates method effects in item wording (via SEM) is presented for comparison.

Educators and researchers in the health field will be interested in the secondary clinical implications of the thesis. High stakes decisions are taken based on the results of large-scale assessment in health care. Modern measurement theory has not been widely applied in that setting. Recommendations are provided to guide precise scoring of the CES-D, and evidence will be presented for score comparability across sub-populations.

Post-hoc analyses will include tests of hypotheses related to conditions of the data that may be specific to this case study, and one alternative method used to examine the sample independence of IRM item statistics in groups that differ on the latent variable.

Specifically, the first post-hoc is a test of the fit the IRM and SEM to a normalized distribution of individual scores by eliminating responses from people who report no symptoms, the second post-hoc test will assess the fit of IRM and SEM to a shortened CES-D where poorly discriminating items are eliminate; the third post-hoc test will present an estimate of the sample independence of IRM item statistics when contrast groups are represented with ability scores on a shared metric. The case of real data selected for this study is described in the following section.

The Case Study

The idiosyncrasies of real response data are important and provide necessary evidence for the validity of any model used for individual scoring. Responses to the Center for Epidemiologic Studies Depression (CES-D) scale are selected for this case study. This was, in part, a pragmatic choice; a large sample of data was available. However, depression and its measurement are also relevant in the educational field. Depression is one mediating factor between instruction and learning. Evaluation theorists in education link mental and physical capacity with achievement. Affect is a primary element in many models for educational evaluation, including those presented by Tyler (1950) and by Hammond (1973). Goodlad (1979) has listed 12 goals for American schools, one of these was "emotional and physical well-being." Emotional health has also been described as a primary purpose of education; Worthen and Sanders (1987, p. 67) have operationally defined it in terms of the "...attitudes, feelings and emotions" of the student. The emotional experience of the learner is an important consideration. It is a pre-requisite and an outcome of effective instruction.

The CES-D is used to measure depression in adults and provides an indicator of the motivation and susceptibility of adults for interventions that require the acquisition of coping skills for emotional and cognitive adjustment. In this sense, the counseling and theraputic approaches used to treat depression based on interactive guidance strategies are similar to student-centered teaching models. The capacity for learning is central to the educational process. Clearly, emotional well-being is an important element and this is made explicit by research in the discipline of educational psychology. Humanistic models of education explore the impact of emotional health on learning; Hamachek (1990) summarizes five popular theories from educational psychology:

> All in all, research and clinical evidence strongly suggest that psychological health and a sense of competency are goals within reach of every school-age youth and every adult. Perfection is not our goal. Improvement is, and this is a goal toward which each of us can aim ourselves and the students we teach (Hamachek, 1990, p. 69).

Clearly, affective measures have an important role in measuring student readiness for learning and this information may be used to help teachers work towards their educational goals with students.

The CES-D is intended for use with an adult population and may be useful in planning (or evaluating the impact of) community-based educational programming for this group. Lifelong learning has also been an important development in models of education (Knowles, 1977). Adult education is one avenue for improving the quality of life of our elderly population. Issues related to providing educational and health programming for the elderly have been a major focus in the last decade in Canada and

internationally. This is evident in the number of funded community and national studies of aging populations, the assemblage of advisory committees for seniors' issues, and the decision of the United Nations that 1999 be declared the International Year of Older Persons.

In addition, the results of this case study related to the invariance of item statistics from different test models will have implications for some achievement measures used in education. Study findings related to sample invariance of item statistics based on responses to the 20 item CES-D may be generalized to specific testing situations in education. Shorter self-report measures, scholastic aptitude tests, and affect measures used as diagnostic or screening tools in evaluation have data properties very similar to the CES-D. One example is a measure of test-taking anxiety, described by Ferrando (1999). The Questionnaire of Anxiety and Performance (QAP) evaluates adult students for anxiety related to educational test taking and fear of consequences from performance. Students respond to items presented in a format equivalent to the response scale used with the CES-D, and the QAP includes 18 items. Both the CES-D and the QAP contain items that represent subtle, and possibly diverse aspects of the broader domain of an underlying trait, and the QAP has been examined using an IRM in several validation studies (Ferrando, 1999).

A second example for comparison with this case study may be drawn from Benson (1987) who examined item bias in affective scales. The measure was designed by Educational Testing Service for a national integration study and consists of 32 items assessing self-concept and racial attitudes. In that study a SEM was applied to determine the similarity of response data properties across contrast groups of eighth grade students.

The group sizes selected for analysis ranged from 1264 to 377. Those results were interpreted as evidence for the appropriateness of the SEM methodology for studies of measurement invariance in sub-populations. In this example, the affect measure and the selected analysis are similar to this case study, and implications for educational measurement are made explicit. It seems reasonable that a study of invariance techniques and alternative test models based on CES-D responses would be equally relevant.

If this study demonstrates that the response data from the CES-D meet assumptions required for each model of test scores, it is reasonable to expect the study results to be relevant for other applied situations such as the affective test designed by ETS (Benson, 1987); or the QAP (Ferrando, 1999), or other similar short binary response screening tests. There is no theoretical limitation on the nature of the latent variable modeled using these techniques; therefore, these results may be generalizable to other tests where the response data have similar structure.

The CES-D includes 20 items that reflect affective, attitudinal, and somatic elements symptomatic of depression. The scale contains four positively worded items (describing positive feelings), while the remaining items have negative phrasing. Researchers have not yet examined the possible impact of item wording on the structure of responses to the CES-D. It seems reasonable that these potential method effects would lead to communalities in the response data similar to case-based questions used in proficiency testing. These linked items are often recognized as testlets and scoring is adjusted to accommodate shared response variance. These features of the CES-D (items with expected associations irrelevant to the latent variable, and complex structure), and the importance of applications of this measure, were important considerations in selecting

this case study. The CES-D has become one of the most popular screening measures used in studies of population and community health (Fechner-Bates, Coyne, & Schwenk, 1994) and has not been examined using IRM or SEM scoring methods.

A large sample of responses to the CES-D was available, and this provided an opportunity to examine sample invariance properties across successive random samples and across selected comparison groups. The response data are contained in a large survey database that includes a total of 6974 responses to the CES-D. A cross-validation of the fit of the IRM and SEM is also possible in a sample of this size, as are statistical tests of model fit and structural invariance that make use of large sample theory. The case study provides a unique and timely contribution to our knowledge by extending educational measurement methods to an allied discipline where potential benefits are substantial.

Modern Measurement Models

A variety of authors have presented comparisons of alternative measurement models, and described the mathematical relationships among them (e.g., Hambleton & Jones, 1991; McDonald, 1982). Brennan (1998) has placed the development of new test models (e.g., IRM or SEM) within the classical test theory framework. In his presidential address to the annual meeting of the NCME, Brennan described complex measurement models as specialized and fallible derivations from traditional conceptions of test scores:

> With the development of new models, there is a tendency to think of them as a replacement for classical theory....However, classical theory is alive and well, and it will continue to survive, I think, for both conceptual and practical reasons. Classical test theory is an incredibly simple, but not simplistic, model. It postulates that an observed score can be split into two latent parts (a true score and an error score). This is a tautology, and, as such, it is not capable of being

> proved or disproved. Yet, this simple notion is fundamental to all of measurement. I believe, therefore, that it is misleading to view new models as replacements for classical theory-extensions or liberalizations, yes, but not replacements. (Brennan, 1998, p.6).

Others have also described the simple linear CTT model as a special case of the general IRM (Goldstein & Wood, 1989). In the simple linear (true score) model certain assumptions are made. These assumptions require only that observed scores are composed of a true score representing examinee ability and some amount of measurement error, and that the error estimate is uncorrelated with the true ability score (Nunnally, 1978). This is often expressed as X=T+E, as given in texts such as Crocker and Algina (1986). In the simple form of this model, correct items are simply added together for a total score. Almost fifty years ago, Lord (1953) made the critical observation that the true ability score must be independent of the particular items on a test. Thus, began the search for other useful models of test scores that would differentiate the peculiarities of test items from the true ability of examinees.

An alternative model for test scores was later described using linear factor analysis (McDonald, 1985b). This IRM allows items to have unequal errors of measurement and provides a weighted total score. In that same chapter, McDonald led the reader through the development of many forms of the modern IRM based on non-linear or harmonic FA. The potential benefits of the IRM are only achieved when the response data are appropriate and fit the IRM well. Many approaches to testing IRM suitability or fit have been developed (McDonald & Mok, 1995). This requirement for good fit to the data

also holds for the simple linear CTT model, although the assumptions are usually tested indirectly.

Item Response Model

The initial development of item response models has been attributed to Lawley (1943), who proposed that responses to items on a given test represent an underlying trait (or ability). This assertion was formalized in early item response theory (IRT) via the specification of a variety of mathematical models for the prediction of responses based on the properties of test items and on the underlying trait or ability of the person (e.g., Bock & Aitkin, 1981; Lord, 1953). Methods of estimation have been made more efficient through application of maximum likelihood estimation and the use of the normal ogive curve to describe the model (Bock & Lieberman, 1970).

Two necessary assumptions for the simple form of the IRM are unidimensionality of the latent trait and local independence of items (Hambleton et al., 1991). Model appropriateness depends on how well the response data meet these assumptions and on substantive and empirical evidence for specified parameters in the IRM. Analysis of the fit of successively more complex IRM to a data set provides information about the properties of items and the validity of score interpretations based on each model. A review of essential concepts for the IRM will introduce some characteristics of response data that are appropriate for these methods.

In the simple unidimensional case, a set of items measure a common underlying factor and the IRM is used to describe responses to items as a monotone function of the latent trait. The probability of a person endorsing an item is predicted from this model with parameters to represent any or all of: ability (θ), threshold (b), discrimination (a),

and pseudo-guessing or chance endorsement (c) for items. IRM parameter estimation is generally based on a form of non-linear factor analysis, involving marginal maximum likelihood estimation. Various other methods of estimating item parameters have also been studied, including Bayesian and iterative least squares methods and non-parametric methods (e.g., van der Linden & Hambleton, 1997; Thissen & Steinberg, 1984).

The function relating item parameters to the latent trait may be plotted as a normal ogive curve representing the likelihood of endorsement of the item as a function of the latent trait. In a simple example, the response format for the item is assumed to be binary. A general mathematical form of the two-parameter logistic IRM is attributed to Birnbaum by Lord and Novick (1968). The general IRM formula for a unidimensional 3 parameter logistic model is shown below:

$$P(i=1|\underline{\theta}) = c_i + (1-c_i)\frac{e^{1.7(\underline{a'\theta}-d_i)}}{1+e^{1.7(\underline{a'\theta}-d_i)}},$$

where θ is the vector of abilities on the latent trait;

a' is the vector of discriminations of item i on the latent trait;

c_i is the lower asymptote for item i; and,

d_i is a scalar related to the difficulty of item i.

In IRT the nature of the underlying trait that is measured is not specified; rather, the focus is on mathematical models for item and test properties useful in test development and scoring. The distinction between theories defining latent traits, and IRM

used to predict responses to items is argued by Hulin et al. (1983). The term IRM seems most appropriate when these methods are used to examine responses reflecting a broad range of constructs. This convention is also followed by Goldstein and Wood (1989) and serves to de-emphasize conceptual definition of traits.

IRM analysis has been adopted in other disciplines, and has a long history in cognitive and psychological measurement. Reise and Waller (1990) presented a study of IRM applied to a widely used personality measure (the Multidimensional Personality Questionnaire, MPQ). In their study, the question of IRM appropriateness when assessing typical performance was contrasted explicitly with the context of IRM for maximal performance assessment in educational testing.

Whereas educational tests elicit a high proficiency sample of behavior to represent ability, personality and health measures are intended to elicit a sample of responses that characterize normal functioning for an individual at the time of testing. The MPQ consists of 11 sub-domains that each consists of a small set of items (ranging from 20 to 34 items each). These small item sets were analyzed as separate tests by application of a two-parameter IRM. The authors found little evidence for mis-fit of the IRM; in fact 95% of the MPQ items had good fit to the model based on a modified χ^2 test of residuals. The authors concluded, "...researchers engaged in the assessment of normal-range personality processes have much to gain from exploiting item response models" (Reise & Waller, 1990). Their work served to underline the suitability of IRM analyses for a variety of applications where typical performance is measured (e.g., health screening tests).

Comprehensive reviews of the development and use of IRM have been produced by several authors (Crocker & Algina, 1986; Hambleton et al., 1991; Hulin et al., 1983;

van der Linden & Hambleton, 1997). A class of models has been developed that have increasingly broad application, extending from items requiring binary responses, to items with polytomous graded or nominal response formats (Hambleton & Jones, 1993; McDonald, 1982; Wainer, 1989). Although the IRM was originally developed for educational testing, there is no theoretical limitation on the underlying trait or construct. However, even when statistical evidence indicates that a more complex measurement model fits the response data well, the usefulness of any scoring model must be examined in a validation study with applied data (Linn, 1990).

The usual terms defining item and person estimates in the IRM have not been standardized for health scales. One interpretation of IRM parameters for mental health has been described by Zumbo, Pope, Watson, & Hubley (1997), and that convention is used here. In this case study, responses to a measure of depression are to be examined. Therefore, the latent variable, θ, will represent the severity of the underlying condition (here, self-reported depression). The item discrimination parameter, *a*, will refer to the accuracy of an item in classifying individuals with high and low overall depression. The *b* parameter for item difficulty will correspond to the threshold on the underlying condition where the likelihood of endorsing an item rises above chance levels.

Within the framework of the IRM, endorsement of an item in the CES-D occurs when the depressive state of the respondent is at or above the threshold measured by the item. A higher response option (one, for a binary item scored zero or one) indicates the presence of the symptom. It is correct to describe depression items via the item threshold, and this term is used in several item analysis programs (Fraser & McDonald, 1988; Mislevy & Bock, 1990).

The *c* parameter usually represents the likelihood of endorsement of an item just by chance, and forms the lower asymptote or intercept for the item characteristic curve (ICC). When responses are given to a health scale the *c* parameter may represent an endorsement of an item when the underlying severity of depression is low. For example, a symptom may be present due to a co-existing condition (e.g., some depression items may measure somatic symptoms). Therefore, IRM parameters for health scales are described by the underlying individual health state (θ), the severity measured most accurately by an item (*b*), item discrimination (*a*), and the intercept for responses to an item (*c*). The IRM allows for a larger number of item statistics than either CTT or SEM. The following section provides a description of the common form of SEM that is applied to item response data.

Structural Equation Model

Two benefits of the SEM analysis for measurement are the availability of statistical tests of hypothesized structure across groups and the potential to account for method effects in item wording. This section provides an introduction to the estimation methods related to SEM, prior to an introduction to the topic of method effects related to item wording.

In the general form of SEM, responses to a measurement instrument are analyzed using a linear form of confirmatory factor analysis (CFA); this procedure belongs within the broader class of available SEM techniques. SEM has been applied to test hypotheses that specify the relationship between observed variables (responses to items) and the latent trait(s) defined by a measurement instrument (Bensen, 1987; Bollen, 1989; Byrne, 1997; Ferrando, 1996; Muthen, 1984; 1985; 1989). SEM techniques have been used in

studies of construct validity, where meaningful groups of items define the structure of responses to items. A broad range of useful estimates may be obtained using SEM, including item discrimination and total score reliability.

The CFA portion of a full SEM represents the structural relationships between items and the latent factor. When causal relationships are included in the SEM between two or more latent factors, in addition to the measurement model or CFA, a full SEM has been defined (Byrne, 1997). It is also possible to specify a non-linear SEM that is mathematically equivalent to the IRM (McDonald, 1985b; Muthen, 1984). However, the common linear form of SEM is applied in this study, and an introduction to the specification and estimation of the linear SEM is presented in this section. A brief review of the basic issues involved is offered here. Interested readers are referred to Byrne (1997) for a thorough treatment of this topic.

SEM (or CFA) is appropriate when there is a strong theoretical foundation whereby it is possible to identify the underlying factor structure. Exploratory factor analysis (EFA) is conducted in order to reduce the number of variables that describe response patterns in a data set. This is often contrasted with CFA methods. SEM has been termed a theory-testing technique, whereas EFA is theory-generating (Stevens, 1996).

A path model diagram may be created by the analyst using LISREL to represent the hypothesized SEM. The latent factors are linked to item responses by regression paths. Error terms are included, and these would have unique regression paths to variables representing each observed variable or factor. The model corresponds to a LISREL syntax file, where each element and relationship in the model is presented. Item regression paths, error terms, variances, and correlations in SEM are usually represented using the symbols

from the Greek alphabet. Each parameter corresponds to a particular set of estimates used in the mathematical analysis. These are grouped in the form of vectors or matrices. In a CFA model that represents responses to a measure, latent factors are usually identified using ξ (ksi), paths to items are identified as λ (lambda) and error variance terms for items are δ (delta). Correlation paths between latent factors are identified as Φ (phi), the correlation paths between error terms are termed $\Theta\delta$ (theta-delta). Observed values for item responses are usually represented as X.

Each estimated or fixed path (or variance term) is a parameter in the SEM, and belongs to a matrix that defines all possible parameters of that kind in the model. The rectangular Lambda matrix will have dimensions where the number of rows equals the number of items in the test, and has a column for each latent factor postulated. The delta matrix will be a vector with a column entry for each item to represent unique error variances. The theta-delta matrix is square, equal in row and column number to the number of items. Off-diagonal elements in the theta-delta matrix represent correlations between item error terms. The ksi matrix will be a vector containing the variances of the latent factors. Together, the full set of SEM matrices contains the total set of model parameters. Byrne (1997, p. 32) defines the general CFA model relating these matrices to observed variables as:

$$X = \Lambda_x \xi + \delta$$

where,

X represents the vector of x observed variables;

$\Lambda_x \xi$ represents all λ regression coefficients multiplied by their respective ξ factor; and

δ represents measurement error associated with each of the x observed variables.

Model specification with LISREL requires that the analyst define the type, form, and mode of each matrix. The type of matrix will refer to the Greek name for the elements. Whether the elements are full, symmetric, or diagonal is dependent on form. The matrix mode will be indicated by the analyst to denote whether elements are fixed, free, or constrained to any value. LISREL has default types and forms for each matrix, and some conventions are followed. For example, the Theta-delta matrix is fixed unless otherwise indicated. Elements in each matrix (or vector) in the model are identified via item and row subscripts and by the name of the matrix. Any member of a matrix where elements are defined as fixed (not estimated) may be freed in the subsequent syntax of the program. Model fitting is usually preformed in successive runs where starting values or re-parameterizations are incorporated to improve the fit of the model. This method is used to obtain a best baseline model, where substantive and statistical criteria are considered to arrive at the best fit to the response data.

In a test of structure across two groups, the syntax program is written to simultaneously fit a structure to the data representing each group. Each matrix in the model for the second group may be fixed as exactly equal to that of the first group, or of similar form, or be estimated independently. Usually, a stringent SEM (equal forms and elements for all matrices in both groups) is compared to successively less stringent models over successive runs for a two-group validation test. A general discussion of some issues related to mathematical estimation may clarify steps required for any SEM analysis.

A set of linear equations is used to represent the SEM; these specify the relationships between model parameters. These equations are soluble via matrix algebra (identifiable) when there are sufficient observations relative to the number of unknown parameters in the model. The SEM is identifiable when there is a unique solution possible for the model. Arbitrariness in the model is avoided when there are a greater number of data variances and covariances than there are parameters in the model. This desirable situation results in what is termed an over-identified model. A just-identified model (also, an under identified model) lacks the positive degrees of freedom to yield a unique solution for parameters in the model. However, a model with sufficient degrees of freedom must also have a specified scale for estimation.

The scale for each latent variable and regression estimate must be determined to identify the SEM. Latent variables have no intrinsic metric; therefore, the variances, regression paths, and error variances may only be estimated when a scale is supplied. This may be accomplished in one of two ways. Each latent variable in the SEM may be linked to one observed variable with a fixed regression path constrained to 1.0; usually a variable with high reliability is selected for this purpose. This variable is termed a "reference" variable in the model. Alternatively, the variance of the latent factor may itself be standardized based on the estimated population standard deviation. This method is preferred by some researchers, but it does not allow the variance of the latent factor to be estimated as a parameter in the model. It is not possible to estimate all the regressions of items on a latent factor, in addition to the variance related to that factor. Therefore, a decision must be made by the analyst to fix either one regression path for a reference variable, or to standardize the variance for each latent factor in the SEM.

The observed responses to items are analyzed to solve the set of algebraic equations and obtain values for all specified SEM parameters. The recommended data for input is the variance-covariance matrix based on observed variables (Cudek, 1989). If the data are on a non-interval scale (e.g., nominal or binary) the analysis will be based on an asymptotic matrix of item covariances, calculated using PRELIS. It is also possible to construct a SEM that incorporates tests of mean structure, when item mean scores from the sample are provided as input.

Several statistical estimation methods are available, including some distribution-free methods, weighted least squares, unweighted least squares, and maximum likelihood estimation. The latter is usually preferred for large samples and is based on an iterative procedure that minimizes discrepancies between expected and observed response patterns in the data. In the case where binary data are analyzed using an asymptotic matrix of item covariances, a weighted least squares analysis method must be specified. There are several authors who provide technical descriptions of these mathematical estimation methods and recommendations appropriate for SEM analyses (Bentler, 1986; Jöreskog & Sörbom, 1999; Muthen, 1984).

Generally, the overall fit of the SEM is evaluated based on a comparison between the SEM and a null model for item covariances (that no structure is present). The residuals obtained when the SEM and null models are compared result in a minimum fit χ^2 value for the data. This statistic is known to be sensitive to sample size and to any departure from multivariate normality in the item response data (Bentler & Bonett, 1980). A variety of alternative indices have been developed. The current convention followed by most authors is to report several indices of model fit (Byrne, 1997).

Popular choices include the root mean square error of approximation (RMSEA), the goodness of fit index (GFI), and the comparative fit index (CFI). The RMSEA is a measure of the discrepancy for the SEM applied to the estimated population covariance matrix, taking into account model complexity via degrees of freedom. A RMSEA value under .05 is interpreted as an indication of good fit. The GFI is an absolute index of fit, and compares the SEM with the null model hypothesis that no structure is present in the data. The GFI computation estimates the relative amount of variance and covariance in the input data that is jointly explained by the implied population covariance matrix. Model parsimony is not accounted for with the GFI, therefore it is most often reported along with other (adjusted) indices. The CFI has been selected by Bentler (1990) as the index of choice for SEM. The CFI is adjusted for model parsimony and for sample size. The fit of the SEM is compared with the fit of a specified "independence" model (representing zero correlations among all variables) in calculating the CFI. Values above .90 for the CFI and GFI are considered evidence of good fit for the SEM.

SEM may be applied to response data elicited using a variety of item formats, including continuous, binary, or categorical responses (McDonald, 1982; Muthen, 1984) and with survey data (Andrews, 1984). Bollen and Lennox (1991) have described possible advantages of SEM techniques for measures that are composed of items covering diverse content, citing the CES-D in particular. They stated that SEM is an appropriate method to examine heterogeneous scales; the CES-D contains some "causal" (e.g., loneliness) and some "effect" (e.g., sadness) indicators. Indeed, the heterogeneity of items and possible multidimensionality in the CES-D may violate assumptions of CTT-based evidence for reliability and validity.

Method Effects.

Andrews (1984) noted that survey measures may contain systematic response patterns related to wording or response option formats for items. These may be termed method effects, and it is possible to specify these features in the SEM. One example of method effects was presented by Higgins, Zumbo, and Hay (1999), who modeled context dependent item sets from a measure of attributional style (a risk factor for depression). They described the identification of method effects:

> Factor analysis using structural equation models attempts to reproduce the covariation among items comprising a scale by postulating latent variables that account for the covariation. If there is a noteworthy amount of unaccounted-for covariation over and above that attributable to the latent variables, it is commonly assumed that this is due to an insufficient number of latent variables and/or an incorrect factor patterning (i.e. some or all items do not measure on the specified factors). However, this is only true if we assume that there are no other sources of covariation in the observed data than the individual differences (latent) variables. (Higgins et al., 1999, p. 7).

When method effects are not accounted for, factor solutions may be misleading. This is a particular concern when using scales that contain both negatively and positively worded items. Carmines and Zeller (1979) emphasized the importance of method effects (also termed response sets), in their consideration of the use of factor analysis in reliability and validity assessment. Using a self-esteem measure as an example, they demonstrated that the correct structure and dimensionality of the scale was accurately

identified when one accounts for method artifacts in the specification of latent factors (Carmines & Zeller, 1979, p.70).

In a similar study, Tomas and Oliver (1999) recently examined alternative test models for a self-esteem scale. They concluded, "...if method factors are not modeled in factorial validity studies in which positively and negatively worded items are present, the underlying structure of the responses to the scales can be obscured by method bias" (Tomas & Oliver, 1999, p.94). The inclusion of method effects in a scoring model based on SEM is now possible. This type of scoring, and tests of possible group bias based on SEM, has not been prevalent in educational measurement research (or, using health scale data).

The incorporation of method effects in the SEM would provide a unique contribution to educational measurement theory. The variance shared by items where method effects are present in responses to items is specified separately in the SEM from the shared variance in item responses that is attributed to the principal construct of the measure. A description of how this may be accomplished for the CES-D is included in the methods section.

SEM analyses may lend important information about potential group-related differences in measurement properties, and scoring based on a specified SEM that incorporates method effects may yield more accurate individual scores. Clearly, there are important practical implications for item development, scale scoring, and validation studies of applied measures containing positively and negatively worded items.

Item Statistics

This section includes a general description of the relationship between item statistics from different test models, and some evidence from the literature where these item statistics have been compared. Item and test statistics based on CTT depend on the characteristics of the sample that provide responses. The section concludes with a presentation of studies where the hypothesis of sample invariance of item statistics has been tested. Particular emphasis is given to the mathematical techniques appropriate to determining the sample-invariance properties of these statistics. This information forms an important step in developing the analytic methods for the case study, to be proposed and described in the methods chapter.

Theoretical treatments of the differences between CTT and IRM have been presented by several researchers (e.g., Crocker & Algina, 1986; Hambleton & Jones, 1993; Hulin, Drasgow, & Parsons, 1983; Linn, 1990). The ability of an individual is presented mathematically with CTT at the level of the observed test score. The simple linear total score is calculated as a sum of responses to items, and item statistics are based on the proportion of correct responses (difficulty, p) and the correlation between item and total scores (discrimination, r or r_{pb}). These statistics are sensitive to the average ability level in the sample of responses and to the variance and scale of item and total scores. CTT item and test statistics and individual scoring have been described as sample dependent, and therefore are less precise when compared to invariant IRM analyses and scoring (Crocker & Algina, 1986).

Lord (1980) gave a formula representing the relationship between CTT and IRM item statistics. This relationship was described as approximate by Hambleton and Jones

(1993), because the distribution of assigned scores in each model has a different shape and they are related non-linearly. Also, the CTT observed score includes measurement errors, whereas the IRM ability score does not (Hambleton & Jones, 1993, p. 43). The relationship between IRM item discrimination and CTT item-total correlations is:

$$a_i \cong \frac{r_i}{\sqrt{1 - r_i^2}}$$

where a_i is the item discrimination parameter value for item i used in IRT and r_i is the item biserial correlation. The relationship between CTT and IRM item thresholds is partly dependant on the discrimination of items. A general relationship for this estimate in each model is also given by Lord (1980):

$$b_i \cong y_i / r_i$$

where b_i is equal to the IRM item difficulty parameter value for item i and y_i is the normal deviate for the ability score beyond which p_i of the examinees fall.

The notion of item discrimination exists in SEM as the regression of each item on the latent trait. Because the class of IRT models has developed from the foundation of linear factor analysis, it has been possible to define item statistics in modern SEM in a similar way. Ferrando (1996) compared the structure of a measure across multiple groups using a SEM based that was equivalent to a two-parameter IRM.

In an extension of the work by Thissen et al. (1984), Ferrando (1996) presented an analysis of a 22-item affect scale for anxiety with a graded response format. The structure of responses from male and female groups was compared using a model that included a threshold parameter for each item. The item threshold parameter was estimated as the mean observed item score for subjects at the latent trait value of zero when the distribution of the latent trait was constrained as bivariate normal N(0,1). Because item thresholds are actually intercepts in a SEM analysis of mean structures, the terminology is rather confusing. However, Ferrando summarized the relationship between IRM and SEM threshold parameters:

> From the point of view of item analysis, test theory generally considers two parameters in the items: difficulty and discrimination. Explicitly or implicitly, models without intercepts only take into account the discrimination parameter, and so, only make use of partial information for analyzing items (Ferrando, 1996, p. 424).

The inclusion of a threshold SEM parameter allows a test of the hypothesis that differences in the mean value for the latent trait were present for men and women, given invariant measurement structure. This was accomplished in that study by specifying the covariance matrix, representing the latent trait, as equal across groups analyzed in a single SEM run. Ferrando (1996) makes the point that this method allows for a test of item statistics across groups: "according to both the classical test theory and the item response theory, the invariance in the Ψ^2 [variance-covariance] matrices is related to the invariance in the item's accuracy of estimating latent trait values" (Ferrando, 1996, p. 434).

An additional link between IRM and SEM exists because the reliability of SEM item scores may be estimated as the ratio between squared factor loadings and error variances for items. This has been shown equivalent to the item information function in IRM for continuous responses (Mellenbergh, 1994). Ferrando (1996) advocates the application of SEM that includes threshold values and mean structure estimates in studies of invariance for continuous response items. McDonald (1985b) had also presented a non-linear factor analysis model equivalent to IRM for binary responses. However, this method has not been widely applied to real binary response data for generating optimal individual scores. In addition, popular SEM computer programs such as LISREL do not provide a method of incorporating tests of mean structures to estimate item thresholds using binary response data.

Therefore, although it is possible to specify a SEM that is mathematically equivalent to the one- or two- parameter non-linear IRM, this methodology is not accessible to the general measurement practitioner. A study of the sample-invariance of item parameters and individual scores that applies the common form of SEM, in comparison with the appropriate IRM, would have broader relevance for educators and researchers interested in selecting an optimal test model.

Sample Invariance

A review of studies in the measurement literature does not clearly establish the superiority of IRM over CTT for sample-invariant item statistics with real data. One reason for this may be that different summary statistics and indices have been applied in studies of invariance. An overview of research using IRM, CTT, and SEM studies of item

invariance is included here, and methods of determining the invariance of item statistics are introduced.

The majority of researchers who provided comparisons of CTT item statistics with IRM item estimates have used educational data (Becker & Forsyth, 1992; Cook, Eigner, & Taft, 1988; Fan, 1998). Some authors have identified a lack of invariance in item parameters derived from the IRM (Cook et al., 1988; Miller & Linn, 1988). Linn (1990) observed that different implications arise regarding the pattern of learning as children progress through school, depending on the scaling method used. Becker and Forsyth (1992) supported this finding in their study of test equating with high-school achievement scores. Work in this area has had important implications for theories of learning and has been used to select optimal scoring methods to estimate educational progress.

Fan (1998) studied CTT and IRM item statistics and individual scores across different samples. Fan examined CTT, one- and two- parameter logistic IRM and concluded that estimates based on all methods were similar. CTT item discrimination estimates were found to be subject to sampling, whereas IRM and CTT item difficulties were more stable across gender and ability groups. Evidence provided by Fan (1998) can be interpreted as partial support for the sample-independence of IRM item statistics. Fan pointed out the lack of empirical evidence provided to date:

> Ufortunately, the view that the argument is moot seems to have occurred largely in the vacuum of empirical evidence, because the literature fails to show that this important premise has been subjected to systematic and rigorous empirical investigation. (Fan, 1998, p.379).

Fan (1998) also gave an example of one method of summarizing the sample-invariance of item statistics from each model for comparison. In that study, a Pearson correlation was used to estimate the association between IRM scores and linear scores. CTT item difficulty values were subtracted from unity, and then z-transformed to a scale similar to IRM difficulty values. Item discrimination values for CTT models were also normalized via a z-transformation, but results based on original r_{pb} values were nearly identical. A z-transformation of this kind does not normalize the distribution of CTT item discrimination values. The justification and rationale for this choice of technique to produce estimates of sample-invariance for item statistics was unclear to this reader.

It is also possible that the selected IRM did not provide good fit to the data in some sub-samples presented in Fan (1998). For example, unidimensionality was assessed via linear FA. Other model fit criteria (such as standard errors of estimates) were not reported for comparison groups. Model fit was assessed using the Likelihood Ratio χ^2 from calibration runs using random samples; mis-fit was identified only for the Rasch model. Interestingly, the author made no comments regarding the appropriateness of CTT linear scoring. This seemed a reasonable concern given mis-fit of the Rasch model. In the Rasch IRM, the analyst imposes the assumption of equal discrimination for items. Therefore, the test of fit can be interpreted as an explicit method of examining the hypothesis that items merit equal weight for total scoring. A simple linear score sums items with equal weights, so, the Rasch model may be viewed as a test of the validity of the simple CTT scoring method. Finally, it is not possible to evaluate the relative validity of alternative scoring methods without some additional evidence.

Research studies of the sample invariance of IRM item statistics use a variety of summary statistics to draw conclusions. Some methods are not useful for real data, as they depend on indices based on simulated data where invariance of item statistics was known (Oshima & Miller, 1990). Others were descriptive and often complicated by the presence of equating across test forms or levels of ability (Becker & Forsyth, 1992; Cook, Eigner & Taft, 1988). A simple correlation coefficient has also been applied to IRM parameters across samples (Cook et al., 1988; Fan, 1998). However, this may not be a useful method because of the non-linearity and lack of a fixed scale for IRM ability estimates in any two samples. Even when invariance is present, item estimates may differ up to a linear transformation of scale (Rudner, 1983).

It is essential that a summary index be selected that will yeild a valid in comparison of invariance properties across CTT, IRM and SEM techniques. CTT item difficulty estimates are expressed as a percentage of the sample (zero to 100, or zero to one), CTT item discrimination estimates are in a similar range based on a correlation coefficient (zero to one). The SEM discrimination estimate is expressed as a standardized regression weight, also ranging from zero to one. IRM item statistics are located on a logistic scale selected for the latent trait (-2 to 2 is most common). The issue of scale dependency for an index of the similarity of item statistics across samples has not been adequately addressed. A summary index based on ANOVA for repeated measures designs will be suggested in the methods chapter to examine the invariance of item statistics across samples in the present study.

The CES-D Scale

The CES-D is one of the most popular brief measures for depression and has been used worldwide (Fechner-Bates et al., 1994; Furukawa, Anraku, Hiroe, Takahashi, & Iida, 1997). The scale has been translated into many languages, including Russian, Spanish, French, Japanese, Italian, American Indian languages, Cantonese, and Mandarin. The 20 item scale was designed to measure recent depression in an adult community sample and scores are intended to inform health planning and epidemiological study of risk factors for depression (Radloff, 1977; Radloff & Locke, 1986). Items for the CES-D were drawn from clinical literature and research that identified the major components of depression symptomology.

Affective components are emphasized, including depressed mood, worthlessness, helplessness, hopelessness, and somatic symptoms (see Table 1). Four items describing positive affect are included in the scale. The scale is administered in an interview or may be completed by the respondent alone. Each item is preceded by a prompt; "please indicate if these thoughts, feelings or events have occurred in the last seven days." Several scoring options have been used for scale items. A binary scoring option is favored for efficiency in large survey administrations. This is termed the "presence" item format; items are scored zero "not at all" or one "at least one time."

Some authors have described the CES-D as a measure of generalized distress, they note that the content of the measure does not conform to that required for a medical diagnosis of depression (Fechner-Bates et al., 1994; Orme, Reis, & Herz, 1986). Radloff (1977) stated that the scale was "...designed for use in studies of the relationships between depression and other variables across population subgroups" (Radloff, 1977,

p.386). The primary use of the scale continues to be self-reports in large community surveys. Scores are used to predict depression in populations, and to plan intervention and treatment strategies.

Table 1 Center for Epidemiologic Studies Depression Scale Items

#	Item Name	
1.	BOTH	I was bothered by things that don't usually bother me.
2.	APPE	I did not feel like eating; my appetite was poor.
3.	SHAK	I felt that I could not shake off the blues even with help from my family or friends.
4.	GOOD	I felt that I was just as good as other people.*
5.	MIND	I had trouble keeping my mind on what I was doing.
6.	DEPR	I felt depressed.
7.	EFFO	I felt that everything I did was an effort.
8.	HOPE	I felt hopeful about the future.*
9.	FAIL	I thought my life had been a failure.
10.	FEAR	I felt fearful.
11.	REST	My sleep was restless.
12.	HAP	I was happy. *
13.	TALK	I talked less than usual.
14.	LONE	I felt lonely.
15.	UNFR	People were unfriendly.
16.	ENJO	I enjoyed life. *
17.	CRYS	I had crying spells.
18.	SAD	I felt sad.
19.	DISL	I felt that people dislike me.
20.	GETG	I could not "get going."

*These positively worded items are reverse coded before scoring.

Properties of CES-D scores have been examined using responses from inpatient and community samples around the world (Fechner-Bates et al., 1994; Furukawa, Hirai, Kitamura, & Takahashi, 1997; Geisser, Roth, & Robinson, 1997; Orme et al., 1986; Radloff, 1977; Radloff & Locke, 1986; Santor et al., 1995; Schein & Koenig, 1997). This overview will be focused on a general description of scoring the CES-D, factor analytic studies of the structure of response data, and the comparability of scores across subgroups.

CES-D Item and Total Scoring

The CES-D was originally designed with four options in each item. These options made up a graded response format. Anchor terms were usually specified as: (1) rarely or none of the time/ less than 1 day; (2) some or a little of the time/ 1-2 days; (3) occasionally or a moderate amount of the time/ 3-4 days; (4) most of the time/5-7 days (Radloff, 1977). This scoring format was intended to identify the presence and the severity of depressive symptoms (Radloff & Locke, 1986).

Two dichotomous scoring methods have also been applied, the most popular is termed the "presence" scoring method. The presence method differentiates between no symptoms and any occurrence of the symptom in the last seven days (the second option or higher). An alternative binary format, termed the persistence method, only counts symptoms that have occurred for at least one or two days (the third option or higher). These item option formats contain only two choices and are easier for self-report or interviewing in surveys. Studies of dichotomous scoring formats for the CES-D have been presented by Furukawa, Hirai, Kitamura, and Takahashi (1997), and by Santor, Zurloff, Ramsay, Cervantes, and Palacios (1995).

The possible range of total scores depends on the response option format used. When graded responses are presented, the four options are scored 0, 1, 2 or 3, respectively. The positively worded items are then reverse coded (ratings are subtracted from 3) and all item ratings are summed for a maximum score of 60. Dichotomous scoring awards one point for each symptom that meets the endorsement criteria defined in the anchor options. The positively worded items are reverse scored (ones recoded as zeros, and zeros recoded as ones). Then each item score is added for a possible total score of 20 for the scale.

There has been evidence based on internal consistency and re-test studies that responses to the graded item response format show acceptable reliability (Radloff, 1977; Radloff & Locke, 1986; McDowell & Newell, 1996). Alpha values for CES-D item scores usually ranged from .85 to .91, and score stability evidence from test-retest and inter-rater comparisons result in Pearson correlations ranging from .76 to .85 (Radloff, 1977). The test-retest or inter-rater reliability for dichotomous scoring options has not yet been presented in the literature. This is an important consideration, as the dichotomous format for the scale item is popular in large surveys and an estimate of reliability provides a good indication of measurement error in individual scores.

The reliability of individual CES-D items has not been well examined, particularly for the binary item format. This may be a special concern because of the presence of positively and negatively worded items. Radloff and Locke (1986) suggested that the four items measuring positive affect may be somewhat unclear to participants when a self-report or interview administration format is used. Callahan and Wolinsky (1994) also noted that non-response in their self-reported survey was highest for those

four items (almost 20% of participants did not complete those items). Response patterns to these four items appear to differ from responses to the other scale items, simply as a function of positive wording.

Furukawa et al. (1997) provided evidence for the properties of some alternative item response formats for the CES-D; these included the graded response format, presence binary scoring, and an alternative binary scoring method indicating the persistence of symptoms. The impact of each scoring method was evaluated using the level of agreement between cut-score classifications and a diagnosis criterion with a first-visit psychiatric sample. The authors concluded that the graded response and presence methods were equally valid for the identification of depressive disorders. The measurement properties of scales composed of different item option formats were not fully described in that study.

Santor and Coyne (1997) examined the usefulness of each of the four graded response options. When significantly more depressed respondents endorsed an option at a higher rate than the next lowest option, the authors described this item option as highly discriminating. The authors concluded that the second and third options discriminated best for most items. A binary item option format was recommended for a revised version (9 items). Santor and Coyne did not explore the impact of item option formats on the distribution or ranking of individual scores.

The conclusions reached by Santor and Coyne (1997) may have been dependent on their choice of CTT item analysis methods. Item discrimination was evaluated by examining the proportion of people in diagnosed or normal groups who endorsed an item option. This method introduced the problem of criterion sufficiency, as item responses

were assumed to be indications for diagnosed cases of depression. This clinical application is not strictly in line with the original purpose of the CES-D as described by Radloff (1977). However, this study gives some empirical evidence for a shortened (9 item) version of the CES-D. The validity of score interpretations from a shorter form of the scale requires additional support.

Structure of CES-D Responses

Many researchers have examined the structure of responses to the CES-D scale, beginning with Radloff (1977) who identified four general factors. Principal components analysis was performed using the inter-item correlations from graded responses to items. The four factors corresponded to items describing depressed affect, positive affect, somatic symptoms, and interpersonal relations. Radloff and Locke (1986) identified the same four factors in a later study, and noted a high degree of inter-correlation among factors. This general structure has been replicated in a variety of later factor analytic studies (Callahan & Wolinsky, 1994; Chapleski, Lamphere, Kaczynski, Lichtenberg, & Dwyer, 1997; Dershem, Patsiorkovski, & O'Brien, 1996; Hertzog, Van Alsten, Usula, Hultsch, & Dixon, 1990; Knight, Williams, McGee, & Olaman, 1997; Orme et al., 1986; Ying, 1988).

An example of typical structure for the CES-D was presented by Knight et al. (1997) who tested the fit of a confirmatory factor analysis model to their data. They obtained a good fit to responses from a large sample of women residing in the community. The four factors were strongly associated with each other, with correlations between depressed affect, somatic, and interpersonal factors ranging from .59 to .84. Positive affect was negatively associated with all three symptom factors (-.27 to -.55).

Item-factor loadings were all significant, ranging from .44 to .82. A hierarchical model with a super-ordinate construct for depression was also tested. A second order model was supported based on examination of some statistical criteria; however, the authors concluded that this complex model was not a substantively meaningful improvement over the original model proposed by Radloff (1977).

Confirmatory factor analysis solutions for the CES-D have also been presented in the literature with data from different population sub-groups as evidence of construct validity. Strommel, Given, Given, Hripsime, Kalaian, Schultz, and McCorkle (1993) tested the invariance of CES-D responses across gender groups. Two items were identified as possibly biased (talked less, and cried more than usual). Unfortunately, in that study a non-equivalent sample was used to cross validate the findings; cancer patients were compared with caregivers and the results were not replicated. In addition, the clinical sample and caregivers responded to the four-option item format of the CES-D. The question of gender bias and bias due to age or health factors in the binary item format used for large surveys merits rigorous study.

Researchers have not yet examined the possibility of method effects related to the four positively worded items in the CES-D, although the possible importance of item wording has been noted (Callahan & Wolinsky, 1994; Radloff, 1977; Radloff & Locke 1986). In studies where exploratory or confirmatory factor analysis methods have been applied, the dimensionality of responses has been the focus, and statistical tests of hypothesized structure or invariance across groups were not provided.

SEM replication of typical CES-D structure that includes tests of model invariance across large random sub-samples and tests of correlated errors for the positive

affect items would add valuable evidence of scale score validity. Analysis of covariance structures also provides estimates of the strength of association between item responses and latent traits (item sensitivity and reliability), tests for scale dimensionality (Byrne, 1994), and provides a model for scoring that may reduce errors of measurement. In addition, the factor analysis based SEM analysis provides a useful theoretical link between earlier research from simple linear CTT models for the CES-D and the non-linear factor analysis (IRM) analyses to be applied in this study. As Brennan (1998) stated, new measurement technology holds promise as a supplement to existing methods and should not be viewed as a replacement for CTT. IRM utility for scoring and test construction may be evaluated via an examination of both statistical and practical evidence for validity (van der Linden & Hambleton, 1997).

Research Questions

The sample invariance of item statistics has been an important contribution of the IRM to testing practice. Properties of tests and items derived using an appropriate IRM have the theoretical advantage of independence from the particular sample that provided responses. Ideally, only a difference in the level of the trait measured by the test would be reflected by differing individual scores. Respondent characteristics (e.g., gender, age, or health) would not be expected to influence scores given underlying equivalence in the level of depression.

Real data examples are lacking to describe the proposed sample invariance of item statistics associated with the IRM. In addition, alternative models for responses to test items are rarely compared to support the theoretical superiority of the IRM due to this invariance property. Given the wide acceptance of IRM techniques in educational

measurement, and the observation that IRMs are preferred because of the invariance property, it seems important to provide an explicit test of the hypothesis that IRM item statistics are more similar across samples of response to a test, when compared with CTT item statistics.

Theoreticians of test validity have emphasized the need for practical demonstrations of the merit of alternative techniques for developing tests and scoring individuals. This must be accomplished using specific cases of real data. This study makes use of data from a health screening test to examine hypotheses related to the sample invariance of item statistics. The choice of an analytic model for responses to this test may influence item selection, scoring individuals, and ultimately effect an important classification decision where consequences for examinees are significant. Furthermore, this test is similar in important respects to short educational screening tests and the results of this study will contribute to our understanding of the relative value of these test models for similar educational measures. Therefore, three research questions are posed to test the sample invariance of item statistics from the IRM, the CTT, and the SEM in this case of real response data:

1. Is there evidence for the superiority of the IRM over CTT and SEM in random replication samples when item discrimination and item difficulty estimates from binary response data are examined for invariance?
2. Is there evidence for the superiority of the IRM over CTT and SEM across comparison groups when item parameters are tested for invariance?

3. Is there a difference in the reliability of scores and the ranking of individuals that is derived from these three test models?

The invariance of item parameter estimates derived using CTT, SEM, and IRM analyses will be tested with random samples in response to Question 1. In response to Question 2, the invariance of item parameter estimates will be tested for each of the three models across substantively important subpopulation groups; specifically these are gender, age, and health categories. The practical impact of each scoring method on score reliability and on ranks for individuals will be tested in response to Question 3.

Chapter 3: Methodology

The methodology chapter is organized to present general study procedures first, followed by specific details of the analysis in later sections. The method of sampling from the original Established Populations for Epidemiologic Studies of the Elderly survey (EPESE) for inclusion in this study is described in the first section. The next section gives an overview of the two levels of analysis involved, and is followed by a more detailed description of these steps.

There are four general steps in testing item parameter invariance across samples, and comparing scoring properties, based on the three analytic models. The steps may be categorized as: testing model assumptions; generating item statistics; conducting invariance tests; and, comparing the rank of individuals and the reliability of scores based on each model. A table summarizing the analysis tasks is provided in the final section of this chapter (page 66). The final section of the chapter describes methods used to test three alternative hypotheses. These are included to extend these results to cases with normalized response data, shorter (or optimal) tests, and group calibration studies that examine sample invariance of IRM item statistics. The source of data examined in this study are described below.

Population and Samples

Two sections of a large longitudinal population-based study were used for this study. Responses from a total of 6974 cases were available from the EPESE, where the 20 item CES-D was administered in two cohorts (Taylor, Wallace, Ostfeld, & Blazer, 1998).

The public use data were available to the Inter-University Consortium for Political and Social Research (ICPSR), of which the University of Ottawa is a member.

The EPESE was undertaken to identify predictors of mortality, hospitalization, need for long term care for the elderly, and to identify risk factors for chronic diseases and disability. Four locations in the United States were surveyed: (1) East Boston and Massachusetts; (2) New Haven and Connecticut; (3) Iowa and Washington counties, Iowa; and (4) several counties in North Carolina.

The sampling frame for the entire EPESE study and coding for all variables are described in detail by the principal investigators (Taylor et al., 1998). Only the New Haven and Duke (North Carolina) survey data include responses to the 20 item CES-D collected in personal interviews, comprising 2812 and 4162 participants, respectively. The binary (presence) item option format was used to represent CES-D responses for all people included in this analysis. The EPESE data also includes information on general health status and demographic factors, such as age and gender, used to identify sub-samples for this study (see Appendix A).

Levels of Analysis

Three of the four analysis steps are based only on the raw response data from the sample. In contrast, the test of sample-invariance of item statistics depends on a summary analysis that uses the item parameter estimates from each model as input. These two phases can be viewed as separate components of study methodology. Some clarification of these components is presented in this section, to provide a distinction between analyses performed using the raw item response data and later analyses that are based only on the item statistics estimated. After the explanatory discussion of the level of analysis used in

each component of the methodology, the four steps of the methodology are described in the sequence that they naturally occur.

The raw data analysis includes drawing repeated samples of responses from the raw data to form sub-samples of data representing random and comparison groups. An examination of the assumptions required for the IRM and SEM is based on the raw response data. Item statistics and individual scores were calculated for CTT, SEM, and IRM using the raw response data, and this data also form the basis for the post-hoc tests reported in this study.

A separate summary data set was constructed to provide a basis for determining the invariance of item statistics across sampling conditions. The data set contained item statistics generated from each of the CTT, the IRM, and SEM techniques. It was necessary that all item statistics be recorded in a rectangular data matrix from each test model. For example, in the random sampling condition each matrix was 10 x 20. These dimensions refer to the 20 CES-D items on the vertical axis, and the horizontal axis consisted of the 10 item statistics obtained from each of the random samples. Overall, summary data sets were produced to record item statistics from CTT, IRM, and SEM in 10 random samples and 12 comparison samples. A method of examining the closeness of item estimates was applied to analyze these summary data sets. This component of the analysis yielded a single estimate of invariance for each item statistic derived from a test model across samples.

Analysis Steps

The analysis of the raw responses to CES-D items are described in the following section. Specifically, details on re-sampling and tests of model assumptions are described.

Next, the methodology applied to obtain model estimates, tests of sample invariance of item statistics, and finally, the methods used for post-hoc tests (see Table 3 for a summary of each major analysis step, page 66).

Re-Sampling

Ten random calibration samples, each representing 7.6 % of the pooled New Haven and Duke data (500 cases in each), were drawn to establish baseline estimates for each model (see Table 2). These samples were analyzed separately to test the invariance of item parameter estimates across random replication samples. Tests of model assumptions and fit were based on analyses of the first random 500 cases. Individuals in this sample were also scored using each analytic model to compare ranking of individual respondents, and score reliability based on each test model.

Table 2 Re-Sampling Method

Analysis Step	Replication x size	Available cases[a] with complete data
Calibration of each model	1 x 500	6621
Invariance in random samples	10 x 500	6621
Invariance in subgroups[b]		
Gender	2 x 500	2176
Age	2 x 500	2088
Health	2 x 500	2459

[a]Available cases in pooled data with no missing CES-D, age, gender, or health items.

[b]Available cases in smaller of the two categories, pooled data.

The total sample was used to draw responses from 500 men and 500 women at random to allow for invariance tests across gender groups. This method was repeated, without replacement of the original sample, for cross-validation of the gender contrast invariance test with a new sample of 1000 people. In a similar fashion, 500 people less than 74 years of age were selected to represent a group of younger seniors, and 500 people aged 74 or older were selected to represent older adults in this elderly group. Again, this process was repeated without replacement of the initial two samples so that the invariance tests could be cross-validated. The same method was again repeated so that responses from 500 people reporting good or very good physical health could be compared with 500 people who reported only fair, poor, or very poor health. This pairing was also re-sampled for cross-validation.

Tests of Model Assumptions

A set of implicit or explicit assumptions is associated with each test model applied to the data in this case study. The correct dimensionality must be evaluated, if known, for all models. Both IRM and SEM require assumptions related to multivariate analyses of data, in addition to those specific to each test model. Some detail is provided in this section describing the assumptions required for each model.

The CTT model for test scores is often viewed as requiring the fewest assumptions for response data. These assumptions are largely axiomatic and rarely subject to scrutiny. Generally, there must be a linear relationship between the true and observed scores, and error variance is expected to be uncorrelated with true score variance (Crocker & Algina, 1986). When simple linear scoring is used, the items are assumed to be measuring a shared construct with equal accuracy. This assumption of

equal item precision (or equal item weighting in individual scoring) is implicit, although rarely considered empirically.

In this study, the item-total correlations were examined as evidence for the relative precision of item as measures of depression. The consistency of responses to items was estimated with an internal reliability estimate (Cronbach's alpha). The proportion of people who endorsed each item (CTT item difficulty), and the strength of association between item and total scores (CTT item discrimination or reliability) were considered in judging the similarity of items. Different item reliabilities would suggest that a weighted scoring method might improve the accuracy of total scores. Items with unequal difficulties may measure more accurately people who have specific levels (or absence) of depression.

CFA is a special case of the more general SEM. The linear SEM analysis conducted in this study is subject to restrictions on the data for general multivariate analyses that apply to factor analytic procedures. These assumptions include the linearity of relationships between item response variables, multivariate normality of item response distributions, and independence of observations (Stevens, 1996). There are a variety of descriptive and statistical tests appropriate for testing these assumptions. The frequency distributions of item responses are examined as partial evidence for multivariate normality in the response data. Adequate sample sizes are also important; Tinsley and Tinsley (1987) state that 500 subjects provide a very good sample size for general linear factor analysis techniques.

Factorability of the item covariance matrix depends in part on adequate variance for each item. The item correlation matrix was examined prior to SEM analyses for

evidence of adequate associations between item responses. It has been shown that communality estimates based on factor analysis of the matrix of inter-item covariances from dichotomous variables will be attenuated compared with analyses of matrices of similar estimates based on continuous variables (Tinsley & Tinsley, 1987). It is important to select an appropriate form for the input data and to examine the strength of inter-item covariances. SEM analysis of the asymptotic covariance matrix based on tetrachoric item correlations is appropriate for binary data, such as responses to the CES-D. The program PRELIS 2.3 (Jöreskog & Sörbom, 1999) was used to prepare these matrices prior to SEM analysis.

Two restrictions on the response data are common for IRM analyses. These are correct dimensionality (unidimensionality in simple models) and local independence of item responses under the specified structure (Hambleton et al., 1991). The size of residuals after non-linear factor analysis was calculated using the computer program NOHARM (Fraser, 1988) and were then tested with an approximate χ^2 implemented in the CHIDIM program (De Champlain & Tang, in press). Dimensionality assumptions were also tested indirectly via tests of model fit, and these were available in the estimation process for both IRM and SEM (e.g., RMSE, χ^2 tests for the IRM; and minimum fit function χ^2, CFI, and NFI for SEM).

Independence of observations was also considered for these analyses. The EPESE survey was based on a randomized multi-stage probability sampling design, and the possibility of inaccurate variance estimates and standard errors exists. This could lead to incorrect inferences when statistical hypothesis tests are conducted. There are procedures available to correct variance estimates from complex survey samples (Skinner, Holt, &

Smith, 1989). Alternatively, the size of effect due to the survey design may be used as a correction for analyses. Unfortunately, the EPESE data do not include analytic weights that are appropriate to adjust the variances of responses to the CES-D.

The risk of failing to account for the non-random survey design is that estimated variances based on the sample data may not be equivalent to variances derived from a simple random sample of the population (Jessen, 1978). This may result in inaccurate (usually smaller) standard errors around point estimates. The possible impact of complex sampling designs on multivariate analyses, such as SEM or IRM, remains unclear. Many techniques are available to adjust for the sampling design (e.g., devise appropriate weights for each observation via replication methods); however these depend on sampling information that is not available in the EPESE public data release files.

The survey sampling design could not be accounted for in this analysis because information that described the probability sampling units (PSU) for each case was not available in the public release data. Therefore, it was important to examine the characteristics of the included sample as one method to determine the generalizability and clinical meaningfulness of the study. The demographic characteristics of the sample are compared with elderly Canadians to facilitate generalization of the American clinical profile to a similar Canadian sample. Also, as the pattern and possible impact of missing item responses are potentially confounding factors in any research design, an analysis of cases excluded due to missing data was conducted. People represented in the original Duke and New Haven samples of the EPESE data who were not included in this analysis because of missing data were identified and compared to the included sample.

Estimation

Item properties and total scores based on simple linear scoring for the CES-D were calculated using SAS 7.12 for Unix (SAS Institute, 1997). One- and two-parameter logistic models were applied using the computer program BILOG 3.0 (Mislevy & Bock, 1990). The correct dimensionality of responses for IRM was tested via an approximate χ^2 test of residuals after IRM analysis with the programs NOHARM (Fraser & McDonald, 1988) and CHIDIM (De Champlain & Tang, in press). The SEM for responses based on a linear structural relations model was estimated with PRELIS 2.3 and LISREL 8.3 (Jöreskog & Sörbom, 1999). Some details on estimation procedures for each analytic model are provided below.

CTT

Item total-correlations were estimated as point-biserial correlation coefficients. Threshold estimates were represented by a simple proportion; specifically, the number of people who endorsed an item divided by the total number of respondents. Individual scores were calculated as a simple sum of item responses, and the distribution and rankings based on individual scores were estimated using SAS. The closeness of CTT item discrimination and item threshold statistics across samples was calculated using the scale reliability option for ICC provided in SPSS 8.0 for Windows (SPSS Inc., 1999).

IRM

The computer program BILOG was used to generate maximum likelihood estimates of item parameters and individual scores. The distribution of the latent variable was specified based on the posterior distribution of responses. This was a precaution

against expected non-normality in the distribution of total scores (and possibly the latent trait) in the population.

CTT and SEM runs are more comparable with IRM when each random replication is conducted without fixing item estimates based on a prior run. No item parameters were fixed during estimation for IRM analyses in this study. This appeared reasonable because item based equating may provide an unfair advantage in the closeness of IRM estimates across replications if an equating method was used. The CTT item statistics depend in part on the metric of the items (e.g., total scores on the CES-D range from 0 to 20). Both IRM and SEM require that the metric be selected by the analyst (or a default is implemented by the computer programs). In order to aid interpretation, the CTT metric was selected for linear transformed IRM and SEM scores. This was accomplished using a linear transformation for the IRM individual scores (and item statistics) to a distribution equal to the mean and standard deviation of total scores from CTT analyses. A similar strategy was used to standardize the linear SEM scores to the same scale as CTT and IRM total scores.

The best fitting IRM was selected based on the results from fitting unidimensional one- and two- parameter IRM (tested with a χ^2 change statistic). Model fit and sufficiency were assessed using substantive and empirical criteria (e.g., improvement in χ^2 from hierarchical models, CTT item discrimination estimates, evaluation of the size and range of item parameters). The fit of individual items to the model was examined using the residual RMSE values and the relative size of standard errors to respective estimates.

SEM

Weighted least squares analysis of the asymptotic covariance matrix and tetrachoric correlation matrix was conducted according to recommendations for binary (ordinal) response data (Jöreskog & Sörbom, 1999). Estimates from the completely standardized solution were interpreted. A single latent factor was specified for the underlying factor representing depression. This choice appeared reasonable based on the scoring recommendations given by Radloff (1977) and others, and the typical interpretation of CES-D individual scores as representing a single trait. The regression path for item 6 (I felt depressed) was fixed equal to one to aid identification and set the metric of estimates. This item was expected to have a strong relationship to the latent construct based on past research (Santor et al., 1995). Error variances were then estimated for items and for the latent trait.

The baseline model included paths for possible correlated errors among the four positively worded items due to expectation that method effects would result in correlations among responses to these items unrelated to underlying depression. Only those correlated errors that contributed to a significant improvement in model fit (based on the change in χ^2) were retained. The statistical significance of each estimate and a small value for the associated modification index were used as additional criteria for specification of the baseline model. Both empirical (statistical) and substantive criteria (meaningful relationships based on item wording) were used to examine the overall fit of the SEM, including CFI, AGFI and RMSEA estimates. After establishing the baseline model, two-group validation runs were preformed for gender, age, and health contrast groups as an additional test of the invariance of the overall SEM.

Invariance of Item Statistics

IRM invariance studies in the literature have described a variety of summary statistics used to draw conclusions about sample invariance (Fan, 1998). A simple correlation coefficient may not be adequate due to the non-linearity and lack of a fixed scale for IRM and SEM ability estimates. Even when invariance is present, item estimates may differ up to a linear transformation of scale (Rudner, 1983).

The intraclass correlation coefficient (ICC) for absolute agreement appears a reasonable model for examining the proximity of item statistics across samples. This analysis method has also been described as a two-way mixed effects model for examining rater reliability by Shrout & Fleiss (1979). The technique has been used primarily in inter-rater reliability analyses. It is favored when variance across persons is expected and judges are representative of any possible judges, but variation between raters is considered measurement error (Spence-Laschinger, 1992).

A discussion of how the ICC terminology has been adapted for this study of invariance may be useful. Item estimates from different samples were treated in the same way as observations from different judges would be in a reliability study. Instead of comparing the consistency between observations that reflect ratings made by separate judges, in this study the item estimates obtained for each item were compared across samples of data.

Chinn (1990) cautions that valid interpretations of the ICC are possible only when observations are on an equivalent scale. It is not clear how sensitive ICC value may be to across-sample differences in the metric of item parameters from SEM, CTT, or IRM, respectively. It may be most important to consider these differences in the metric of IRM

item statistics when there is an expected lack of invariance. In contrast to CTT and SEM, the IRM metric is selected by the analyst for each run and is partially dependent on mean performances in the sample. Where invariance is present, the metric of IRM item statistics will be equivalent across samples. It is not clear how the ICC estimate may be affected by moderate violations of the invariance assumption required by the IRM. However, it is reasonable to assume that ICC values that represent invariance in item parameter estimates across samples will be closer to unity.

A single ICC was produced using summary data sets for each item statistic using the scale reliability analysis procedure provided in SPSS 8.0 (SPSS Inc., 1999). The ICC summarized invariance in the random sample condition across all 20 items and 10 samples. Therefore, the invariance of item discrimination estimates across random samples for each model was represented by three ICC coefficients, one each for IRM, CTT, and SEM. The invariance of item threshold estimates across random samples is represented by two coefficients: one from IRM; and one from CTT (item thresholds were not modeled in the SEM).

Invariance tests across contrast groups were developed in a similar fashion. The observed item discrimination and threshold statistics were compiled in a summary data set as observations for each item. One ICC was calculated to represent the association between item estimates from males versus females, one for older versus younger groups, and one for those in poor versus good health. These three comparisons were replicated for cross-validation with a new selection of paired groups. This resulted in CTT and IRM item discrimination and item threshold invariance estimates for a total of six pairs, and item discrimination estimates for SEM using the same six pairs of groups.

Individual Scoring

Total scores for the first 500 people were calculated via a simple sum of item responses representing CTT scoring. IRM theta estimates based on the best-fitting IRM and SEM factor scores were linear transformed to a distribution with mean and standard deviation equal to CTT individual scores. This transformation was performed to facilitate comparisons of the individual score distributions on a similar range metric. The baseline SEM was applied to responses from the same 500 people and factor scores were calculated using a method based on Anderson and Rubin (1956). This method of factor scoring exactly reproduces the estimated sample covariance matrix for items in the pattern of individual scores. The similarity in ranks given individuals across the three test models was evaluated using a Spearman correlation coefficient. Also, the shape of the distribution and reliability of total scores was described for each model.

Post-hoc Analyses

Post-hoc analyses were conducted to enhance the generalizability of these findings to other cases of applied data in achievement testing. Some rival hypotheses related to possible idiosyncrasies of this case study warranted special attention. These post-hoc tests were conducted to examine three important issues; the potential improved model fit to a normalized shape of total score distributions in the response data, the possible improved fit to a shortened CES-D with poorly discriminating items deleted, and the potential evidence of lack of invariance based on a group calibration study of sample invariance for IRM item statistics. These issues are elaborated in the next few paragraphs.

First, the possible improvement in the fit of the IRM and SEM was examined when people who did not report any symptoms on the CES-D were removed from the

total sample. Second, the fit of the IRM was also examined after eliminating weak items from the scale. Elimination of people from the sample who did not experience any symptoms would make these results more similar to those found in maximal performance testing.

Third, a total group calibration run was conducted for the IRM to examine possible confounding effects related to generating IRM estimates in health contrast groups separately. The ability estimates from the total group calibration were used to generate new item parameter estimates in each group. The invariance of these item parameters was examined across the health groups to control for average differences in depression in that condition. These three post-hoc tests are described in more detail below.

A distinction may be made between typical performance tests where responses are gathered to determine performance during the educational process, and optimal performance tests where intensive preparation leads to the best possible demonstration of skill attained after a program of study. The first post-hoc test was constructed to examine the possible influence of the use of typical response data in this study, when compared with data that represent optimal performances. The CES-D would be more equivalent to a typical performance tests as the emotional status at a given time is not a function of an intervention (such as instruction) or specific preparation. Truncating the sample has the effect of selecting people who experience more symptoms, the distribution of response data is less skewed and may be more comparable to optimal performance data. In effect, this post-hoc test examines the potential confounding of the distribution of total scores that might be expected from typical versus optimal testing situations.

The shorter version of the CES-D contained only items that were maximally discriminating. The second post-hoc test made use of an optimal item set from the CES-D. Although the CES-D has been refined several times by different researchers over its long history (over 20 years), it seems reasonable to use what is learned from the item analysis in this study to suggest further refinements to the measure. One important step in scoring achievement tests in high stakes assessment includes item elimination after administration. In particular, maximally discriminating items may be selected to enhance the validity of classification decisions when criterion-referenced interpretations are made from test scores.

The third post-hoc test was constructed to examine the invariance of IRM item statistics in the health group comparison based on a common metric for ability and item parameters. This comparison group was selected because it was reasonable to assume that different mean depression scores would be present in groups that have good, compared to poor, overall health. As Chinn (1990) notes, a comparison of item statistics across measures with different metrics may not be meaningful. This post-hoc test involved obtaining IRM ability scores for both health groups in a single IRM scoring run. Next, a logistic regression was used to obtain item threshold and discrimination values for each item based on ability values from good and poor health groups separately. These item statistics were then compared using intra-class and biserial correlations to enable a comparison with the methodology used in the original study methods (e.g., based on separate IRM calibrations for each contrast group). This post-hoc test is similar to studies where item parameter invariance is studied based on high and low ability groups and is sometimes used to examine IRM fit (Hambleton & Swaminathan, 1991).

These three post-hoc analyses were included to aid interpretation of these results with respect to similar short educational or affective screening tests, where criterion referenced decisions are made and similar item selection would be used prior to obtaining final scores and test statistics. A summary of these analysis tasks is presented in Table 3.

Table 3 Analysis Task Summary

1. Calibrate each test model.

    1.1 Pool data sets and randomize cases.

    1.2 Draw one random sample (500 cases) for calibration.

    1.3 Estimate CTT item statistics.

    1.4 Select the appropriate IRM.

    1.5 Fit the baseline SEM.

2. Test the invariance of item parameters in replication sample pairs.

    2.1 Select an additional 9 random samples (500 cases each).

    2.2 Generate IRM, SEM, and CTT estimates in all samples.

    2.3 Estimate ICC values for all item statistics across samples.*

3. Test the invariance of item parameters in comparison samples.

    3.1 Select two samples of 500 each in comparison groups (gender, age, and health).

    3.2 Generate IRM, SEM, and CTT estimates in comparison groups.

    3.3 Estimate ICC values for item statistics across comparison groups.*

        3.3.1 Repeat each comparison sampling for cross-validation and generate ICC values.

Table 3 Analysis Task Summary cont.

4. Compare reliability and ranks of individual scores from each test model.

    4.1 Generate and describe moments of score distributions.

    4.2 Identify changes of ranks for individuals across scoring models.

    4.3 Calculate reliability estimates for total scores from each model.

5. Conduct post-hoc tests.

    5.1 Eliminate responses from people without symptoms and examine IRM and SEM fit.

    5.2 Eliminate low discrimination CES-D items and examine IRM and SEM fit.

    5.3 Compare IRM item statistics across health samples derived from a logistic regression based on ability estimates from a combined IRM scoring run.**

*In the original proposal a simple correlation was suggested to examine invariance. Due to potential bias related to the different scales of item statistics, the ICC summary index appeared an improved analysis method.

** This post-hoc test was included at the request of the external examiner.

Chapter 4: Results

The study findings are presented in this chapter in several sections. These generally follow the same sequence as the analysis steps. First, a description of the sample and tests of the impact of missing response data are provided. Next, the results of tests related to model assumptions are given. The model estimation steps are followed by a description of the empirical and substantive rationale for selecting the specified SEM and IRM and a description of the fit of baseline models. A description of the item statistics derived from each model based on the calibration sample follows. Intraclass correlations are then described that summarize the sample invariance of item statistics among random replications and contrast pairings. Finally, individual score distributions are examined for any impact of ranks or score reliability. The last section of the results chapter is a report of findings from three tests of post-hoc hypotheses.

Sample

The total available sample when the New Haven and Duke EPESE baseline data were pooled contained responses from 6974 people. Of this group, 1253 people did not provide responses to one or more CES-D items or to questions for age, gender or health ratings. The eliminated responses included 476 people from New Haven and 777 people from the Duke sample. A total of 5721 cases were included in the analysis sample, representing people who responded to all CES-D items and to the demographic questions needed to create the contrast groups. The analysis sample is compared to a similar Canadian sample from the National Population Health Survey Canadian (NPHS94-95, Statistics Canada and Health Canada, 1996) in Table 4.

Table 4 Characteristics of the EPESE Study Sample Compared to the Canadian NPHS94-95 Sample

Demographic characteristics	EPESE Total 5721 % (N)	NPHS94-95 Total 2393[a] % (N)
Age at time of interview		
<70 years	35.5 (2033)	34.0 (814)
70 to74 years	28.0 (1600)	29.3 (702)
75 to79 years	18.1 (1035)	19.1 (457)
80 years or over	18.4 (1053)	17.6 (419)
Gender		
Female	62.0 (3545)	57.0 (1363)
Male	38.0 (2176)	43.0 (1029)
Health compared to others same age		
Excellent	13.9 (794)	12.5 (299)
Good	43.1 (2468)	60.9 (1456)
Fair	32.3 (1848)	20.5 (491)
Poor or bad	10.7 (611)	6.1 (146)
Income group (household)[b]		
<5,000	34.6 (1974)	
5,000 to 9,999	28.4 (1623)	no equivalent
10,000 to 14,999	9.5 (542)	data
15,000 or more	12.9 (733)	
Marital status		
Married	41.8 (2234)	59.5 (1424)
Separated, divorced or annulled	10.1 (539)	5.1 (121)
Widowed	48.1 (2571)	35.4 (847)

[a] NPHS sample and proportions are weighted to represent the Canadian population.
[b] 926 people refused to answer or did not know their household income.

A comparison with the Canadian population of elderly was conducted in order to frame any clinical implications of this study. Questions about income asked in the NPHS and EPESE were different as Canadians reported income from all sources for their household, whereas Americans were asked only for current personal income. Therefore, a comparison of income is not possible. Marital status and health questions were similar, and age and gender comparisons were equivalent. The sample included in this study appeared fairly close in demographic features to a comparable Canadian population.

Most elderly people in this study reported good or fair health, when asked to report based on their comparison with others of their age. They belonged to a low earned income (retired) sample, and 63.5% were between 65 and 74 years of age. More than half the sample was female (62%), and this is similar to the expected proportion of women in this age group in the general population in Canada. The mean sample CES-D score in the total sample was 3.79, (SD 3.9), based on simple linear scoring. Response distributions for items and total scores are fully described in the results section for tests of model assumptions.

Missing Responses

The people who were excluded due to some missing data on items or covariates were compared with the included sample on essential variables to ensure that the study analysis reflected the original EPESE stratified random sample. It is possible to analyze data with missing responses using IRM or SEM techniques; however, an equivalent approach is not possible for generating CTT item statistics or CTT total scores for individuals with missing response data. It seemed a more valid comparison of test models would result from a study sample of complete response data. Therefore, several steps were followed to examine the possible impact of excluding people with missing responses.

First, the data from the group of people who omitted responses for items were examined to determine if those patterns were related to particular items. The proportion of missing responses in the original sample ranged from 4% to 5 % for all but one item. Item 8 ("I felt hopeful about the future") had 9% missing responses, somewhat higher

than other items. Otherwise, there does not appear to be a pattern whereby people tended to miss certain items more often than other items.

The proportion of men who did not provide complete data was not significantly different from the proportion of women ($\chi^2 = 1.83$, $p > .05$). A bivariate test of the relationship between reported health and age of people in the included and the original sample indicated a pattern. Older people, and those in poor health had a statistically significantly greater tendency to omit items.

A logistic regression model for the predictive association between age, health, and gender categories for incomplete or missing responses was tested to examine the representativeness of the included sample. The logistic model was significant based on an omnibus χ^2 test ($\chi^2_{(6)} = 71.5, p < .05$). Interaction effects between predictor variables were also included in the model. Results indicated that only health status was a significant predictor of exclusion due to missing data. The odds ratio for health status indicated that those in poor or very poor health were nearly twice as likely to have incomplete data (odds ratio was 1.98). The regression coefficient for health was significant based on the Wald χ^2 test ($\chi^2 = 5.1, p < .05$). Gender, age, or interactions among age, gender, and health factors were not significantly related to missing responses. Fifty-seven percent of the original New Haven and Duke EPESE sample reported poor or very poor health, whereas the included sample for this study contained 46% of people who reported ill health.

Study findings based on people who provided complete response data would represent a somewhat healthier group than the larger EPESE sample. An examination of the responses from Canadians to the NPHS94-95 survey indicates that Canadians report

higher levels of good health than do the American EPESE sample. Therefore, the included sample may in fact be more similar to the Canadian population of elderly than the original EPESE sample. This supports clinical interpretations based on the CES-D score properties in this sample to Canadian elderly.

The first random sample of 500 people drawn from the combined data set was used as a calibration sample to specify the baseline IRM and SEM for this analysis. Tests of model assumptions were also conducted using responses from these people. Particular assumptions that are necessary for each test model are considered in the next section. A discussion of the appropriateness of the response data for each test model is included later in the model fit and estimation summaries.

Model Assumptions

The distribution of simple linear total scores was unimodal, with a mean value (3.74) within the range of 20 possible points. This mean score was consistent with means scores in other North American and European community samples (Fava, 1983; Orme, Reis, & Herz, 1986; Radloff, 1977). The standard deviation around mean scores was 3.90, indicating some variability in total scores. The modal score in the calibration sample was zero, and the maximum score was 20. The value of the skew for total scores in this distribution was 1.30. Tabachnick and Fidell (1989) provide a simple test of the significance of skew (that the value differs from zero) for scores based when multivariate models are applied. The standard error of the skew was computed as the square root of the argument given when 6 is divided by the sample size (Tabachnick & Fidell, 1989, p. 72). A z-test was used to test the ratio of the skew to the obtained standard error. This formula resulted in a statistically significant value for skew in total scores ($z = 11.8$, $p >$

.05). This skew in total scores may attenuate coefficients used to compare individual scores from each model. However, this non-normality is not expected to lead to more favorable invariance findings for any particular test model; any attenuation in total score variance would lead to more conservative findings when rankings from scoring model are compared.

CTT

Classical reliability analyses provided some information about the suitability of the CTT model. The internal consistency index (Cronbach's alpha) for CES-D items was α= .85. Item variances ranged from .058 to .213, with a mean of M=.15. Table 5 lists the proportion of people in the calibration sample who endorsed each item. Less than 30% of the sample endorsed any single item. The items ranged from a low of 6.2% ("I felt that people dislike me") to a maximum of 30% ("I felt hopeful about the future"). This evidence of non-normality or negative skew in item responses is due to the low prevalence of depressive symptoms in the general population. As a result, some attenuation in the covariances between items may be expected.

The covariation among item responses indicated that some communality was present, this is consistent with the interpretation of a causal latent trait. Nearly all of the inter-item correlations were positive and statistically significant at $p < .05$ (see Appendix C). The associations between items and the high value for Cronbach's alpha provide some evidence that the basic assumptions for CTT are satisfied in this case. This information is suggestive of possible structure, however dimensionality is explored formally in the next analysis step.

Table 5 Proportion of EPESE Study Participants Endorsing Each Item

# Item name	%	# Item name	%
1. BOTH	19.8	11. REST	27.0
2. APPE	19.8	12. HAPP	20.0
3. SHAK	15.4	13. TALK	16.4
4. GOOD	8.4	14. LONE	26.4
5. MIND	21.6	15. UNFR	11.6
6. DEPR	26.8	16. ENJO	14.0
7. EFFO	29.0	17. CRYS	8.2
8. HOPE	30.6	18. SAD	28.2
9. FAIL	9.6	19. DISL	6.2
10. FEAR	14.2	20. GETG	21.2

IRM

A set of assumptions set out by Hambleton et al., (1991) was considered for IRM, and correct dimensionality is a primary concern. An examination of residuals after a two-parameter IRM was fit to the calibration sample of 500 responses using CHIDIM. The results indicated some complexity in structure for the CES-D when an approximate χ^2 based on the unidimensional IRM was compared with an IRM allowing for two dimensions. The CHIDIM χ^2 difference test was statistically significant ($z = 3.62$, $p < .05$). Additional criteria for correct dimensionality were then considered in order to evaluate the potential importance of a violation of the assumption of IRM unidimensionality.

Results of the NOHARM analyses provided some evidence for a unidimensional IRM for these data. Root mean square residual (RMSR) values that are less than the

typical standard error of the residuals indicate good fit (Fraser & McDonald, 1988). The typical standard error for this sample was .179 (calculated as 4 times the reciprocal of the square root of the sample size). The RMSR value for this sample was .008, and indicates good model-data fit. There was little reduction in the residuals when the unidimensional model was compared to a two-dimensional IRM (RMSR = .007).

The rotated oblique factor pattern for the two-dimensional IRM harmonic factor analysis solution did not appear consistent with distinct domains for items. Items with significant item-factor loadings on the second factor all had minor or moderate item-factor loadings on the first factor (up to .34). The two factors were also highly inter-correlated (r = .72). Unique variances in the unidimensional and the two-dimensional solutions were similar. Items 4, 8, 15, and 19 had the largest unique variances in both solutions. The size of unique variances for these items ranged from .76 to .87 in the one-dimensional solution, compared with a range from .76 to .83 for the same items in the two-dimensional solution. These were two positive affect items and two items describing feelings of alienation. This evidence suggested that a unidimensional IRM was reasonable for CES-D item responses although the underlying construct may be complex, representing facets of depression that are subtly different.

SEM

Linearity of relationships between variables and multivariate normality are common assumptions for multivariate analyses. There was evidence of non-normality present in the patterns of endorsement for items; however, the size of inter-item correlations provided support for the assumption of adequate linear associations among item responses (see Appendix C). Sample sizes were also sufficient (500). Overall, the

principal data conditions necessary for SEM appeared reasonable. Evidence of SEM fit to the data is considered in a separate section of results.

Cudek (1989) has stated that factor analysis of the inter-item covariance matrix is preferable to an analysis of the simple inter-item correlations. As linear factor analysis of dichotomous item responses may be confounded by the unequal difficulties of items and the non-linear relationship to the latent trait, the asymptotic matrix of tetrachoric coefficients is more appropriate for binary data (Byrne, 1997). Use of this matrix for SEM will account for the asymptotic properties of binary fixed response data. In this case study, the asymptotic matrix of inter-item covariances and the matrix of tetrachoric item correlations were used as input for all SEM analyses. This consideration was important in that the item thresholds could not be incorporated into SEM for the CES-D. The means required to estimate item thresholds were not produced by the available PRELIS program when the asymptotic matrix needed for LISREL analysis of binary data was prepared.

Estimation of Models

This section describes evidence and decisions made in the analysis steps required to obtain baseline models for the SEM and IRM. CTT item and test statistics from the calibration sample are presented first to offer some information about decisions made for subsequent IRM selection and SEM specification. Selection criteria and fitting procedures are then described with the results for IRM and SEM. A full description of the selection and fit of the IRM and SEM appears for each, along with the obtained item and test statistics in the calibration sample. Results of the invariance tests across samples follow this description of estimation using the three test models.

CTT

The CTT item statistics showed a wide range for item discrimination and threshold estimates. The corrected item-total Pearson correlation may be used with dichotomously scored variables; with this formula, the response to the given item is not part of the total score computed. The biserial correlation formula is preferred when the underlying trait is assumed to be normally distributed in the population (Crocker & Algina, 1986). Either coefficient may be interpreted as an item discrimination statistic. The biserial coefficients were consistently higher, and are summarized in Table 6.

Discrimination of items ranged from r_{pb} =.37 (item 4) to r_{pb}=.81 (item 3). The most highly discriminating item was item 3 ("I felt that I could not shake off the blues..."), followed by item 6 ("I felt depressed"), and then items 18, 9, and 20 (describing sadness, feelings of failure and lack of motivation, respectively). Less discriminating items were 4, 8, and 19; these items reflect feeling as good as others, feeling hopeful about the future, and feeling disliked by others. These items represent less precise measures of the depression construct.

The percentage of people who endorsed each item is the common form for estimating CTT item thresholds (with a potential range from 0 to 100%). As noted in the investigation of model assumptions, the proportion of people who endorsed any item ranged from 6% (item 18) to 31% (item 8). The most commonly endorsed items were 8 ("I felt hopeful about the future") and 7 ("I felt everything I did was an effort"). Items 18, 6, and 14 were also endorsed by a large proportion of people, describing sadness, depression and loneliness, respectively. Items with low endorsement levels included 17, 4, and 9; these describe crying spells, feeling one is just as good as others, and feelings of

failure. These items describe less frequently reported symptoms and may be associated with more severe levels of depression.

CTT item statistics are often considered when selecting an appropriate IRM. For example, when there are a range of CTT item difficulties, an IRM that accounts for both item difficulty and item discrimination would be expected to provide the best fit to the response data. This evidence and other data considerations are contained in the next section for the selection of an appropriate IRM.

IRM

A correctly specified IRM may account for any or all of, a range of item features. The three-parameter IRM would include parameters for item thresholds, discrimination, and the presence of guessing in the response data (Hambleton et al., 1991). The possibility of guessing exists when items have a forced choice option format. The CES-D response data was evaluated for potential guessing using response patterns from low scorers. If people with low scores endorsed items that have high item thresholds (few others reach the item), then spurious responses may be present.

People who had the lowest CES-D scores did not endorse any items. In fact, the modal simple linear score was zero in the calibration sample. Therefore, a non-zero lower asymptote (*c* parameter) in the IRM for possible guessing does not appear to be justified. The range of CTT item threshold and discrimination statistics may be interpreted as support for a two-parameter logistic model for responses to the CES-D.

Table 6 Classical Item Statistics

Item Number	Name	%	Correlations	
			Corrected Pearson	Biserial
1	BOTR	19.8	.469	.672
2	APPE	19.8	.412	.591
3	SHAK	15.4	.534	.816
4	GOOD	8.4	.208	.374
5	MIND	21.6	.438	.615
6	DEPR	26.8	.595	.800
7	EFFO	29.0	.403	.534
8	HOPE	30.6	.327	.429
9	FAIL	9.6	.434	.750
10	FEAR	14.2	.469	.729
11	REST	27.0	.386	.518
12	HAPP	20.0	.511	.731
13	TALK	16.4	.438	.656
14	LONE	26.4	.479	.646
15	UNFR	11.6	.323	.530
16	ENJO	14.0	.411	.641
17	CRYS	8.2	.405	.734
18	SAD	28.2	.590	.785
19	DISL	6.2	.218	.430
20	GETG	21.2	.522	.736

A variety of criteria for model-data fit were considered when the best IRM was selected for the CES-D calibration sample. Model selection included a comparison of the fit of a one-parameter IRM to a two-parameter IRM via an examination of residuals using an overall χ^2. In addition, the size of item parameter estimates and the residuals for individual items based on different IRM were considered.

The difference in overall χ^2 indicated a superior fit to the data for a two-parameter IRM when compared with a one-parameter IRM. The omnibus test of the two-parameter IRM based on twice the negative log likelihood was $\chi^2(117) = 149.3$, whereas the two-parameter IRM resulted in a larger $\chi^2(111) = 175.1$. A simple difference test of the change in residuals associated with the more restrictive model resulted in $\Delta\chi^2(6) = 25.1$. This exceeded the critical value for a χ^2 statistic at the $p < .05$ level of significance.

Obtained item estimates appeared reasonable from both models, with larger standard errors from the less restrictive (one-parameter) IRM. Item discrimination parameters are usually estimated in the interval between 0 and 2, and item threshold estimates would lie approximately between –2 and 2. The one-parameter IRM analysis resulted in an estimate for a common item slopes of a=.86, and item discrimination values were in a reasonable range. The average slope for the two-parameter IRM was a = .90, and values ranged from a =.52 (item 4) to a = 1.40 (item 18). Estimates from the two-parameter IRM are presented in Table 7.

Table 7 IRM Item Statistics

Item	Slope *(SE)*	Threshold *(SE)*	Chisq df *(p)*
1. BOTR	0.891	1.304	10.2 6.0
	(0.113)	(0.120)	(.12)
2. APPE	0.767	1.418	2.7 6.0
	(0.099)	(0.144)	(.84)
3. SHAK	1.270	1.339	4.3 5.0
	(0.175)	(0.098)	(.51)
4. GOOD	0.520	3.069	6.7 6.0
	(0.105)	(0.523)	(.35)
5. MIND	0.770	1.308	6.8 6.0
	(0.102)	(0.143)	(.34)
6. DEPR	1.393	0.787	8.4 5.0
	(0.171)	(0.072)	(.14)
7. EFFO	0.694	0.980	16.6 7.0
	(0.093)	(0.127)	(.02)
8. HOPE	0.534	1.072	15.1 7.0
	(0.076)	(0.158)	(.04)
9. FAIL	1.016	1.889	6.1 5.0
	(0.143)	(0.169)	(.30)
10. FEAR	0.990	1.565	3.9 6.0
	(0.133)	(0.140)	(.70)
11. REST	0.658	1.122	16.5 7.0
	(0.087)	(0.142)	(.02)
12. HAPP	1.004	1.218	9.2 6.0
	(0.122)	(0.105)	(.16)
13. TALK	0.844	1.549	5.7 6.0
	(0.112)	(0.151)	(.46)
14. LONE	0.965	0.931	3.9 6.0
	(0.113)	(0.091)	(.69)
15. UNFR	0.711	2.109	7.1 6.0
	(0.116)	(0.249)	(.31)
16. ENJO	0.849	1.709	3.4 6.0
	(0.116)	(0.154)	(.76)
17. CRYS	0.988	2.046	7.6 5.0
	(0.155)	(0.201)	(.18)
18. SAD	1.396	0.734	10.7 5.0
	(0.163)	(0.064)	(.06)
19. DISL	0.644	2.948	3.5 5.0
	(0.129)	(0.455)	(.62)
20. GETG	1.049	1.134	1.1 6.0
	(0.124)	(0.098)	(.98)

The residuals around item estimates are used as a summary for the closeness of predictions based on the IRM with observed responses. Item χ^2 statistics with a probability of less than 5% were judged to indicate poor fit of the specified IRM to the item. The one-parameter IRM showed poor fit to 5 items (items 4, 7, 8, 15 and 18). There was poor fit of the two-parameter IRM for only two of the same items and one additional item (items 7, 8, and 11). The range of discrimination (*a)* values from the two-parameter IRM also provided support for the more complex model.

Finally, the overall reliability of optimal scores based on each IRM was considered additional evidence of model fit. The IRM reliability estimate is computed using the average standard error around individual IRM scores for a specified range of ability. These IRM reliability indices indicate average precision for individual scores. IRM reliability provides an exact estimate for the reliability of CES-D scores, compared to traditional lower bound estimates such as Cronbach's alpha (Mislevy & Bock, 1990). The IRM reliability values for the CES-D calibration samples were similar (.84 and .83, respectively). Based on all the evidence of model fit considered, the two-parameter unidimensional IRM appeared to provide slightly better fit to these data. Invariance tests of IRM item statistics and IRM individual scores calculated for the remaining steps of the study were based on the two-parameter IRM only.

The IRM discrimination estimates can be interpreted for items 18, 3, 9, and 20 as more precise indicators of depression. These items describe sadness, inability to shake off blues, feelings of failure, and lack of motivation. Items with lower *a* estimates were 4, 8,

11 and 19. These items describe feeling as good as others, feeling hopeful, having restless sleep, and feeling disliked.

Ranks based on the IRM threshold values indicate their ordering on a continuum of severity of depression. Items with lower *b* values are interpreted as more accurate measures of severe depression, whereas responses to items with higher threshold estimates are more precise indications of lower levels of depression (or absence of depression). Item ordering based on IRM and CTT thresholds are compared in Table 8.

Table 8 IRM and CTT Threshold Ranks

	IRM Rank Order (*a* estimate)	Order	CTT Rank Order (% endorsed)
	SAD (.73)	1	HOPE (.31)
	DEPR (.79)	2	EFFO (.29)
More severe	LONE (.93)	3	SAD (.28)
depression	EFFO (.98)	4	DEPR (.27)
	HOPE (1.07)	5	REST (.27)
	REST (1.12)	6	LONE (.26)
	GETG (1.13)	7	GETG (.22)
	HAPP (1.22)	8	MIND (.22)
	BOTR (1.30)	9	HAPP (.20)
Moderate	MIND (1.31)	10	APPE (.20)
	SHAK (1.34)	11	BOTR (.20)
	APPE (1.42)	12	TALK (.16)
	TALK(1.55)	13	ENJO (.14)
	FEAR (1.57)	14	SHAK (.15)
	ENJO (1.71)	15	FEAR (.14)
Low or	FAIL (1.89)	16	UNFR (.12)
no depression	CRYS (2.05)	17	FAIL (.09)
	UNFR (2.11)	18	GOOD (.08)
	DISL (2.95)	19	CRYS (.08)
	GOOD (3.07)	20	DISL (.06)

The order of items based on CTT ranks was close to that found with IRM item threshold values. While the ranks did not correspond exactly, most CTT threshold values

placed items only two or three positions from the rank derived from the IRM threshold value. The greatest difference occurred for HOPE ("I felt hopeful about the future"). This item was ranked at the most severe end of the CTT threshold scale, and was ranked fifth in severity by IRM. The item SAD "I felt sad" was placed at the lowest IRM threshold.

SEM

A single underlying factor for depression was estimated from the asymptotic matrix of tetrachoric correlations between responses to CES-D items. The measurement model was initially specified without paths between error terms for items. It was expected that paths between positively worded items would be statistically significant and would improve the fit of the model. The regression path of item 6 "I felt depressed" on the latent factor was constrained to unity to aid model identification. Successive models were then tested and compared to the more parsimonious SEM in order to evaluate possible method effects for items. The initial model resulted in a $\chi^2(170)=683.8$ ($p > .05$).

The improvement in model fit obtained with the inclusion of correlated item errors is shown in Table 9. Indices of mis-fit for comparing successive models that were examined included the comparative fit index (CFI) and change in χ^2 the associated test of statistical significance at a the 5% level. There were six possible correlated error terms for positively worded items. In addition, a path between items 15 and 19 was tested based on model fit information provided in the earlier results. An expected change in the minimum fit χ^2 is presented for each parameter not estimated in the SEM. Based on initial estimation runs, a large value (modification index) representing a correlation between error terms linking UNFR and DISL was found. That path between alienation items 15 and 19 was also freed for estimation for the baseline SEM.

A parsimonious baseline model was therefore specified to incorporate meaningful error terms that also led to significant changes in the overall χ^2. The variance shared by items with positive wording, or by items that describe alienation, may be differentiated from the common variance in CES-D items that is explained by the underlying trait (depression). Method effects were accounted for in the SEM via correlation paths among the error terms for these items. This model incorporated four of the possible six method effects for positively worded items, and one correlated error term between items 15 ("People were unfriendly") and 19 ("I felt that people dislike me"). These latter items reflected feelings of alienation; therefore similarity in substantive meaning justified the additional inclusion of a path among error terms for these items.

Table 9 χ^2 Tests and CFI for Correlated Errors of the SEM in the Calibration Sample

Item Errors Correlated	Minimum fit χ^2	Change in χ^2	Change in df	Significance of Change χ^2 ($p < .05$)	CFI
None	683.8	-	-	-	.94
GOOD/HOPE	674.0	9.8	1	yes	.94
GOOD/HAPP	673.9	.01	1	no	.94
GOOD/ENJO	666.1	7.8	1	yes	.94
HOPE/HAPP	665.8	.03	1	no	.94
HOPE/ENJO	629.9	35.9	1	yes	.95
HAPP/ENJO	539.0	90.9	1	yes	.96
UNFR/DISL	437.0	95.4	1	yes	.97

The baseline model selected for the SEM showed good fit to the data, although the overall χ^2 remained statistically significant at the 5% level for Type I errors ($\chi^2(165) = 455.9$). The completely standardized SEM solution is presented in Figure 1. Each estimate is adjusted by its respective standard error in this solution. The path for item 6 (DEPR) was not estimated, and was set equal to one. However, this item regression is adjusted to set the metric for the model, and it reflects error variance in the latent construct of depression for a standardized solution. Therefore the path for item 6 is expressed as .94. In addition, the variance of the error term for the latent construct for depression was fixed at 1 to allow an estimate for the variance of depression.

Model mis-fit appeared minimal for the baseline model, based on the following global indices. Conventions in the literature have established that RMSEA values around

.05, and CFI or GFI above .90 are associated with good model-data fit (Byrne, 1997). The obtained values for this SEM for CES-D responses in the calibration sample were all indicative of good fit: RMSEA = .059, GFI = .97, and CFI = .97. This baseline SEM was used to generate item statistics for invariance tests in random and comparison samples for the later stages of the study.

The estimates presented for the baseline SEM are also reported in Table 10, along with the significance tests for each item regression term. The model appeared to be meaningful based on the size of squared multiple correlations for each item (most ranging from .31 for GOOD to .98 for SHAK; the smallest was .15 for DISL). The error variances obtained were all within a reasonable range; the smallest were associated with SHAK (.02) and GETG (.10). The most highly discriminating items were 3, 10, 18, and 20. These reflected feeling blue, fearful, sad, and without energy, respectively. Weaker items appeared to be items 4, 8, 15, and 19. These items described feeling good, hopeful, or that people were unfriendly, and feeling disliked. Table 10 contains each item-factor regression and the T test statistic to indicate statistical significance.

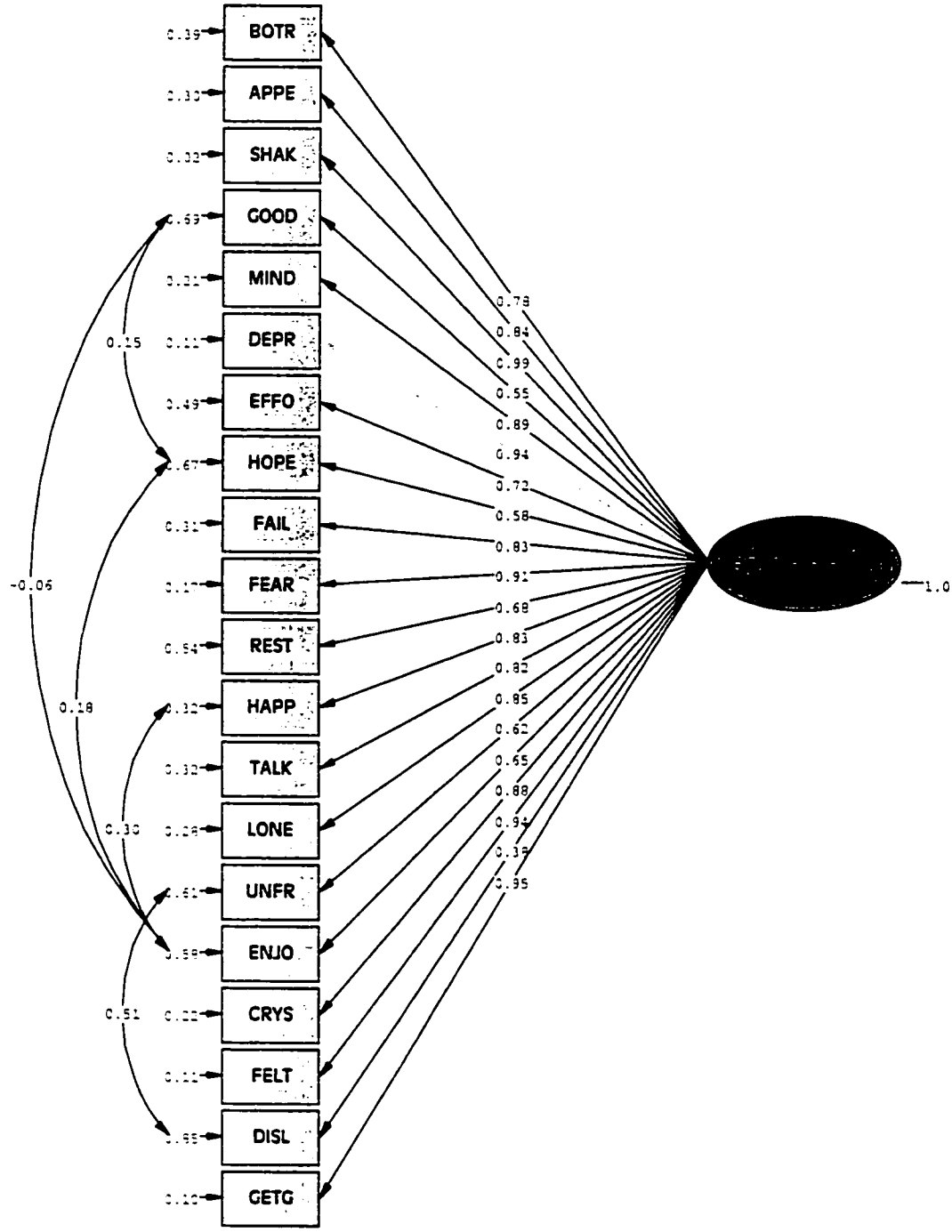
Table 10 SEM Item Regressions and Significance Tests

# Item	Estimate	T Statistic (p)
1. BOTH	.78	28.3 (.03)
2. APPE	.84	26.4 (.03)
3. SHAK	.99	37.6 (.03)
4. GOOD	.55	11.4 (.05)
5. MIND	.89	26.1 (.04)
6. DEPR*	.94	-
7. EFFO	.72	24.5 (.03)
8. HOPE	.58	17.0 (.04)
9. FAIL	.83	23.5 (04)
10. FEAR	.91	26.5 (.04)
11. REST	.68	21.0 (.03)
12. HAPP	.83	27.2 (.03)

13. TALK	.82	24.3 (.04)
14. LONE	.85	32.8 (.03)
15. UNFR	.62	16.0 (.04)
16. ENJO	.65	20.3 (.03)
17. CRYS	.88	27.9 (.03)
18. FELT	.94	41.0 (.02)
19. DISL	.38	8.60 (.05)
20. GETG	.95	32.2 (.03)

*Item 6 was selected as a reference variable for model identification.

Four of the five SEM paths linking item error terms were statistically significant at $p <$ .05. The completely standardized estimates between correlated error terms were .15 for GOOD/HOPE, .16 for ENJO/HOPE, and .30 for ENJO/HAPP. The error term linking alienation items UNFR/DISL was largest of any correlated item terms, 0.51. Only the path between items ENJO/GOOD resulted in a small and non-significant correlation path (-.06). A high and statistically significant variance estimate was obtained for the latent construct of depression (.89, t = 27.0, $p <$. 05). This baseline SEM appeared appropriate for these data and is applied in later SEM tests for the sample-invariance of item statistics in this study.



Fit: Chi-Square=455.94, df=165, P-value=0.00000, RMSEA=0.059

Figure 1 SEM Standardized Item Estimates

Fit of SEM and IRM in Selected Samples

The fit of SEM and IRM to responses from each sub-sample was considered to be an important mediating factor for invariance findings for item statistics. If the IRM or SEM were not correctly specified, item estimates may be expected to lack invariance. Evidence for model fit (or mis-fit) in the random sample condition is presented in Table 11. Both baseline models appeared to fit well in each of the 10 random samples, and there were only small differences in fit across the 10 samples for either SEM or IRM.

A convention has been suggested (Mislevy, 1990) that IRM overall χ^2 values divided by relative degrees of freedom that result in a ratio of 1.5 or less are considered evidence of good model-data fit. IRM modified reliability values are computed from the average standard error of estimated scores (Mislevy & Bock, 1990). This provides an exact estimate of the precision of individual scores across the range of depression, given the specified IRM. All reliability values in these 10 samples exceeded .80, and, therefore, met the criterion for acceptable measurement error in test scores (Crocker & Algina, 1986).

SEM overall minimum fit χ^2 values are interpreted in Table 11 in relative terms. This is due to the caution that χ^2 tests of statistical significance are less useful global indicators of model-data fit in large samples (Byrne, 1997). The overall χ^2 was significant for all 10 samples at the 5% level for Type I error. However, the χ^2 values were small, and CFI was .96 or more in all random samples. These CFI values represented good to very good fit of the SEM. RMSEA values are also small and indicated good fit, ranging from .040 to .075.

A similar examination of fit for SEM and IRM to responses from contrast pairs was conducted and is reported in Table 12. The original 500 cases in any condition are numbered 1, and the cross-validation sample drawn without replacement for the same condition is numbered 2 in this table. The fit of the IRM and SEM in each of the 12 contrast samples met accepted criteria for good fit.

Table 11 IRM and SEM Fit in Random Samples

	IRM			SEM		
Sample	χ^2 (df)[a]	Score Reliability	Ratio χ^2 /df	χ^2 (df=165)	RMSEA	CFI
1	149.3 (117)	.84	1.3	458.9	.059	.97
2	143.8 (117)	.85	1.2	426.7	.056	.97
3	140.5 (117)	.86	1.2	497.2	.064	.97
4	166.8 (110)	.83	1.5	497.2	.064	.96
5	143.6 (116)	.84	1.2	551.3	.069	.96
6	143.4 (111)	.85	1.3	622.9	.075	.96
7	136.2 (111)	.83	1.2	488.9	.063	.97
8	157.6 (116)	.86	1.4	464.0	.060	.98
9	171.0 (117)	.85	1.5	475.6	.061	.97
10	175.5 (111)	.85	1.6	295.1	.040	.99

[a] Degrees of freedom for the IRM depend on the model and on the number of different response patterns in the data.

Overall, the IRM ratio χ^2 /df was 1.5 or less for 9 of the 12 samples, and modified reliability values were all .81 or greater. The SEM χ^2 values were small, ranging from 334.5 to 646.5 with 165 degrees of freedom. The estimated mis-specification

between model and data as summarized by the RMSEA values was also small in each comparison group (from .045 to .080). CFI values were all in the moderate to high range, and consistent with good fit of the model to data from each sub-sample.

Table 12 SEM and IRM Fit in Contrast Samples

	IRM			SEM		
Contrast samples	χ^2 (df)[a]	χ^2 /df	Reliability	χ^2 (df=165)	RMSEA[b]	CFI[c]
Male 1	137 (118)	1.2	.82	411.2	.055	.97
Male 2	192 (106)	1.8	.83	395.9	.058	.98
Female 1	164 (122)	1.3	.87	463.3	.060	.97
Female 2	136 (119)	1.1	.84	441.2	.058	.97
Older 1	157 (120)	1.3	.84	465.3	.060	.96
Older 2	167 (126)	1.3	.85	420.3	.056	.97
Younger 1	139 (109)	1.3	.84	690.5	.080	.96
Younger 2	176 (111)	1.6	.86	646.3	.076	.96
Good health 1	199 (117)	1.7	.84	510.3	.065	.96
Good health 2	152 (101)	1.5	.81	510.3	.065	.97
Poor health 1	169 (130)	1.3	.86	334.5	.045	.96
Poor health 2	139 (128)	1.1	.85	497.2	.064	.96

[a] Degrees of freedom for IRM depend on the model and on the number of different response patterns in the data.
[b] Root mean square error of approximation.
[c] Confirmatory fit index.

Responses from all contrast samples proved to have good fit for both baseline IRM and SEM. Based on this evidence, it appeared reasonable to use estimates from

these models to test the sample invariance of item statistics. Given that the models selected were sufficient for these data, the hypothesis of sample-invariance of item statistics can be tested with minimal confounding due to inappropriate model specification. The next phase of the study concerns summary evidence for the sample-invariance in item statistics based on item statistics from applications of these CTT, SEM, and IRM analyses.

Invariance of CES-D Item Statistics

Two different tests of the sample invariance of item statistics were conducted. First, an intraclass correlation coefficient (ICC) was calculated to summarize the closeness of item estimates from each model across random replications and across contrast groups. Second, SEM statistical tests of measurement invariance were performed using responses from paired contrast groups. The results from random and replication samples based on ICC estimates are presented first. SEM two-group tests of measurement invariance for gender, age, and health comparisons follow.

Random Replications

Invariance estimates based on ICC for each test model across random replications are presented in Figure 2 and in Table 13. These calculations were based on summary data sets of item statistics estimated in 10 random samples. One ICC is computed to describe the consistency of one type of item statistic across samples from any single test model.

The ICC representing absolute agreement between IRM threshold estimates from 10 random samples was estimated to be .988, lower than the level of agreement across

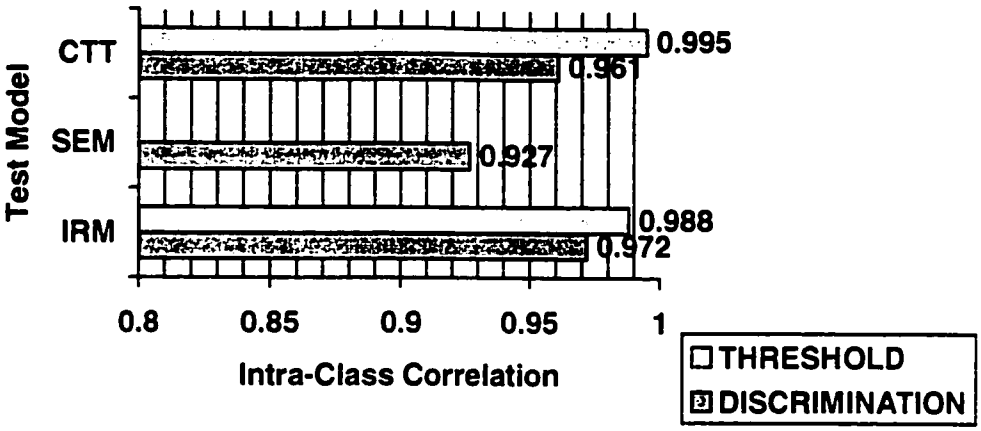


Figure 2 Invariance of Item Statistics from Random Samples

CTT item threshold values, .995. IRM discrimination values, however, were more strongly associated than those based on CTT item statistics. The SEM item discrimination values were associated less strongly at .927. All intraclass correlations were high, positive, and statistically significant at $p < .01$.

Table 13 Intraclass Coefficients for Item Statistics in Random Samples

	Threshold		Discrimination	
Model	Estimate	F value (each $p < .01$)	Estimate	F value (each $p > .01$)
CTT	.995	258.7	.961	26.6
IRM	.988	101.1	.975	40.0
SEM	-	-	.927	14.4

Invariance tests based on CTT item statistics across comparison groups are presented in Table 14. Intraclass correlations were higher in conditions of gender and age contrasts than in health comparisons. The highest invariance estimates (.952) were associated with threshold values across older and younger original samples, and for discrimination values (.934) in the original gender comparisons. The lowest overall

invariance estimates were found for both item statistics in the condition of good versus poor health, with least invariance in the threshold estimate in the cross-validation sample (.566).

Table 14 CTT Sample Invariance of Item Statistics in Contrast Groups

	Original Sample ICC		Cross-Validation Sample ICC	
Contrast Groups	Discrimination	Threshold	Discrimination	Threshold
Male vs. female	.934	.800	.825	.913
Good vs. poor health	.851	.563	.724	.566
Older vs. younger	.663	.952	.855	.912

The ICC estimates for SEM item discrimination statistics are reported in Table 16. These values were lower than ICC values for SEM item discrimination statistics in random samples. They were also lower than the estimates for CTT or IRM statistics in contrast pairs. The lowest ICC values for invariance of SEM discrimination were found across gender and age groups (both about .64). The invariance estimates in cross-validation samples differed significantly from those found in original samples.

Table 15 includes ICC for IRM item estimates in each contrast pair. The ICC estimates in cross-validation samples showed the same pattern as found using the original sample pairs. Item threshold estimates were least invariant in good versus poor health samples, compared with any other contrast condition. The ICC for discrimination values in this condition, however, was high. This may be due to real differences in depression when good and poor health samples are compared; the items appear to have similar accuracy in differentiating people with high or low depression. There was no clear overall

pattern whereby discrimination or threshold values appeared to have greater invariance. The type of contrast pairing (gender or age or health) seemed to provide the widest range in invariance estimates for IRM item statistics.

The ICC estimates for SEM item discrimination statistics are reported in Table 16. These values were lower than ICC values for SEM item discrimination statistics in random samples. They were also lower than the estimates for CTT or IRM statistics in contrast pairs. The lowest ICC values for invariance of SEM discrimination were found across gender and age groups (both about .64). The invariance estimates in cross-validation samples differed significantly from those found in original samples.

Table 15 IRM Sample Invariance of Item Statistics in Contrast Groups

	Original Sample ICC		Cross-validation Sample ICC	
Contrast Group	Discrimination	Threshold	Discrimination	Threshold
Male vs. female	.878	.872	.869	.898
Good vs. poor health	.892	.697	.908	.716
Older vs. younger	.860	.931	.879	.933

Table 16 SEM Sample Invariance of Item Discrimination by Contrast Groups

Contrast Group	Original samples ICC	Cross-validation ICC
Male vs. female	.670	.636
Good vs. poor health	.805	.680
Older vs. younger	.640	.690

In the condition of good versus poor health, the Fisher transformation of independent correlations (based on the common formula presented in Howell, 1982, p.198), resulted in a statistically significant ICC difference of $z = 10.5$ ($p < .05$). Therefore, the lack of SEM invariance in original and cross-validation samples for the same contrast pairing made it difficult to isolate invariance attributed to group membership based on the ICC estimate. The ICC estimates representing the closeness of item discrimination values across contrast samples from all three test models are presented together in Figure 3. In the item threshold estimates based on IRM and CTT test models are presented together. Statistical tests of hypothesized SEM invariance across original contrast pairs are reported in the next section.

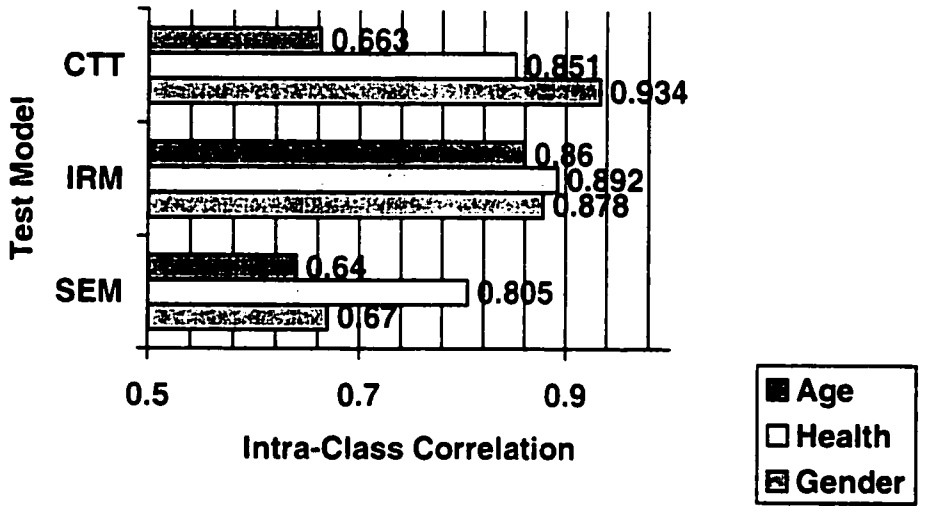


Figure 3 Invariance of Item Discrimination Statistics From Contrast Samples

SEM Two Group Invariance Tests

Multiple-group SEM analyses were conducted as a statistical test of the equivalence of measurement structure across each contrast pair. These invariance tests were hierarchical; each successive run constrained one component of the SEM as invariant. Four separate hypotheses were posed and tested. Each test is based on a significant change in the minimum fit χ^2 statistic (and change in degrees of freedom) between successively more restricted models.

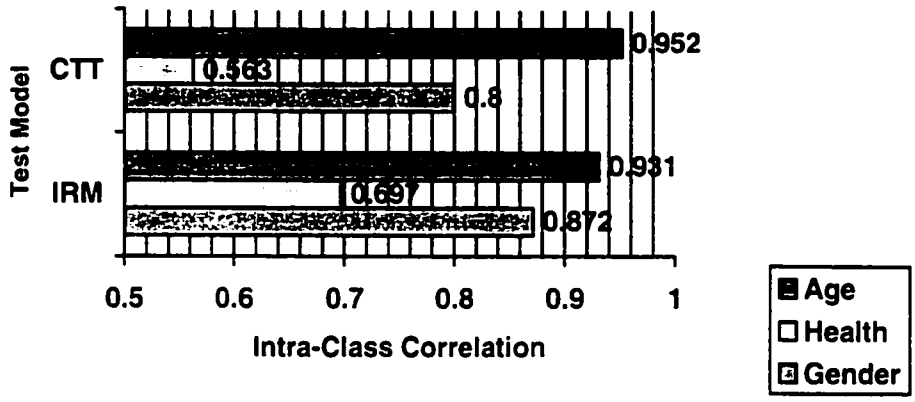


Figure 4 Invariance of Item Threshold Statistics From Contrast Samples

As each new constraint was added to the model, the change in χ^2 was examined for evidence of a significant change (reduction) in model-data fit. Changes in the CFI and GFI were also examined. If the fit was significantly worsened, the added constraint represented a lack of sample invariance in that part of the SEM.

The first (Model A) was used to test the equality of the overall pattern for the SEM across the two groups. For this SEM, the parameters in the second group were freely estimated but had the same pattern as relationships in the baseline SEM. The regression path for item 6 was constrained to one for model identification, and five correlated errors were retained. Therefore, Model A tested that the factor loading matrix, the matrix of error covariances, and the variance of the latent construct in the second group had the same form as that of the first group.

Model B was a test of the hypothesis that estimated factor loadings for items were invariant across groups. Model C tested the invariance of correlated error estimates in the two groups, given equal factor loadings. Model D had the added constraint that the latent construct of depression had equal variance in both groups. This last model represented a

strictly invariant SEM. The results of tests of these four hypotheses in each original contrast pair are described for each comparison group.

All multiple group SEM for gender comparisons showed good overall fit to the data (see Table 17). There was a statistically significant reduction in closeness of fit when item-factor loadings were constrained (Model A compared with Model B). No significant change occurred when the error covariances, or the variance of depression, were constrained to be equal across groups. This lack of significant change in fit indicated that the form of SEM was invariant across gender groups. However, a lack of invariance was obtained in testing the item-factor path estimates. Changes in the GFI of one or more have been considered evidence of a change in model fit associated with lack of invariance (Ferrando, 1996). Based on this convention, and the change in χ^2, the item discrimination estimates do not appear to be equal across gender groups.

Table 17 SEM Invariance Tests of Male and Female Samples

Model	χ^2 (df)	diff χ^2 (df)	χ^2 $p < .05$	CFI	GFI
A	874.5 (330)	-	-	.97	.98
B	950.5 (349)	76.0 (19)	yes	.97	.97
C	964.7 (374)	14.2 (25)	no	.97	.97
D	964.7 (375)	0 (1)	no	.97	.97

Tests of SEM invariance across original groups of older and younger adults are presented in Table 18. All two-group models met criteria for good overall model fit to data from both age groups. However, each successive hypothesized model resulted in a

statistically significant reduction in model fit. There was a reduction in GFI of 1 associated with the stipulation that item regressions on the latent variable were equivalent across older and younger groups. The χ^2 value was significantly inflated with each more stringent invariance model. Therefore, there was evidence that CES-D responses from older and younger samples were not strictly invariant, but may be represented by the same form of SEM.

Table 18 SEM Invariance Tests of Older and Younger Samples

Model	χ^2 (df)	diff χ^2 (df)	χ^2 $p < .05$	CFI	GFI
A	1155.8 (330)	-	-	.95	.97
B	1288.6 (349)	132.8 (19)	yes	.95	.96
C	1351.7 (374)	63.1 (25)	yes	.94	.96
D	1356.3 (375)	5.4 (1)	yes	.94	.96

The results of SEM invariance tests across groups with good and poor health are presented in Table 19. All invariance models resulted in good fit to the SEM. A statistically significant reduction in χ^2 resulted when item-factor regressions were constrained to be equal. There was also a change of 1 in the GFI associated with this new stipulation. There was no significant reduction in SEM fit when correlated error estimates, or the variance for the latent construct representing depression, were equated across groups. Only the regressions of items on the latent factor were not invariant across health groups.

Table 19 SEM Invariance Tests of Samples Reporting Good and Poor Health

Model	χ^2 (df)	diff χ^2 (df)	χ^2 $p < .05$	CFI	GFI
A	844.8 (330)	-	-	.97	.98
B	978.5 (349)	133.7 (19)	yes	.96	.97
C	996.0 (374)	16.5 (25)	no	.96	.97
D	997.0 (375)	1 (1)	no	.96	.97

Individual Scoring Based on Each Test Model

Item response data from the first random sample of 500 were used to generate individual scores based on each test model. SEM factor scores based on the method described by Anderson and Rubin (1956) were compared with optimal theta scores from IRM and with simple linear CTT scores. IRM theta estimates and SEM factor scores were linearly transformed to a metric comparable to that of CTT total scores ($M = 3.7$, $SD = 3.9$). The shape of score distributions, closeness of rankings based on individual scoring and score reliabilities are presented separately in the last section of these results.

Means and standard deviations based on each scoring model were similar due to the choice of IRM and SEM metric transformations. These were selected from the obtained CTT score mean and standard deviation for this sample. Table 20 includes the moments of each score distribution. All score distributions were positively skewed. The IRM score distribution was flatter (with a lower value for kurtosis), and had a smaller positive skew than those based on CTT or SEM scoring. Scores less than one is possible based on both SEM and IRM scoring. The minimum score based on IRM and SEM

scoring resulted in a real number less than zero, whereas the lowest score based on CTT scoring was zero.

Table 20 Individual Score Distributions Based on Each Test Model

Model	Mean (*M*)	*SD*	Minimum	Maximum	Kurtosis	Skew
CTT	3.7	3.9	.00	18.0	1.17	1.30
IRM	3.7	3.5	-.66	13.0	-.56	.48
SEM	3.7	3.9	.69	16.5	1.49	1.60

Score Ranks

The closeness of ranks for individuals based on each scoring model was summarized using a Spearman correlation coefficient. Table 21 includes the correlations between individual total scores from the three models for test scores. The Pearson correlations are provided in parentheses. IRM- and CTT- based scores provided the most similar rankings; the ranks given using SEM factor scoring showed lower agreement with both IRM and CTT score ranks.

Table 21 Correlations Between Individual Scores

Model	IRM	SEM
	Spearman (Pearson)	Spearman (Pearson)
CTT	.992 (.962)	.952 (.886)
SEM	.957 (.859)	

Bivariate relationships between individual scores are plotted in Figures 5 to 7. The plot of IRM and CTT scores showed a strong linear relationship (Figure 5). It is apparent

in Figures 5 and 6 that a range of scores was obtained using IRM scoring when compared with CTT or SEM. People who were assigned the same CTT total score had a broad range of scores using IRM scoring. This pattern was most evident in the lower part of the CTT score scale, and less so in the upper scoring range where data was sparse.

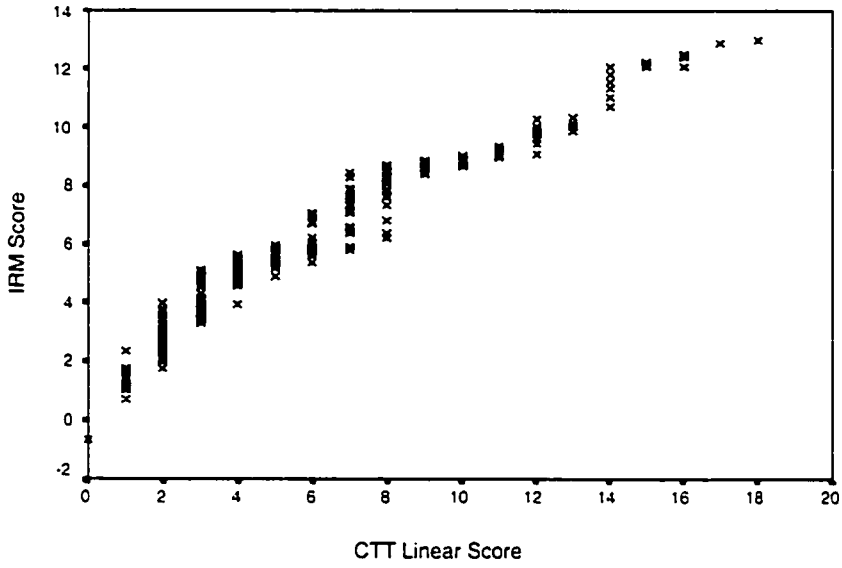


Figure 5 Plot of CTT and IRM Individual Scores in the Calibration Sample

The relationship between CTT and transformed SEM individual scores is apparent in Figure 6. There is a separation in the pattern of relationships for these two total scoring models. A small group of individuals in the total range of CTT scores were assigned substantially higher SEM scores than others in a similar CTT range. Although people were assigned higher SEM scores in these clusters of cases, the linear relationship to CTT scoring is strong (above .95). There was a greater spread in the range of total scores based on SEM scoring compared with CTT scores.

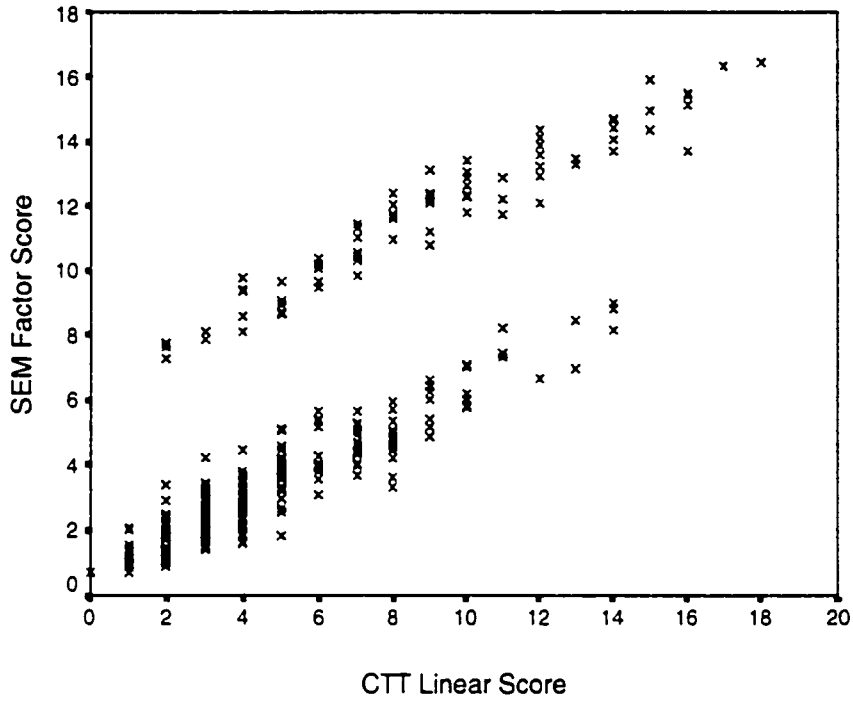


Figure 6 Plot of CTT and SEM Scores in the Calibration Sample

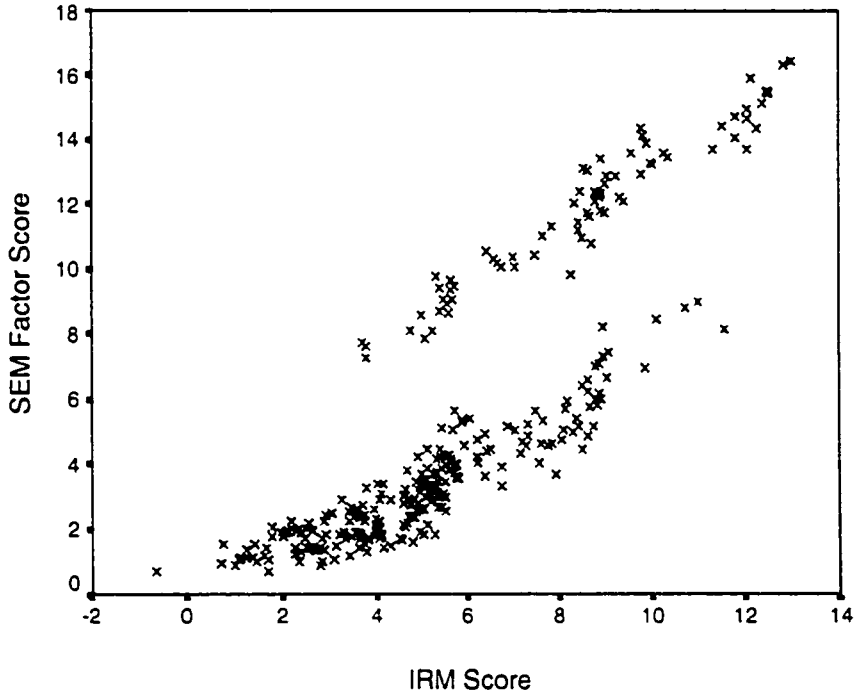


Figure 7 Plot of IRM and SEM Scores in the Calibration Sample

The scatter plot of SEM and IRM total scores is presented in Figure 7. There was also a split pattern in this plot similar to that in Figure 6, but this was apparent for a smaller proportion of cases. This pattern was not evident in the lower range of the IRM scores. A strong linear relationship was evident in both clusters of cases when the score distributions were compared.

Reliability of Scores

Each test model applied in this study is accompanied by an estimate of the measurement error expected for individual scores. Cronbach's alpha is used to summarize the consistency between item responses for CTT reliability and may be considered a lower bound estimate of reliability (Novick & Lewis, 1967). This will be equal to the SEM composite factor score reliability only when the items have equal discrimination in the congeneric model (Ruterberg & Gustaffson, 1992). The SEM composite reliability is calculated as a ratio of the squared sum of factor loadings to that quantity plus the sum of error variances. These estimates are available in the LISREL solution for the measurement model. A modified reliability estimate is calculated based on the standard errors for IRM optimal scores in the range of the sample, given in the program BILOG (Mislevy & Bock, 1990). The CTT, SEM, and IRM total score reliability estimates from the calibration sample are presented in Table 22.

Table 22 Reliability Estimates

Model Estimation Method	Estimate (N=500)
CTT internal consistency	.85
IRM optimal score reliability	.84
SEM composite factor score reliability	.96

The reliability obtained from SEM was higher than either CTT or IRM reliability estimates. This value may be considered with some caution, as only 64% of the variance in observed scores is accounted for with the SEM. With the SEM factor scoring, unequal thresholds or differential accuracy of scores at lower or higher levels of depression are not accounted for. Only with the IRM reliability estimate is precision across different levels of depression in the sample obtained. Alternative methods of determining the precision of total scores from each scoring model that represents the accuracy of CES-D scores with respect to a valid criterion for depression would provide a clearer basis for comparison of models. This would require a true measure of depression and this was not available for the sample in this study. Therefore, these reliability estimates are limited by the parameterization of each model and differential sensitivity to assumptions on the response data. Therefore, these estimates would not be expected to provide equivalent values for score reliability across models.

Post-Hoc Tests

Three important issues were examined to determine the generalizability of these results for response data from criterion-referenced and optimal performance tests that

make use of IRM (or SEM) item statistics. The results of each post-hoc test are considered in the order these were presented in Chapter 3.

First, less discriminating items were eliminated to model a shorter CES-D using item selection rules consistent with criterion-referenced test analysis in achievement. The resulting fit of IRM and SEM to a short form test is presented. The consequences of including poor items in the existing CES-D for sample invariance of item statistics are considered more fully in the discussion section.

Second, responses from people reporting no symptoms were removed (any person who scored zero on the CES-D) from the total sample of EPESE data, and a new sample of 500 was drawn at random. The CES-D may be viewed as a typical performance measure, and this strategy was used in order to determine if truncating the sample to make the distribution of scores more similar to an optimal performance test would result in an improved fit for the IRM and SEM. This new sample could be described as more symptomatic. This pattern of responses to items may be more similar to data collected using a short academic test where optimal performances are measured.

The third post-hoc analysis assesses an alternative method of examining the sample invariance of item statistics where the metric is constant across contrast samples. Only the original sampling from the health contrast condition was selected. This choice was made because the mean depression score in good and poor health groups was expected to differ most, whereas a gender or age contrast would be less likely to have a systematic pattern of different average depression scores. The IRM depression scores were calculated using the 2-parameter logistic IRM with responses from the merged samples of original health groups (N=1000). These theta estimates were then analyzed

using a linear logistic regression to obtain item statistics representing threshold (intercepts) and discrimination values (betas or slopes) for the health groups, separately (each group N=500). Standardized estimates were recorded for each item and a summary correlation was calculated to represent the degree of invariance of these item statistics across samples.

The results of these post-hoc tests are presented as partial evidence for the generalizability of invariance evidence based on the case study to other short achievement or affective tests used in the education setting. An analysis of a short form CES-D after the elimination of some low discrimination items is presented in the next section, followed by an analysis of the symptomatic sample based on responses to the 20-item scale. The third analysis allows extension of the main study results with respect to IRM sample invariance studies where a group calibration is used to generate parameters for each item. Results of the three post-hoc tests are presented in sequence in the following sections.

Short Form CES-D

The potential improvement in IRM and SEM fit obtained by removing some poor items was considered an important factor in this study. Several criteria were used to identify items that might be dropped without reducing the sensitivity of total scores on the CES-D. The item discrimination statistics obtained in earlier analyses, item content, and an index of item discrimination were considered in selecting items for a short form of the scale.

A discrimination index (D) was calculated by comparing the proportion of people who endorsed each item from groups having high or low CES-D total scores. High and

low scoring groups in the calibration sample were defined by the median total score of 2.5, providing a split of the upper and lower 50% of the sample. The proportion of people in the lower total score group who endorsed an item was then subtracted from the proportion of people in the high score group who endorsed the same item. This D formula has been proposed as a valid index for the relative discrimination of items to be used for item trimming using large samples (Crocker & Algina, 1986). Values of D below .30 have been associated with items needing review, and values below .20 are associated with items that merit elimination based on criteria given by Ebel (1965). The proportion of upper and lower groups who endorsed each item and D values are presented in Table 23.

This discrimination index may be compared with the item statistics derived using each test model to provide more information to guide item elimination. There was substantial agreement between low D values and item discrimination statistics derived earlier in this study (consider Table 6, Table 7, and Table 10). The D indices identified 9 items with values lower than .30, of these, there were 4 items with D values lower than .20.

The item statistics from CTT, IRM, and SEM indicated that 5 items (4, 8, 11, 15, and 19) had low discrimination values. Corrected item-total correlations for these items ranged from .21 to .39. These five items were also the weakest CES-D items based on SEM and IRM discrimination values. Standardized item regression coefficients on the latent factor for depression ranged from .38 to .68. The IRM discrimination values for these items ranged between .52 and .71. This combined evidence indicated that these

items had limited value in distinguishing between people with high and low levels of depression.

The content of items with low discrimination estimates and low D index values covered a broad range. It was considered important that any short form of the scale should retain the primary constructs covered by the 20-item CES-D. Four of the 9 items with D values less than .30 described low affect (items 3, 9, 17, and 19) and were judged important inclusions. Only 3 of the remaining 5 items with low D values appeared weak based on IRM, CTT, and SEM discrimination values, and these were eliminated (items 4, 13, and 19). Two additional items were removed from the scale based on low SEM and IRM discrimination values (8 and 15). Therefore, the 5 items removed from the scale included two positive affects (HOPE and GOOD), restless sleep (REST), and two statements reflecting feelings of alienation (UNFR and DISL). The short form CES-D retained 15 items. The same items had been eliminated based on a clinical sample where criterion validity of a shorter form of this scale was tested (Santor & Coyne, 1997).

Table 23 Index of Item Discrimination

#	Item Name	% Upper	% Lower	D Index
1.	BOTR	.36	.03	.33
2.	APPE	.35	.05	.30
3.	SHAK	.30	.01	.29
4.	GOOD	.15	.02	.13
5.	MIND	.38	.05	.33
6.	DEPR	.48	.05	.43
7.	EFFO	.47	.11	.36
8.	HOPE	.46	.14	.32
9.	FAIL	.18	.01	.17
10.	FEAR	.26	.02	.24
11.	REST	.44	.10	.34
12.	HAPP	.37	.03	.34
13.	TALK	.30	.03	.27
14.	LONE	.47	.05	.42
15.	UNFR	.22	.02	.20
16.	ENJO	.26	.02	.24
17.	CRYS	.16	.04	.12
18.	FELT	.53	.04	.49
19.	DISL	.11	.01	.10
20.	GETG	.38	.04	.34

IRM and SEM Analysis

An identical IRM was fitted to response data for the shortened 15- item CES-D. The IRM fit all items well based on an examination of residuals for each item. Chi-square values are not useful indicators of item-IRM fit for tests of less than 20 items. Mislevy and Bock (1990, p 2-2) recommend that root mean square standardized residuals (RMSR) for items be less than the absolute value of 2.5. All items in the short form of the CES-D met this criterion.

The rescaled item statistics were very similar to those found in the calibration sample for responses to the original CES-D. Item slopes in the short test had a mean value of M =.25, and thresholds of M = 9.07. Mean item slope from the 20 item CES-D was M =.23, and threshold M = 9.60. IRM reliability values were also similar, .82 for the short form compared with .84 for the original CES-D. The ratio of these estimates gives a relative efficiency value of .97, confirming that little precision was lost. This can be interpreted to mean that IRM scores retained acceptable precision with fewer items, and that a lower average level of depression would be reliably measured using the short form of the scale.

The SEM of the short form CES-D showed some improvement in fit over the 20-item form. This model retained the path representing correlated errors for two positive affect items HAPP/ENJO and was in other respects identical to the baseline SEM applied earlier in this study to derive item statistics. Indices of model fit included $\chi^2(89) = 135.1$, RMSEA = .032, GFI = .99 and CFI = .99. Standardized regression path coefficients were high for all items, ranging from .61 to .94. All regression paths, the correlated error term, the variance estimates for errors, and the variance for depression were statistically significant at the 5% level. There was a small improvement in the fit of IRM and SEM for the short version of the CES-D. The possible benefit of trimming CES-D items and implications for score validity and sample-invariance of item statistics warrants comprehensive study.

Symptomatic Sample

In order to examine the potential impact of skewed responses to the CES-D in this sample on SEM and IRM model fit, the sample was restricted to people who endorsed at

least one CES-D item. This smaller symptomatic sample included responses from a total of 4492 people. Five hundred people were then selected at random from the symptomatic sample and their responses were analyzed with the baseline IRM and SEM. The mean and median CES-D scores were higher in this group (M = 4.98, *Mdn* 4) compared with the calibration sample (M = 3.74, *Mdn* 2.5). The symptomatic group score distribution was less skewed (1.04) compared with the calibration sample (1.30) and variances were equal.

IRM and SEM Analysis.

Response data from the symptomatic sample were analyzed using the same IRM applied in this study for the 20-item CES-D. All items fit the IRM well, based on χ^2 tests of observed and predicted responses to items (none were statistically significant). Also, the ratio of overall χ^2 to degrees of freedom was small (1.13). In the calibration sample, three items had evidence of mis-fit to the IRM, and the overall χ^2 was larger relative to degrees of freedom. This evidence indicated that the IRM fit data from the symptomatic sample more closely than data from the random sample used for calibration in this study.

IRM item estimates and average standard deviations in the symptomatic sample were similar to those in the calibration sample. The mean item discrimination was .23 for the calibration sample, compared with .17 in the symptomatic samples. This finding suggests that item invariance findings would be similar when tested in a more symptomatic sample. The mean thresholds were 9.60 and 8.74, in the calibration and symptomatic samples, respectively. The actual values of rescaled item thresholds were very close. All but three of the 20 CES-D items in the symptomatic sample received threshold estimates approximately one point lower on the observed score scale. Exceptions were items FAIL, HAPP, and ENJO, where thresholds were higher than the

calibration estimates. IRM reliability for the symptomatic sample was slightly lower, .82 compared with .84 in the original sample.

The SEM for the 20-item CES-D was also fitted to these data representing a symptomatic sample. Fit indices were similar ($\chi^2(165) = 424.2$, RMSEA = .056, GFI = .97, and CFI = .96) in the symptomatic sample compared to the calibration sample. The RMSEA in the calibration sample was larger (.059), whereas other indices indicated slightly better fit when compared to the symptomatic sample (e.g., CFI=.97 in the calibration sample). The item regressions on the latent factor from the standardized solution for symptomatic data were generally lower than those in the calibration sample. The item HOPE had an extremely small regression estimate (.09) compared to that in the original data (.58). This suggested that some items (e.g., positive affect) may be less useful in a more depressed sample, although only minor differences were achieved in overall fit of the SEM to the new data. The sensitivity of SEM invariance tests in contrast groups where data are more normally distributed merits investigation in future research.

Group Calibration for IRM Invariance Tests

Correlation estimates of item statistics from a logistic regression of theta values based on group IRM scoring are presented in Table 24. The original health contrast samples were examined using this alternative methodology because these groups were expected to differ most in mean levels of depression. Therefore, this sampling condition would be most likely to result in inflated ICC values when item statistics were obtained separately to examine sample invariance.

The ICC correlations based on this method were substantially lower than was found in original samples contrasts among health conditions based on separately

calibrated IRMs. The Pearson and ICCs for item thresholds were small negative values. The values reported for Spearman correlations were closer to the values found using the original study methods of deriving IRM item statistics (in the main results the ICC values were .892 for discrimination and .697 for item thresholds). The Spearman values are also lower than the ICCs based on the original method of analysis. However, a similar pattern to the original contrast results for IRM statistics was evident when Spearman correlations were compared; the item thresholds showed the least invariance across good and poor health groups (.420) compared with item discrimination statistics (.749).

Table 24 Correlations of IRM Item Statistics Among Good and Poor Health Groups from a Logistic Regression of Thetas from Combined Original Health Groups

IRM item statistic	ICC correlation	Pearson correlation	Spearman correlation
threshold	-.047	-.079	.420
discrimination	.419	.319	.749

The purpose of deriving these item statistics from thetas, based on a merged calibration of good and poor health groups, was to control for (fix) the metric of item parameters. This was done in order to determine the degree of invariance when mean group differences in depression were taken into consideration. Each of the correlations derived in this post-hoc test are lower compared with ICC values obtained in the same sampling condition using the original study methods. It was possible that sparseness in the data may have been a factor in the range of item statistics that were derived in this post-hoc test; some item statistics may have been estimated imprecisely, and this would

cloud the interpretation of all correlation estimates. The values of each unstandardized item estimate were examined to determine if this computational issue may warrant attention.

The mean and standard deviation obtained in the poor health group were M = 1.65, and $S.D.$ = 2.51 for thresholds, and M = 1.98 and $S.D.$ = 1.29 for discrimination values. The good health sample mean and standard deviation for thresholds were M = 0.92 and $S.D.$ = 0.38, and for discrimination values M = 1.01 and $S.D.$ = .60. These values indicated that the poor health group had a higher overall level of depression compared with those reporting good health, and that the variance in item parameters was greater within the poor health sample, across items. Items with the highest thresholds in the sample reporting poor health were GOOD and DISL. These items represent feeling as good as other people and feeling disliked by others. The item with the highest threshold value in the sample reporting good health was DEPR, representing feelings of depression.

The unstandardized threshold estimates based on the good health group were in the range of -1584 to 7.79, with 13 of 20 items having threshold values less than –1100. The item discrimination values from the good health group ranged from –1.81 to -.51, all negative values. In contrast, the item thresholds from the poor health group were between -8.34 and 8.11, with one exception for the item SHAK receiving a threshold estimate of –1228.5. All item discrimination values were positive in the good health group, ranging from .00 to .61. This broad range in estimates indicates poor estimation in the logistic regression to obtain item statistics, particularly in the good health sample. Implications of this post-hoc test based on a group IRM calibration of thetas and of the preceding post-

hoc tests are discussed with respect to the internal validity of the main study findings in the final chapter.

Chapter 5: Summary and Discussion

This chapter is divided into two components. In the first, the results of the study are summarized with respect to the three research questions and a concise response is offered for each question. This comprises the first three sections of the chapter. The second component is a broad discussion of these results within the framework of relevant literature. This component includes four sections. These four parts are organized to consider some predictions made regarding the sample invariance of item statistics, the impact on individual scoring, limitations of the case study, and implications for research and practice. This chapter concludes with some comments intended to situate the main study findings within existing research from educational measurement.

Responses to the Research Questions

This study was designed to provide an empirical test of the sample invariance of item statistics. Sample invariance is one potential benefit of the item response model (IRM) analysis method. This property allows accurate estimates of items statistics to be made that are not sensitive to the particular sample that provide the response data. This kind of invariance allows for 'person-free item statistics.' The study was not a test of 'item-free person scores,' where individual scores may be compared when people respond to different items. Although there is a close relationship between the invariance of item statistics across samples and issues of item bias or differential item functioning (DIF), this study does not provide a test of DIF. The study was designed to address a lack of empirical examples based on real data for the theoretical superiority of IRM over classical

models for item statistics noted by Fan (1998) and others. A practical application was chosen and some clinical implications and scoring issues were highlighted.

Two conditions of sampling were considered; 10 random replication samples of 500 people, and 12 contrast samples of 500 people who differed in gender, age, or health condition. The case study selected was a large population survey containing responses to the Center for Epidemiologic Studies Depression Scale. The CES-D is a 20 item self-report measure. Items are endorsed if a depressive symptom was experienced during the past week. This short measure has important clinical uses, and yields response data with properties comparable to shorter measures used for screening or affective testing in education.

Three models for test scores were compared for invariance of item statistics among samples. These were classical test theory (CTT), the IRM, and a structural equation model (SEM). The SEM was included because it allows a statistical test of the structure of responses across groups and correlated errors representing method effects due to item wording may be specified. A consistency estimate based on an analysis of variance model used for inter-rater reliability (the intra-class correlation coefficient, ICC) was used to summarize the closeness of item statistics among samples. In addition, the practical impact of each model for individual scoring was examined via a comparison of ranks given to 500 people, and a comparison was made of the reliability estimates from each test model.

The first three sections of this chapter are intended to provide an answer to each research question based on results from the study. For ease of reading, three tables are

included in the summary section where the results of invariance tests are repeated, organized according to size of the ICC.

The original research questions in the thesis were:

1. Is there evidence for the superiority of the IRM over CTT and SEM in random replication samples when item discrimination and item difficulty estimates from binary response data are examined for invariance?
2. Is there evidence for the superiority of the IRM over CTT and SEM across comparison subpopulation groups when item parameters are tested for invariance?
3. Is there a difference in the reliability of scores and the ranking of individuals that is derived from these three test models?

Question 1: The Invariance of Item Statistics Random Samples

The results of this study lend some support to the hypothesized superiority of IRM in random samples. The ICC values for all models indicated that item statistics from any model were very similar when these values are examined across samples (see The ICC values were high, positive, and statistically significant, ranging from .995 for CTT item thresholds values to .927 for SEM item discrimination values. IRM discrimination statistics were more similar across random samples than CTT discrimination statistics, and both showed greater invariance than SEM discrimination values. However, the threshold values based on CTT were more similar than those based on the IRM.

The answer to the first research question would be a qualified affirmative. There is evidence for the superiority of the IRM over SEM and CTT in random replications when item discrimination estimates are examined. However, all test models resulted in a high degree of invariance in this sampling condition, and CTT item thresholds were more

similar across samples than IRM item thresholds. Based on the theoretical formulation of the IRM, one might anticipate that IRM item thresholds would be less sensitive to sampling, compared with CTT item thresholds. This expectation was not supported.

Question 2: The Invariance of Item Statistics in Comparison Samples

Item statistics obtained from substantively meaningful contrast samples were less similar compared with item statistics obtained in random samples. The IRM discrimination values were most similar in all but one contrast condition (original samples, gender contrast), when compared with CTT or SEM. The CTT discrimination values were most similar across conditions, compared to those from the SEM (with one exception in original samples, age contrast). The CTT item discrimination values varied most across health groups compared to other contrast pairings for this item statistic. Generally, IRM discrimination values were more similar across all contrast pairings, when compared with CTT or SEM item discrimination values.

The IRM threshold estimates varied less than CTT item thresholds in four of the six contrast samples. ICC values were higher for CTT item thresholds in one age comparison and in two gender comparison conditions, when compared to IRM values. Item threshold estimates from both test models were more subject to sampling when health categories were compared.

The statistical hypothesis tests of SEM equivalence in contrast groups indicated that the general structure for the covariance of item responses fit well across contrast pairs. There were significant differences in the item regression paths for two contrast conditions (gender and health). The lowest level of equivalence occurred when older and

younger samples were compared; significant differences appeared for item regressions (Lambdas), error variances and covariances (Theta-Deltas), and also in the variance term for the depression factor (Phi). While equivalence in form was present in contrast groups, there was evidence that exact parameter estimates were not strictly equivalent across contrast groups.

The response to the second research question is affirmative with an interesting note. There is evidence for the superiority of IRM over CTT and SEM across comparison samples when item discrimination values are examined and not for item thresholds. The pattern in the original and cross validation data sets of item discrimination results appears clear: IRM item discrimination statistics were most similar across samples, when compared with CTT or SEM. This general observation was shown by statistical tests of item regressions across groups via SEM.

The ICC values representing the similarity of item thresholds did not show a clear pattern based on the choice of test model. The order of ICC values indicates that in some samples CTT analysis results in more similar item threshold estimates; in other pairings, the IRM was least invariant. What is apparent is that the TYPE of contrast pairing is important. Age comparisons led to the most similar item threshold estimates, gender comparisons usually led to greater stability in item thresholds, whereas health contrasts led to the widest discrepancies.

Question 3: The Reliability and Ranks of Individual Scores

Total scores were generated from each test model on a similar metric so that differences in score reliabilities and ranks could be examined. The metric selected specified that SEM factor score and IRM optimal score means and standard deviations

would be similar to those from simple linear CTT scoring. The IRM optimal scores resulted in the most normally shaped distribution, compared with CTT or SEM scores.

The ranks for individuals were very similar across scoring systems. Spearman correlation coefficients were .992 and .957 for IRM with CTT and SEM, respectively; SEM and CTT scores were also strongly correlated at .952. These values were uniformly high. Few individuals would be placed differently with respect to others in the sample based on the total scores obtained with any scoring method. This linear relationship was discernable in scatterplots where scores based on different scoring methods were compared. The weighted scoring from SEM led to greater spread in the upper end of the distribution compared with scores based on CTT or IRM.

Reliability estimates were within acceptable limits. However, the SEM factor score reliability estimate was .96, substantially higher than Cronbach's alpha for CTT scores (.85) or the IRM modified reliability estimate (.84).

A negative response seems reasonable for the third research question. It appears clear that little difference was obtained in the ranking of individuals due to the choice of a model for test scores. Differences found in score reliabilities must be interpreted cautiously, and this issue is discussed in detail below. In spite of the finding that few differences resulted in ranks based on the scoring system, there may be important consequences for individuals who obtained scores in the high range. If SEM factor scoring were applied, a portion of these people would receive relatively higher scores, compared with CTT or IRM scoring where individual rankings would be more similar. Practical differences in the total score reliability may not be obvious, potential impact is illustrated with an example in the next section.

Discussion

This section of Chapter 5 is a presentation of some issues arising these results, and some predictions made based on the literature. Issues arising from tests of item statistics are presented first, followed by issues of reliability and rankings of individual scores. Next, some limitations of the study related to internal and external validity are described. The implications of post-hoc tests are considered in the discussion of internal validity of the study results. The final section of the discussion includes selected implications for research and practice.

Item Statistics in Random Samples

The results of tests of item discrimination across random samples were consistent with the theoretical literature (Crocker & Algina, 1986, Hulin, Drasgow, & Parsons, 1983; Hambleton & Jones, 1993). Specifically, the expectation that IRM item discrimination is less sensitive to sampling when compared with CTT item discrimination estimates was met. However, the prediction that IRM item statistics would be most similar across sampling was only supported by the results of tests of item discrimination values, and less so by results related to item thresholds. Also, the type of contrast pairing had a greater impact on the invariance of item statistics than the choice of test model.

The observed variation of IRM item thresholds and the relative invariance of CTT item thresholds in random samples were also consistent with the literature. Cook et al. (1988) studied large random samples (2,000 to 4,000) of responses to questions on a biology admissions test for high school seniors. They examined small sets of items (29 to 58 in each set) and invariance was examined for common items given in the spring and fall terms. Items with equivalent parameters were defined as parallel. Those authors

concluded "...neither classical test theory nor item response theory is sufficiently robust to provide viable item analysis or equating results when faced with a lack of parallelism" (Cook et al. 1988, p. 43).

Fan (1998) identified invariance for IRM item discrimination, but found that CTT and IRM thresholds were similar across samples. That study made use of large samples of responses from eleventh grade students (1,000 in each group) to the criterion-referenced Texas Assessment of Academic Skills test. The reading and math sub-tests were used; consisting of 48 and 60 items, respectively. There was a notable skew in the distribution of responses, and 20 replication samples were drawn for each comparison condition. Fan notes that ". . . the test score distributions show strong ceiling effects, as is generally the case for minimum competency tests or other criterion-referenced mastery tests...it would be desirable in future studies to replicate the present study using data from norm-referenced testing" (Fan, 1998). The results of that study failed to support the hypothesized superiority of the IRM to produce sample invariant item statistics.

It may be emphasized that the distributional properties of the response data studied by Fan (1998) are similar to those in this case study. In addition, this study of the CES-D provides evidence similar to that found by Fan, whereas the CES-D is a test where norm-referenced interpretations are made. Item thresholds were similar across samples from both CTT and IRM, whereas item discrimination statistics varied less across samples generally. Also, both studies found that IRM discrimination statistics were more robust to sampling compared to CTT item discrimination values. Generally, the results from this case study support and extend the work of Fan (1998).

Item Statistics in Contrast Samples

The lack of invariance of IRM and CTT across contrast groups has been described in earlier research. The work of Oshima and Miller (1990) was based on simulated data samples of 1000 each, where item statistics were modeled on the ACT Assessment Mathematics Usage Test for 40 items. Item threshold values for that study were in a similar range as those from the CES-D 2 parameter IRM analysis, however the items were more highly discriminating. An important similarity, however, is that dimensionality is a factor in invariance testing. Oshima and Miller (1990) give the example of comparisons across cultural sub-groups where language differences add a dimension that may be unique to one group. The authors state: "...IRT methods also tend to indicate substantial deviations from invariance when test data are multidimensional but there is no bias" (Oshima & Miller, 1990, p. 281). Caution is recommended when the response data may be multidimensional, as this complicates interpretations from invariance investigations. These points are equally relevant to our interpretation for the lack of invariance identified for item statistics across contrast groups for the CES-D. Possible explanations for these findings related to group composition and the CES-D are considered in detail below.

Some explanation for differences in the invariance of item statistics with data from specific contrast samples may be offered. The variation in item thresholds that is associated with health may have been related to the fact that some CES-D items describe health states (e.g., item 2, for poor appetite, and item 11, for restless sleep). People who experienced health problems would more frequently endorse these items. These responses may or may not be independent of the level of depression. This finding does not

necessarily mean that there is bias in these items. A complex structure for the construct of depression (of which health is an important feature) may exist. Evidence of the high correlation between factors from the multi-dimensional IRM test using NOHARM supported this assertion.

Social norms for males and females are also a probable reason for differences in response patterns. Stommel et al. (1993), noted this possibility in their examination of gender bias in some CES-D items. It would be reasonable to expect that the items identified by Stommel et al. (1998) reflect social norms for women; women may be more likely to disclose that they experience crying spells, or that they talked less than usual (items 17 and 13). In addition, Stommel et al. (1998) were unable to differentiate gender bias in their response data from real differences in levels of depression. A direct study of bias in CES-D items and possible complexity of the underlying trait would ensure proper interpretation of scale scores from substantially different groups of people.

The lack of invariance in some contrast groups is complicated by the possibility that depression is related to group composition factors (health, gender, or age). A study of bias in CES-D items and total scores would be required to determine the source of the observed lack of invariance across groups. On a more superficial level, the empirical evidence from this study of contrast groups indicates that properties of responses differ for all models for test scores. However, ICC values from the IRM were uniformly high compared with those from CTT and SEM overall. This finding is consistent with recommendations for the use of IRM item statistics to reduce the possible impact of sampling on the selection of items for test development (Hambleton & Jones, 1993).

Generally, IRM item statistics were more similar across contrast groups, while SEM estimates appeared least stable.

Of particular note, SEM discrimination values varied most when compared to any other test model or item statistic. Also, SEM item properties differed most when age groups were compared. This was evident in the ICC invariance estimate (.640), and later in the two-group tests of structure across groups. An important advantage of SEM is the possibility of a statistical test of the location of these differences in the structure across groups.

Two-group SEM invariance tests of the total CES-D measurement model were used to identify differences in item regressions on the latent factor for depression across all comparison groups. However, there was no evidence of statistically significant differences in the general form of SEM structure for inter-item variance-covariances across groups. The greatest discrepancy in particular SEM estimates was obtained when age groups were compared. Not only were the Lambda item regression paths different (based on significant changes in overall fit for successive models), as did the elements in theta-delta and phi matrices. This finding indicates that differences exist in the psychometric properties of response from older and younger adults, and these have an impact on the unique communalities for item pairs and the variance in the latent factor of depression itself. A possible explanation was discussed in the previous section; interaction effects may be present. Groups, such as older samples, are homogeneous in important ways (e.g., more females, or people with poorer health). These factors, and the fact that item thresholds were not estimated, may have led to differences in the item discrimination estimates obtained via SEM.

Individual Scoring

Differences were identified in the invariance properties of item statistics across sampling for each test model. It was logical to ask, then, whether differences would occur at the level of individual scores when test models were compared. The practical impact of each scoring model was examined via a comparison of ranks and total score reliabilities. The lack of substantial differences in ranks given to individuals using these three scoring models is an important finding. Evidence from the literature for sample invariance of IRM item statistics has been used to argue for the superiority of IRM as a scoring model (e.g., Hambleton & Jones, 1993). In this case, little practical difference in rank seems to result from IRM scoring of the CES-D, at least in a random sample.

The largest variation in ranking was present when SEM factor scoring was compared with CTT scores, and this appeared most often for individuals with higher levels of depression. Clinical implications remain and are discussed in the next section (page 138). It is not clear if these results can be interpreted as evidence for reduced bias in IRM scoring. An alternative explanation may be the simple impact of weighted items for individuals who endorsed more items overall. Only minor differences would be expected when IRM or SEM scoring is used to make normative decisions compared with linear CTT scoring of the CES-D. However, the differential reliabilities of items (discrimination values) are consistent with the recommendation for a weighted scoring model.

An unexpected finding was the high SEM reliability estimate compared with the reliability estimates for IRM and CTT total scores. The factor score SEM reliability estimate was substantially higher than the Cronbach's alpha value for CTT scores, or the IRM modified reliability estimate. Increased total score reliability will result when items

have unequal discrimination when a weighted score is computed. One weighted scoring model (SEM) represented increased reliability compared with the CTT scores. However, the CTT and IRM resulted in very close estimates, and the SEM a much higher estimate. This would seem to indicate that SEM factor scores contain a smaller proportion of error. However, some additional criteria are important when evaluating the meaning of these three reliability estimates.

The lack of invariance identified for SEM discrimination estimates causes concern, and the possible complication of failing to account for the differential thresholds of items needs to be considered along with estimates for score precision. There would be less confidence in arriving at an unbiased SEM factor score, as it is evident that response data from groups who differ in gender, age, and health will have different item regressions, and therefore, different factor score weights for responses. The obtained reliability estimate (from a random sample) may not be similar to one from a sample of women or only younger people. It would be important to obtain a separate reliability estimate to cross-validate the SEM factor score reliability estimate obtained in this study.

Cronbach's alpha is known to provide a lower bound estimate of reliability and will not be comparable to reliability estimates based on weighted scoring when item regressions differ (Carmines & Zeller, 1979). This rule is important in interpreting reliability values from this study, as a range of item discrimination estimates was obtained in each phase of the analysis, for any test model. The obtained SEM factor score reliability estimate from this random sample may be misleading because it represents a generic value. Neither the SEM, nor the CTT reliability estimates for total scores

accounts for differences in the level of the latent trait. That means the SEM reliability estimate may not hold for people with different levels of depression.

Only the IRM analysis provides a standard error estimate for each level of the latent trait represented in the sample, where the modified reliability estimate is based on a mathematical average across the range of scores in the sample (Mislevy & Bock, 1990). Taking into consideration these points of concern, it appears reasonable to conclude that the IRM modified reliability estimate is closest to an exact measure of the error in total scores, compared with the approximation methods implemented via CTT or SEM. A study of score reliability and validity based on alternative models for test scores would be an important next step for future research on the CES-D. There are a variety of alternative methods of determining reliability that could be applied to CES-D scores to test the importance of a wider variety of item features and testing situations (e.g., G-theory analysis).

Limitations of the Case Study

Conclusions from this case study are accompanied with some notes of caution. These concerns are grouped under topics of internal and external validity in this section. Internal validity issues are described first, particularly, the included sample and the CES-D measure. External validity issues are considered next and provide a scope for the contribution these results provide to educational measurement, and for educators and researchers in health.

Internal Validity

The usual threats to internal validity in a research study may be broad and include complex interaction effects that may contaminate study findings (e.g., maturation of subjects, attrition of subjects, regression effects, sensitization). In a measurement research project that makes use of secondary data sources, such as this case study, most of these potential forms of confounding are explicitly tested in the design or are examined as an essential analysis step. Some common issues of validity will be considered in this section, followed by a few qualifications of the study results.

Sample characteristics were explicitly examined in the analysis section to establish their similarity to the general population of Canadian elderly. Also, the CES-D measure was designed for administration in large population surveys to support epidemiological and policy studies, consistent with the EPESE study design. Therefore, the source of data appears representative of the relevant population for clinical interpretations and forms a valid source of theoretical tests of CES-D item and score properties.

Empirical tests of the appropriateness of each measurement model were conducted, and the selection of statistical methods was carefully justified. In addition, specific rival hypotheses were studied as post-hoc tests. The second post-hoc test was relevant to issues of internal validity; responses from people who did not report any symptoms were eliminated from the sample. Then, the fit of the IRM and SEM to this symptomatic sample was evaluated to determine if the relatively low prevalence of symptoms (fewer items endorsed) was a significant cause of any lack of fit. This test was a response to any concern that low prevalence in 'typical' response data may invalidate

the use of IRM analyses (more frequently applied to 'optimal' performance data). If a lack of fit was related to the prevalence of symptoms, this would weaken the comparison of CTT with the IRM (and SEM). The results of this test indicated only a minimal change in model-data fit. This can be interpreted as evidence for the appropriateness of the IRM (and SEM) for CES-D responses, and as additional support for the findings of sample-invariance in the case study.

The results from the third post-hoc test suggested that an interpretation of group calibration item statistics may lead to very different conclusions. In particular, the expectation that mean differences in depression would increase the lack of invariance of item statistics across health groups was confirmed by tests using thetas and item parameters derived using logistic regression analysis. It is not clear if an ICC coefficient would be preferable to summarize results from this post-hoc analysis. Individual item statistics were often out-of-range based on this method, and this may be one indication of poor estimation. The item estimates from the unstandardized logistic regression are not bounded as IRM estimates are. For example, the BILOG program applies a default range on out-of-range item parameter values that may result from sparse data, where poor estimation and large standard errors can lead to very large or small threshold or discrimination estimates. There is no similar boundary imposed by the traditional logistic regression model.

An examination of the unstandardized estimates for thresholds identified out-of-range values for 13 of the 20 items based on thetas from the good health sample, compared to estimates from the poor health sample. This lends less credibility to the values obtained, and to the correlation values that were derived from using the ICC or

Pearson biserals. The Spearman coefficient is less influenced by this variation, however, and may be a more useful indication of invariance. However, it must be noted that the item estimates based on this post-hoc method were sufficiently unstable in this example to severely limit our interpretation. Unfortunately, this post-hoc test may be insufficient to indicate a serious confound to the original methodology. A test based on high and low depression samples using a valid criterion may be valuable to replicate this group calibration test.

It has been noted that the comparison of the common form of the SEM with the IRM is limited. A SEM without mean structures would be expected to be more sensitive to real differences in the average performance of comparison groups. This raises the issue of differentiating the observed lack of invariance in SEM tests from tests of bias in the responses from comparison groups. The IRM is the only model that allows the theoretical possibility of obtaining sample-free item statistics, and a threshold parameter was incorporated for the IRM in this study.

The results of SEM tests of invariance may also be qualified in view of some evidence for a lack of multivariate normality in these response data. Univariate tests of the distribution of responses to items indicated that there was a skew related to the relatively low prevalence of any single symptom, and this condition may be related to overly sensitive indices of mis-fit of the SEM (e.g., CFI and changes in the minimum fit χ^2 values). These internal validity concerns do not lead to other significant cautions regarding the results of the study. We turn now to broader questions related to the external validity of these results.

External Validity

The empirical findings in this study are relevant for short educational screening tests, to the extent that similar properties exist in the response data for any comparable measure. A review of the literature yielded some examples of educational tests with similar properties (e.g., the QAP). One motivating factor in this thesis was the advice from many researchers (Andrews, 1984; Byrne, 1996; Carmines & Zeller, 1979; Hambleton & Jones, 1991; Muthen, 1985; Thomas & Oliver, 1999) that modern measurement methods (IRM and SEM) seem appropriate for a wide variety of applications.

The high-stakes in educational testing depend on a rigorous item selection methodology for test development. By comparison, some health measures are relatively unexamined and experimental. The CES-D has a long history and has been well examined by health researchers. In addition, a post-hoc test for a short form of the CES-D was included in the study to identify possible improvement in the fit of IRM and SEM fit when weaker items were eliminated. There was no significant change in the fit of the IRM or SEM based on the shorter version of the CES-D; this finding supports the validity of this application of these analysis methods.

Strommel et al. (1998) found a reduction in the differences between mean scores from males and females after five weak items were removed from the scale (items 9, 13, 15, 17, 19). It was unclear whether the reduction in mean gender score differences was a result of reduced bias, or a function of the restriction in range of the total scores that occurs when some items are removed. However, a very small reduction in reliability resulted from the elimination of items (Strommel et al., 1998). This finding is consistent

with the post-hoc test results for the shortened CES-D in the current study. The next section provides some additional discussion of the external validity of these results as implications for research and practice.

Implications for Clinical Research and Practice

Some study findings merit an interpretation specific to clinical and research practice in health. Some practical consequences of the case study are considered in this section. These extend to clinical use of the CES-D and some theoretical considerations related to this case study.

One clinically important result from the study is the unique evidence for adequate reliability provided for the binary scoring method. This item format is common to large survey administrations of the CES-D. Furukawa et al., (1977) postulated the binary format for items was justifiable, as was the Likert item format, but did not test the reliability of this format. Also, IRM evidence of the discrimination and threshold of CES-D items is useful for scale development (or tailored testing). In addition, this study provides a unique SEM test of the equality of structure across gender, age, and health groups for CES-D responses. Information regarding the stability of the constructs and item properties may be used as a guide for later research on depression in the population. Specifically, more confident comparisons could be made using scores across gender or age groups, whereas health factors may present an interaction effect related to complex structure in CES-D responses.

Individual Score Interpretations

The choice of a test model for individual scoring may be important, even though ranks were found to be similar in this study. This important issue deserves some

elaboration, and can be illustrated with a simple example. Consider a case where John endorses two items, and Bill endorses five items. If a simple sum of endorsed items is taken (CTT scoring), it would appear that John has scored below the group mean and Bill is above the mean for the random sample in this study. When weights are applied to the item responses (IRM or SEM scoring), the exact items that were endorsed becomes a much more important element in total score interpretations.

It is probable in this scenario that the items Bill endorsed may be better indicators of higher levels of depression. Bill selected more items, may therefore be experiencing more symptoms, and items he endorses may be those that reflect higher levels of depression well (e.g., item 6 describing feeling depressed). By comparison, John has endorsed two items and these are likely items that are related to lower levels of depression and may be less discriminating (e.g., item 11, describing restless sleep). As a function of the weights in an IRM or SEM framework, highly discriminating items will result in relatively higher scores than those from CTT scoring. (The IRM score will also take into account the thresholds of those items). Bill may receive a higher score relative to John when IRM or SEM scores were calculated than he would based on CTT scores. Any improvement in the validity of the weighted scoring method would require a study of interpretations made in a real application. In this case study, impact on normative decisions would be minimal (as indicated by the high Spearman coefficients when ranks are compared).

In a criterion-referenced interpretation of CES-D scores, there may also be important consequences that result from weighted scoring. For example, different classification decisions would result for individuals when a weighted (versus un-

weighted) score is used to translate the cut-score on the domain score of the latent variable. The potential impact might be substantial for individuals referred to intervention services, or for assessment of depression. The proportion of people in a population who may be classified as depressed would also depend on the scoring model. Important social and economic consequences exist for classification decisions when they are viewed as outcomes. This information is used to set health care policy, determine institutional funding, and to evaluate existing health programs.

Item Quality and Dimensionality of Responses

The results of this study include a large quantity of information related to item quality in the CES-D. Method effects, item discrimination values and the underlying structure of the response data are interrelated issues and weave through the next few paragraphs. These topics will be described with reference to existing literature in this section.

Some items were strongly associated, including positively worded items and items describing alienation feelings. Positively worded items appeared to have the weakest discrimination values and may merit review or elimination, a suggestion that has been put forward by others (Santor & Coyne, 1997). If some related items are retained, some correlated errors between these will exist. Method effects may be viewed as an extra dimension that can be accounted for in any model, including IRM (e.g., testlets, or content dependent item sets). In this case, the positively worded items appeared to be least precise compared to other items, removing these items may lead to a more efficient CES-D that maintains adequate precision.

Santor and Coyne (1997) identified the same items that were trimmed in this study, and suggested that other items may also merit elimination (social alienation). An analysis of the relative efficiency of a shorter scale from IRM item reliability perspective is one advantage of IRM analysis methods. A shorter scale would be expected to yield more precise scores as weaker items with lower discrimination values have larger errors of measurement. The fit of IRM and SEM to the 15-item scale proposed here resulted in only slightly improved fit an examination of invariance properties of items based on this short form would be an important next step.

A study of bias in items and test scores based on DIF methods from IRM analyses would add to the rationale for trimming CES-D items. The items trimmed in this study were also those that seemed to lack invariance across groups. Callahan and Wollinsky (1994) eliminated seven items (1,5,9,10, 12, 13, and 16) and concluded that gender and race bias in factorial structure was nearly eliminated. In that study, no external criterion for depression was available, so the issue of bias versus real differences in depression remains unanswered.

It may be unrealistic to expect even the "weak" principle of unidimensionality to be met in practice for complex domains. In this case, dimensionality tests such as that implemented in CHIDIM may be overly sensitive. There is support for unidimensionality from CTT and CFA perspective, and this is consistent with Radloff (1977) who recommended the use of a single summary CES-D score. Two-dimensional IRM analysis of these data revealed an alternative structure with highly correlated factors. Complex structure has important implications for IRM score and item interpretations and goes beyond the scope of this study. For some practical purposes, highly correlated

multidimensional structure may be treated as unidimensional (Breithaupt & Gessaroli, 1996).

One interpretation of these results is that there is a relationship between items that lack invariance and the general factorial structure of the scale. Some of the eliminated items shared error variances, consistent with research that proposed these items may belong to a separate construct (an alienation, or interpersonal subscale as described by Strommel et al., 1993).

Conclusions

Evidence from this case study of the sample-invariance of item statistics in random samples supports recommendations in the literature that IRM techniques are superior in the development and calibration phases of test development (Hambleton & Jones, 1993). This application of measurement techniques to a case of applied data is one instance where IRM item parameters varied least across random and contrast samples, compared with CTT and SEM item statistics, in most conditions. These results are an extension of the work of Fan (1998) who identified a need for more research based on real data for properties of the IRM in comparison with CTT models for test scores. Also, the SEM proved a useful tool in identifying the sources of invariance in the response data from different age, gender, and health groups. The SEM analysis allowed for the identification and estimation of error covariances representing method effects due to item wording. An examination of the practical consequences of scoring individuals with a particular model for responses to items is an important activity for score validation (Linn,1990). This study is a first step in examining alternative scoring models for the

CES-D. Some clinical implications that may be useful for researchers and educators concerned with the CES-D were highlighted in the results and discussion chapters.

While some issues raised in this discussion have implications for the CES-D in particular, the case study also meets a need for research on theoretical topics in educational measurement that make use of real data. This case study illustrates some potential benefits of IRM and SEM techniques to reduce sampling bias in item statistics and to generate more precise individual scores. Improvements in test development and scoring practices for widely applied measures, such as this example from health, are warranted. The potential benefit of modern methods in educational measurement is most credible when practical consequences are explored in an applied setting.

References

Anderson, T.W., & Rubin, H. (1956). Statistical inference in factor analysis. Proceedings of the Third Berkeley Symposium, 5. Berkeley CA: University of California Press.

Andrews, F.M. (1984). Construct validity and error components of survey measures: a structural equation modeling approach. Public Opinion Quarterly,48, 409-442.

American Educational Research Association, American Psychological Association & National Council for Measurement in Education (1999). Standards for Educational and Psychological Testing. Washington: AERA.

Becker, D.F., & Forsyth, R.A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. Journal of Educational Measurement, 29 341-354.

Bensen, J. (1987). Detecting item bias in affective scales. Educational and Psychological Measurement, 47 55-67.

Bentler, P.M. (1986). Structural modeling and psychometrica: an historical perspective on growth and achievements. Psychometrika, 51 35-51.

Bentler, P.M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 112 400-404.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structure. Psychological Bulletin, 88 (3) 588-606.

Bock, R.D., & Aitkin, M.A. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46 443-459.

Bock, R.D., & Liberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35 179-196.

Bollen, K.A. (1989). Structural Equations with Latent Variables. NY: Wiley.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. Psychological Bulletin, 110 (2), 305-314.

Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. Educational Measurement: Issues and Practice, 17 (spring) 1-9, 30.

Breithaupt, K., & Gessaroli, M.E. (1996). A comparison of Stout's T and an approximate χ^2 in conditions of complex structure and pseudo-guessing. Paper presentation at the annual meeting of the American Educational Research Association: New York, NY.

Byrne, B.M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. Multivariate Behavioral Research, 29, 289-311.

Byrne, B.M. (1997). Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. NY: Springer-Verlag.

Callahan, C.M., & Wolinsky, F.D. (1994). The effect of gender and race on the measurement properties of the CES-D in older adults. Medical Care, 32 341-356.

Carmines, E.G., & Zeller, R.A. (1979). Reliability and Validity Assessment. Sage University Paper Series: Quantitative Applications in the Social Sciences. Newbury Park: Sage.

Chapleski, E.E., Lamphere, J.K., Kaczynski, R. Lichtenberg, P.A., & Dwyer, J.W., (1977). Structure of a depression measure among American Indian Elders: Confirmatory factor analysis of the CES-D scale. Research on Aging, 19, (4) 462-485.

Chinn, S. (1990). The assessment of methods of measurement. Statistics in Medicine, 9 351-362.

Cook, L.L., Eignor, D.R., & Taft, H.L., (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25, (1) 31-45.

Crocker, L., & Algina, J. (1986). Introduction to Modern and Classical Test Theory. FL: Holt Rinehart & Winston.

Cronbach, L.J. (1988). Validity. In H. Wainer, & H.I. Braun (Eds.), Test Validity (pp3-18). NJ: Lawrence Erlbaum & Associates.

Cudek, R. (1989). Analysis of the correlation matrix using covariance structure models. Psychological Bulletin, 88 317-327.

Dali, S. (1942). The Secret Life of Salvador Dali. New York: Dial Press.

De Champlain, A.,F., & Tang, K.L. (in press). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. Educational and Psychological Measurement.

Dershem, L.D., Patsiorkovski, V.V., & O'Brien, D.J. (1996). The use of the CES-D for measuring symptoms of depression in three rural Russian villages. Social Indicators Research, 39 89-108.

Ebel, R.L. (1965). Measuring Educational Achievement. Englewood Cliffs, N.J.: Prentice-Hall.

Engelhard, G. Jr. (1992). Historical views of invariance: evidence from measurement theories of Thorndike, Thurstone, and Rasch. Educational and Psychological Measurement, 52 275-291.

Fan, X. (1998). Item response theory and classical test theory: and empirical comparison of their item/person statistics. Journal of Educational and Psychological Measurement, 58, (3) 357-381.

Fava, G. (1983). Assessing depressive symptoms across cultures: Italian validation of the CES-D self-rating scale. Journal of Clinical Psychology, 39 (2) 249-251.

Fechner-Bates, S., Coyne, J.C., & Schwenk, T.L. (1994). The relationship of self-reported distress to depressive disorders and other psychopathology. Journal of Consulting and Clinical Psychology, 62, (3) 550-559.

Ferrando, P.J. (1999). Likert scaling using continuous, censored, and graded response models: effects on criterion-related validity. Applied Psychological Measurement, 23 (2) 161-175.

Ferrando, P.J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodel. Multivariate Behavioral Research, 31 (4) 419-439.

Fraser, C., & McDonald R.P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Furukawa, T., Anraku, K., Hiroe, T., Takahashi, K., & Iida, M. (1997). Screening for depression among first-visit psychiatric patients: comparison of different scoring methods for the Center for Epidemiologic Studies Depression Scale using operating characteristic analyses. Psychiatry and Clinical Neurosciences, 51 71-78.

Furukawa, T., Hirai, T., Kitamura, T., & Takahashi, K. (1997). Application of the Center for Epidemilolgic Studies Depression Scale among first-visit psychiatric patients: a new approach to improve its performance. Journal of Affective Disorders, 46 1-13.

Geisser, M.E., Roth, R.S., & Robinson, M.E.(1997). Assessing depression among persons with chronic pain using the Center for Epidemiologic Studies-Depression Scale and the Beck Depression Inventory: a comparative analysis, The Clinical Journal of Pain, 13 163-170.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. British Journal of Mathematical and Statistical Psychology, 42 139-167.

Goodlad, J. (1979). What Schools Are For. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Gorsuch, R.L. (1983). Factor Analysis, 2nd Edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gumpel, T., & Wilson, M. (1996). Application of a Rasch analysis to the examination of the perception of facial affect among persons with mental retardation. Research in Developmental Disabilities, 17, (2) 161-171.

Hamachek, D. (1989). Psychology in Teaching, Learning, and Growth (4th Ed.). Boston MA: Allyn and Bacon.

Hambleton, R.K. (1984). Criterion-referenced measurement. In Husen & Postlethwaite (Eds.) International Encyclopedia of Education. NY: Pergamon Press.

Hambleton, R.K, Swaminathan, H. & Rogers, J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice. Instructional Topics in Educational Measurement Series (Fall) 38-47.

Hammond, R.L. (1973). Evaluation at the local level. In B.R. Worthen & J.R. Sanders, Educational Evaluation: Theory and practice. Belmont, CA: Wadsworth.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement,9 139-164.

Hertzog, C., Van Alstine, J., Usala, P.D., Hultsch, D.F., & Dixon, R. (1990). Measurement properties of the Center for Epidemiologic Depression Scale (CES-D) in older populations. Psychological Assessment, 2 64-72.

Higgins, N.C., Zumbo, B.D., & Hay, J.L. (1999). Construct validity of attributional style: modeling context-dependent item sets in the attributional style questionnaire. Educational and Psychological Measurement, 59 804-820.

Howell, D.C. (1982). Statistical Methods for Psychology. Boston MA: Duxbury Press.

Hulin, H.L., Drasgow, F., & Parsons, C.K. (1983). Item Response Theory: Application to Psychological Measurement. Holmwood, IL: Dow Jones-Irwin.

Jessen, R.J. (1978). Statistical Survey Techniques. New York, NY: John Wiley & Sons.

Jöreskog, K.G., & Sörbom, D. (1996). PRELIS3.8: User 's Reference Guide. Chicago, IL: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (1996). LISREL8.3: User 's Reference Guide. Chicago, IL: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D., du Toit, S., & du Toit, M. (1999). LISREL 8: New Statistical Features. Chicago, IL: Scientific Software International, Inc.

Kirisci, L., Moss, H.B., & Tarter, R. (1996). Psychometric evaluation of the situational confidence questionnaire in adolescents: fitting a graded item response model. Addictive Behaviors, 21, (3) 303-317.

Knight, R.G., Williams, S., McGee, R., & Olaman, S. (1997). Psychometric properties of the Center for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life, Behavioral Research and Therapy, 35 (4) 373-380.

Knowles, M.S. (1977). A History of the Adult Education Movement in the United States. Huntington, NY: Krieger Publishing.

Lawley, D.N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61, 273-287.

Linn, R.L. (1990). Has item response theory increased the validity of achievement test scores? Applied Measurement in Education, 3, (2) 115-141.

Linn, R.L. (1993). Educational Measurement, 3rd Edition. National Council on Educational Measurement and ACE, Oryx Press; Pheonix, AZ.

Lord. F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13 517-548.

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

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

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, (4) 378-396.

McDonald, R.P. (1985a). Unidimensional and multidimensional models for item response theory. In Weiss, D.J. (Ed.) Proceedings of the 1982 Item Response Theory and Computer Adaptive Testing Conference. Department of Psychology, University of Minnesota.

McDonald, R.P. (1985b). Factor Analysis and Related Methods. Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R.P. (1989). Future directions for item response theory. International Journal of Educational Measurement, 13 (2) 205-220.

McDonald, R.P., & Mok, M.M. (1995). Goodness of fit in item response models. Multivariate Behavioral Research, 30, (1) 23-40.

McDowell, I., & Newell, C. (1996). Measuring Health: a Guide to Rating Scales and Questionnaires (2nd Edition). NY: Oxford University Press.

McHorney, C. (1997). Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Annals of Internal Medicine, 127, (8) part 2, 743-750.

McHorney, C.A., Haley, S.M., & Ware, J.E. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10) II: Comparison of relative precision using Likert and Rasch scoring methods. Journal of Clinical Epidemiology, 50, 451-461.

Mellenburgh, G.J. (1994). A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research, 29, 223-237.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment, Educational Researcher, 18 (2) 5-11.

Messick, S. (1993). Validity in Linn, R.L. (Ed.) Educational Measurement. American Council on Education, Macmillan Inc., NY: 13-103.

Miller, M.D., & Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. Journal of Educational Measurement, 25 205-219.

Mislevy, R.J., & Bock, R.D. (1990). Bilog 3.0 Item Analysis and Test Scoring with Binary Logistic Models (2nd Edition). Scientific Software Inc., Mooreville, IN.

Moss, P.A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practice. Summer, 5-13.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49 (1) 115-132.

Muthen, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. Journal of Educational Statistics, 10 (2) 121-132.

Muthen, B. (1989). Dichotomous factor analysis of symptom data. Sociological Methods and Research, 18, (1) 19-65.

Novick, M., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32 1-13.

Nunnally, J.C. (1978). Psychometric Theory. McGraw-Hill, NY.

Orme, J.G., Reis. J., & Herz, E.J. (1986). Factorial and discriminant validity of the Center for Epidemiological Studies Depression (CES-D) Scale. Journal of Clinical Psychology, 42, (1) 28-33.

Oshima, T.C., & Miller, M.D. (1990). Multidimensionality and IRT-based item invariance indexes: the effect of between-group variation in trait correlation. Journal of Educational Measurement, 27 (3) 273-283.

Radloff, L.S. (1977). The CES-D scale: a self-report depression scale for research in the general population. Applied Psychological Measurement, 1, (3) 385-401.

Radloff, L.S., & Locke, B.Z. (1986). The Community Mental Health Assessment survey and the CES-D scale. In M.M Weissman, J.K. Myers & C.E. Ross (Eds.) Community Surveys of Psychiatric Disorders, (pp 177-189). NJ: Rutgers University Press.

Reise, S.P., & Waller, N.G. (1990). Fitting the two-parameter model to personality data. Applied Psychological Measurement, 14 (1) 45-58.

Rudner, L.M. (1983). A closer look at latent trait parameter invariance. Educational and Psychological Measurement, 43 951-955.

Ruterberg, S., & Gustafsson, J. (1992). Confirmatory factor analysis and reliability: testing measurement model assumptions. Educational and Psychological Measurement, 52 796-811.

Santor, D.A., Zuroff, D.C., Ramsay, J.O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. Psychological Assessment,7, (2) 131-139.

Santor, D.A., & Coyne, J.C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. Psychological Assessment, 9 (3) 233-243.

SAS Institute Inc., (1997). SAS® Procedures Guide, Version 7. Cary, NC. Author.

Schein, R.L., & Koenig, H.G. (1997). The Center for Epidemiological Studies-Depression (CES-D) Scale: assessment of depression in the medically ill elderly. International Journal of Geriatric Psychiatry, 12 436-446.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86 (2) 420-428.

Skinner, C.J., Holt, D., & Smith, T. (1989). Analysis of Complex Surveys. New York: John Wiley & Sons.

Statistics Canada & Health Canada (1996). National Population Health Survey (NPHS94-95): Special Research Initiative, Ottawa: Statistics Canada.

Stevens, J. (1996). Applied Multivariate Statistics for the Social Sciences . Mahwah, NJ: Lawrence Erlbaum Associates.

Streiner, D.L., & Norman, G.R. (1995). Health Measurement Scales: a practical guide to their development and use (2nd Edition). NY: Oxford University Press.

Strommel, M., Given, B., Given, C., Hripsime, A., Kalaian, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Research, 49 239-250.

Spence-Laschinger, H.K. (1992). Intraclass correlations as estimates of interrater reliability in nursing research. Western Journal of Nursing Research, 14 (2) 246-251.

SPSS (1999). SPSS 8.0 User 's Guide. Chicago IL: SPSS Inc.

Tabachnick, B., & Fidell, B. (1989). Using Multivariate Statistics, 2nd Edition, NY: Harper Collins.

Taylor, J., Wallace, R., Ostfeld, A. & Blazer, D. (1988). Established Populations for Epidemiologic Studies of the Elderly, 1981-1993: East Boston, Massachusetts, Iowa and Washington Counties, Iowa, New Haven, Connecticut, and North Central North Carolina. Codebook ICPSR 9915. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Teresi, J.A., Golden, R.R., Cross, P., Gurland, B., Klienman, M. & Wilder, D. (1994). Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. Journal of Clinical Epidemiology, 48, (4) 473-483.

Thissen, D., & Steinberg, L. (1984). Taxonomy of item response models. Psychometrika, 51 (567-578).

Tinsley, H.E., & Tinsley, D. (1987). Uses of factor analysis in counseling psychology research. Journal of Counseling Psychology, 34 (4) 414-424.

Tomas, J.M., & Oliver, A. (1999). Rosenberg's self-esteem scale: two factors or method effects. Structural Equation Modeling, 6 (1), 84-98.

Tyler, R.W. (1950). Basic Principles of Curriculum and Instruction. Chicago, IL: University of Chicago Press.

van der Linden, W.J., & Hambleton, R.K. (1997) Handbook of Modern Item Response Theory. NY: Springer-Verlag.

Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement, 26, (2) 191-208.

Worthen, B.R., & Sanders, J.R. (1987). Educational Evaluation. White Plains, NY: Longman.

Ying, Y. (1988). Depressive symptomology among Chinese-Americans as measured by the CES-D. Journal of Clinical Psychology, 44, (5) 739-746.

Zumbo, B., Pope, G., Watson, J., & Hubley, A. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. Educational and Psychological Measurement, 57 (6) 963-969.

Appendix A: Questionnaire Items for Contrasting Groups

General health

Q26.1 How would you rate your health at the present time?

1=excellent 2=good 3=fair 4=poor, very poor or bad.

Demographic Characteristics

Table 1. Q1.6 Gender m/f

Q2.2 Age 1= < 70 2=70-74 3=75-79 4=80-84 5=85+

Q2.21 Are you married, (legally) separated, or divorced?

1=married 2=separated, divorced or annulled 3=widowed

Q9.1 Which of these income groups represents your own (your husbands/wife 's/spouse 's) personal income for the past month/year?

1= < $5,000 2=$5,000 to 6,999 3=$7,000 to 9,999 4=$10,000 to 14,999

5=$15,000 or more 8=don't know 9=refused

Appendix B: An Approximate χ^2 statistic

The Approximate χ^2 statistic for assessing the dimensionality of a set of examinee responses to test items uses nonlinear factor analysis and the "weak" principle of local independence. Gessaroli & De Champlain (1996) provide a detailed description of the calculation of this Approximate χ^2. In general, the Approximate χ^2 tests the null hypothesis that the off-diagonal elements in a matrix of residual correlations are equal to zero. This examination of residuals can be performed after the fit of a model of specified dimensionality (e.g., *m* dimensions), so that the fit of the *m*-factor model may be evaluated. Computations are available in De Champlain (1992), and in De Champlain and Gessaroli (1994). The five steps in calculating the χ^2 statistic can be summarized as follows:

1. For all pairs of items, determine the proportion of examinees who correctly answered item i, item j, as well as both items. These quantities are referred to as $p_i^{(o)}$, $p_j^{(o)}$, and $p_{ij}^{(o)}$, respectively.
2. Based on the results of the *m*-factor model for all pairs of items determine the expected as well as residual joint-proportions of examinees who correctly answered items i and j. The estimates of the residual joint-proportions are provided by the computer program NOHARM (Fraser, 1988) and are referred to as $p_{ij}^{(r)}$.

3. Calculate the estimated residual correlations ($r_{ij}^{(r)}$) for each pair of dichotomous items with the following formula:

$$r_{ij}^{(r)} = \frac{p_{ij}^{(r)}}{\sqrt{p_i^{(o)}(1 - p_i^{(o)})p_j^{(o)}(1 - p_j^{(o)})}}.$$

4. Transform each of the estimated residual correlations to a Fisher z ($z_{ij}^{(r)}$) using

$$z_{ij}^{(r)} = .5\log_e(1 + r_{ij}^{(r)}) - .5\log_e(1 - r_{ij}^{(r)}).$$

5. Calculate an approximate χ^2 statistic defined as

$$\chi^2 = (N - 3)\sum_{i=2}^{p}\sum_{j=1}^{i-1} z_{ij}^{2(r)},$$

where $z_{ij}^{2(r)}$ is the square of the Fisher z corresponding to the residual correlation between items i and j, (i, j=1, . . . ,p) and N is the number of subjects in the sample. This statistic is approximately distributed as a central χ^2 with df = .5k (k - 1) - t where k is equal to the number of items and t is the total number of independent parameters estimated.

Appendix C: Inter-Item Correlation Matrix

	BOTR	APPE	SHAK	GOOD	MIND	DEPR
BOTR						
APPE	0.50					
SHAK	0.62	0.42				
GOOD	0.37	0.27	0.17			
MIND	0.47	0.36	0.54	0.30		
DEPR	0.55	0.52	0.74	0.24	0.44	
EFFO	0.35	0.33	0.47	0.24	0.41	0.45
HOPE	0.28	0.26	0.24	0.31	0.42	0.36
FAIL	0.37	0.24	0.58	0.25	0.50	0.56
FEAR	0.53	0.38	0.58	0.20	0.45	0.52
REST	0.39	0.29	0.37	0.20	0.41	0.38
HAPP	0.46	0.36	0.48	0.40	0.39	0.56
TALK	0.51	0.45	0.43	0.15	0.39	0.47
LONE	0.37	0.48	0.67	0.28	0.33	0.65
UNFR	0.25	0.30	0.32	0.23	0.26	0.34
ENJO	0.39	0.26	0.43	0.29	0.37	0.49
CRYS	0.41	0.53	0.51	0.35	0.41	0.56
FELT	0.52	0.49	0.62	0.22	0.39	0.79
DISL	0.25	0.10	0.29	0.04	0.17	0.29
GETG	0.45	0.55	0.53	0.09	0.52	0.52

	EFFO	HOPE	FAIL	FEAR	REST	HAPP
EFFO						
HOPE	0.19					
FAIL	0.53	0.40				
FEAR	0.46	0.31	0.41			
REST	0.39	0.34	0.45	0.40		
HAPP	0.41	0.41	0.50	0.48	0.42	
TALK	0.39	0.20	0.53	0.49	0.31	0.56
LONE	0.31	0.24	0.49	0.35	0.22	0.44
UNFR	0.25	0.33	0.35	0.57	0.16	0.35
ENJO	0.33	0.44	0.38	0.33	0.43	0.73
CRYS	0.28	0.41	0.48	0.51	0.50	0.38
FELT	0.30	0.33	0.56	0.55	0.41	0.53
DISL	0.26	0.11	0.35	0.41	-0.07	0.33
GETG	0.59	0.25	0.45	0.50	0.53	0.48

	TALK	LONE	UNFR	ENJO	CRYS	FELT
TALK						
LONE	0.44					
UNFR	0.44	0.40				
ENJO	0.32	0.35	0.36			
CRYS	0.45	0.42	0.24	0.34		
FELT	0.46	0.72	0.42	0.49	0.59	
DISL	0.36	0.34	0.75	0.17	0.14	0.50
GETG	0.50	0.48	0.21	0.48	0.56	0.60

	DISL
DISL	
GETG	0.22

```
*pgm getdata;
*-----------;
%macro less(nn); less&nn=2-(mmms2<&nn); %mend less;
%macro lotsless(lonn,hinn);
  %do nn=&lonn %to &hinn; %less(&nn) %end; %mend lotsless;

*-----------;
data cut.models;
  set csha.link2(keep=id csha1age sex s1status c1status
                      s1_3ms n1_3ms ic
                      vital inst s_status c_status scrn_3ms nurs_3ms
                 rename=(csha1age=age));
  if age>=70 and ((0<=s1_3ms<=100) or (0<=n1_3ms<=100))
       and (vital=2 or (0<=scrn_3ms<=100) or (0<=nurs_3ms<=100))
       and c1status^=3;
  mmms1=s1_3ms; if not (0<=mmms1<=100) then mmms1=n1_3ms;
  mmms1=mmms1/100; age=age/100;
  mmms12=mmms1**2; age2=age**2; mmmsage=mmms1*age;
  mmms2=scrn_3ms; if not (0<=mmms2<=100) then mmms2=nurs_3ms;
  if vital=2 then mmms2=.;
  %less(50) %lotsless(80,96)
  if s_status>9 and vital=1 then vital=2;
  clinref=0;
  if (ic=1 or s1status in (2,3)) and c1status=0 then clinref=1;
  if c_status=1         then cogimp=2; else
  if c_status in (2,3) then cogimp=1; else
  if s_status in (1,2) and mmms2>85
    then cogimp=2; else cogimp=.;
  keep id age age2 sex ic mmms1 mmms12 mmmsage clinref
       vital less50 less80-less96 cogimp; run;

*=================================================================;
data cut.project;
  set csha.link2(keep=id csha1age sex c1status inst educ
                      vital s_status c_status scrn_3ms nurs_3ms);
  if vital=1 and ((1<=s_status<=6 and c_status^=3)
                   or c_status in (1,2,10,11,12))
     and ((0<=scrn_3ms<=100) or (0<=nurs_3ms<=100))
     and inst^=8 and 0<=educ<=40;
  age=csha1age+5; if inst=0 then inst=1; if educ>20 then educ=20;
  mmms=scrn_3ms; if not (0<=mmms<=100) then mmms=nurs_3ms;
  clinref=0; if c_status>3 then clinref=1;
  prevclin=0;
  if c1status in (1,2,3) or c_status in (1,2,3) then prevclin=1;
  keep id age sex mmms inst c_status clinref prevclin educ; run;

*-----------;
title; footnote;
*endpgm getdata;
```