Testing for Sample Invariance of IRM Item Parameter Estimates: A case study on real response data

Krista Breithaupt,
Senior Consultant, Assessment Strategies, Ottawa, Ontario; and
Faculty of Education, University of Ottawa.

Bruno D. Zumbo,
Professor of Mathematics & Psychology,
University of Northern British Columbia, Prince George, B.C.

When citing this paper please use the following format (for APA style):

One important accomplishment in educational measurement has been the specification of item response models (IRM). The primary reason IRM techniques have received wide acceptance is the desirability of a central model feature, namely the theoretical invariance of test and item statistics and individual scores (Engelhard, 1992; Rudner, 1983). This study is an investigation of the sample-invariance of IRM item parameter estimates using a large data set of responses to test items. Sample-invariance of item parameter estimates must be distinguished from item-invariance of person scores; only the closeness of item parameter estimates across different samples of responses to items is considered in this study.

This type of invariance is fundamental to techniques that employ IRMs. For example, IRM item parameters are assumed to have invariance when item statistics are used in the identification of differential item functioning (DIF), item selection for test assembly, test equating, and in calculating optimal scores of individuals (Crocker & Algina, 1986; Hambleton, Swaminathan & Rodgers, 1991; Kolen & Brennan, 1995). However, the sample-invariance of item parameter estimates based on alternative models for responses to test items is seldom examined using real response data (Anderson, 1999; Cook, Eignor & Taft, 1988; Fan, 1998). As the selection of IRM over classical item analysis techniques may depend on the sample-invariance property, an empirical investigation using real response data appears warranted. Two possible IRMs are compared in this study via an examination of item parameter invariance across random and stratified samples using a case study of real response data.

When sample-invariance of item estimates holds, the properties of items derived from the IRM (e.g. item difficulty and discrimination values) are not sensitive to examinee characteristics unrelated to ability (such as gender, or average group performance). Specifically, item difficulty and discrimination estimates based on separate subpopulations will be equivalent up to a linear transformation of scale (Rudner, 1983). The degree of invariance of the IRM is dependent, in part, on the fit of the selected IRM for the data. A well-fitting IRM will result in smaller standard errors of item parameter estimates, and in more similar item parameter estimates across samples (Hambleton et al, 1991). One underlying assumption for the common logistic IRM is that items are determined by a single latent factor (unidimensionality). This condition may not be met in many cases of real response data; particularly tests of complex domains with correlated factors. IRM dimensionality has been examined using a variety of techniques, including linear factor analysis, and harmonic factor analysis based on a model equivalent to the logistic IRM (Hattie, 1984; 1985; De Champlain & Gessaroli, 1996).

The degree of correlation between dimensions in multidimensional response data with complex structure will influence statistical tests of unidimensionality, and also affect the invariance of item parameters across samples (Breithaupt & Gessaroli, 1996; Oshima & Miller, 1990). Clearly, the invariance of IRM item parameter estimates across samples will rely on both the dimensionality of the response data, and on the sufficiency of parameters in the model to adequately describe responses to items. This study was focused on a case of real data to contrast unidimensional 1- and 2-parameter IRMs in the presence of some complex structure. This work also provided an opportunity for an application of modern test theory techniques to an allied discipline where the potential benefits are substantial.

The practical advantages of IRM sample-invariance properties make it possible to generate and compare optimal individual scores across samples, to tailor tests for new samples using calibration information, and generally, to examine score validity via a wider range of item and test statistics. IRM scoring reduces score bias related to group composition, and allows for a comparison of individuals across different tests. Applications of these techniques for the validation of tests in population health have been limited, incomplete, and sometimes incorrect. One form of the IRM dominates in the health measurement field (the Rasch model), and has been used primarily to examine relative item validity. Important issues such as IRM dimensionality, fit and appropriateness have been largely ignored.

**Methods**

The sample-invariance of item parameter estimates from 1- and 2-parameter logistic IRM was examined in a large sample of real data. The response data were expected to exhibit some multidimensionality, as the scale contains items that different aspects of depression. However, a unidimensional treatment of the response data and single interpretation of scores has been recommended by the author of the scale (Radloff, 1977), and a single score is most often reported (McDowell & Newell, 1996).
The dimensionality of the data were explored using an IRM-based dimensionality test appropriate for binary responses based on McDonald's weak principle of essential unidimensionality (McDonald, 1985). This analysis was accomplished with the computer program NOHARM (Fraser & McDonald, 1988). The fit of the unidimensional and 2 factor 2-parameter logistic IRMs were tested via an Approximate Chi-square analysis using the method implemented by the CHIDIM program (De Champlain & Tang, in press). Model fit for the 1 and 2 parameter IRM was examined using a calibration sample of 500. Next, item statistics were calculated for each IRM applied to 10 random sub-groups of 500 people, and also to 3 pairs of contrast groups of 500 people who differ on important characteristics (gender, age and health status). The IRM item statistics and model fit were calculated with Bilog 3.0 (Mislevy & Bock, 1990).

Measure

The idiosyncrasies of real response data provide important evidence for the validity of any technique used to evaluate or score test items. The Center for Epidemiologic Studies Depression (CES-D) scale was selected as a case, partly because depression is an important mediating factor between instruction and learning. The CES-D contains 20 items that have been broadly described as affective, attitudinal and somatic aspects of depression. In large-scale administration (population studies) the items are scored using on a binary response format. There has been some evidence for subscale factors for the CES-D to represent categories for item sets (Callahan & Wolinsky 1994; Knight, Williams, McGee & Olaman, 1997; Radloff, 1977). These factors are strongly associated with each other, and recommendations for scoring warn against reporting sub-scores (McDowell & Newell, 1996).

The CES-D has become one of the most popular screening measures used in studies of population and community health internationally (Fechner-Bates, Coyne & Schwenk, 1994). Item qualities derived from responses of individuals with different health, gender or age groupings have often been examined as evidence for the validity of broad application of this scale. Previous studies of scale structure and group bias for the CES-D have not yet applied IRMs, or considered the possible relationship between the structure of response data and findings of invariance across groups.

Population and Samples

Two longitudinal population-based studies provided data for this study. Responses from a total of 6974 cases were available in the Established Populations for Epidemiologic Studies of the Elderly (EPESE), where the 20-item CES-D was administered in the Duke and New Haven cohorts (Taylor, Wallace, Ostfeld, and 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 organization.

Ten random calibration samples representing 7.6 % of the pooled New Haven and Duke samples (500 people in each) were drawn, the first was used to compare the fit of 1- and 2-parameter IRM and to test the fit of a one-dimensional and two-dimensional IRM. Next, all samples were analyzed separately to generate item parameter estimates across random replication samples.

The total sample was then used to draw at random 500 men and 500 women for the examination of the invariance of item parameters across gender groups. Contrasting age and health groups of 500 individuals were then drawn, after replacement of the total sample. The cut-point for older adults was 74 years of age (younger elderly were 65 to 74), while those in the good health category reported good or very good health (poor health was defined as those reporting fair, poor or very poor health). 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 responses. Similar comparisons were made across age and health contrast samples, where re-sampling was also used for cross-validation.

Invariance of Item Estimates

The invariance of item parameters from different samples was estimated via calculation of an intraclass correlation of item estimates across replications, separately for each IRM. Researchers of IRM sample-invariance of item estimates have applied a variety of summary statistics to draw their conclusions. Some of these are not useful for real data, as they depend on indices from simulated data conditions where invariance in the underlying parameters is known (Oshima & Miller, 1990). Others are descriptive, and may be 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 estimate the closeness of IRM item parameters across samples (Cook et al, 1988; Fan, 1998), although, the use of a simple correlation coefficient may not be adequate. For example, a simple measure of linear association may not be appropriate when there is a non-linear relationship between the IRM item estimates, or when the ability estimates lack a fixed scale across groups to be compared. Even when sample-invariance of item estimates is present, item estimates may differ up to a linear transformation of scale (Rudner, 1983). In view of these issues, an alternative method was identified to summarize the closeness of item estimates.

The intraclass correlation coefficient for absolute agreement is an alternative summary statistic for agreement among item statistics. This analysis method was described as a two-way mixed effects model, used to examine rater reliability by Shrout & Fleiss (1979). This approach has been recommended when variance across persons is expected and when raters are representative of any possible raters, but variation between raters is evidence of measurement error (Spence-Laschinger, 1992).

In this study, item estimates derived from different data samples are treated as data from different raters would be in a traditional rater reliability study. Instead of observations that reflect ratings for individuals, we have estimates obtained for items from each data sample. Chinn (1990) states that the intraclass correlation is an appropriate estimate of agreement in cases where the metric for scores is not on an equivalent scale, a concern for IRM estimation. High agreement among item parameter estimates is strong evidence of sample-invariance. This evidence would be reflected by a high, positive and statistically significant intraclass correlation value.

Results

Dimensionality tests of the response data are presented prior to a description of the fit of 1- and 2-parameter IRMs. The intraclass correlation estimates from random and comparison groups are then presented as evidence of the agreement among item parameters across samples for each IRM. The results in random samples are presented initially, followed by the analyses of item parameter estimates from sub-groups defined by age, gender and health status.

Dimensionality

Harmonic factor analysis of item responses showed adequate fit for a unidimensional 2 parameter IRM. Root mean square residual (RMSR) values below .05 are considered evidence of good model-data fit. The estimate of residuals in the off-diagonal covariance matrix of item responses was acceptably small (RMSR=.008).

The Fisher z test was calculated based on a difference Chi-square from 1- and 2-dimensional IRM produced by CHIDIM (De Champlain & Tang, in press). The improvement in fit based on a 2-dimensional IRM was significant at the 5% level (z=3.62), when compared to the unidimensional IRM solution. There was, however, a small reduction in the size of residuals (the 2-dimensional model resulted in RMSR=.007). The rotated oblique factor pattern for the 2 dimensional factor solution did not reveal distinct domains. Two of the 5 items with significant item-factor loadings (> .30) on the second factor had moderate item-factor loadings on the first factor (.20 and .34). The two factors in the multidimensional model were also highly correlated (r = .72).

The unique variances in the unidimensional and the 2-dimensional solutions showed similar patterns across items. The same four items with the largest unique variances identified in the unidimensional IRM analysis were identified again with the 2-dimensional IRM. This evidence suggests that a unidimensional IRM may be reasonable for these item responses although the underlying construct appears complex, representing sub-groups of depression items.

Fit of 1- and 2-Parameter IRM

A calibration sample of 500 randomly selected responses was drawn to compare the fit of the 1- and 2-parameter IRM. A 3-parameter IRM was not considered for this study, although a guessing parameter present in ability tests may be argued for health tests, based on concurrent symptoms. In this study, it was judged that a guessing parameter would not be justified without a criterion measure that provided an alternative substantive rationale for responses (as unrelated to depression). Therefore, only the 1- and 2-parameter IRM were evaluated and applied to the CES-D response data.

The overall model Chi-square for the 2-parameter IRM indicated a superior fit to the data when compared the same statistic based on a 1-parameter IRM. The omnibus test of the 2-parameter IRM based on twice the negative log likelihood was $\chi^2 = 149.3, df=117$; and the 2-parameter IRM resulted in a
larger $\chi^2 = 175.1, \text{df}=111$). A simple difference test of the change in residuals associated with the less restrictive model resulted in $\Delta \chi^2 = 25.1, \text{df}=6$, this exceeded the critical value for chi-square statistic with 6 degrees of freedom at the $p<.01$ level of significance.

The size of item estimates appeared reasonable from both IRM, with slightly larger standard errors around estimates from the more restrictive (1-parameter) model. Item discrimination and difficulty parameters were in a reasonable range for both models. The 1-parameter IRM estimated slopes for all items to be $a = .86$. The average slope for the 2-parameter IRM was $a = .90$ (ranging across items from $a = .52$ to $a = 1.40$).

Item residuals were summarized using a test of observed and predicted response patterns based on each IRM. Chi square statistics with a probability of less than 5% were judged as indicative of poor fit of the specified IRM to any item. The 1-parameter IRM showed poor fit to 5 items, whereas there was poor fit of the 2-parameter IRM for a total of 3 items. In addition, the finding of a wide range of estimated discrimination ($a$) parameters based on the 2-parameter IRM, and on classical item-total correlation statistics, provided support for the 2-parameter IRM.

### Invariance of Item Estimates

Item thresholds were very similar across random samples based on both IRMs. The 1-parameter IRM resulted in an interclass correlation coefficient among threshold estimates of .993 ($F=258, p<.01$). The same calculation based on 2-parameter IRM threshold estimates among random samples was .989. The estimate across discrimination values for all items in the same 10 random samples was .972 based on the 2 parameter IRM. Both invariance estimates from the 2 parameter IRM were statistically significant at $p<.01$.

The intraclass correlation coefficients representing agreement among threshold estimates in comparison samples for the 1 parameter IRM are presented in Table 1. Comparisons across health groups resulted the least agreement between threshold estimates, whereas estimates based on older or younger samples led to the most similar item thresholds. The comparison of data from males versus females led to a high level of agreement in only the cross-validation replication. The estimate of agreement based on original gender samples showed less similarity between threshold values.

#### Table 1: Agreement of 1-parameter item threshold estimates

<table>
<thead>
<tr>
<th>Sample conditions</th>
<th>Original samples (intraclass correlation)</th>
<th>Cross-validation samples (intraclass correlation)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>male vs. female</td>
<td>.780</td>
<td>.924</td>
</tr>
<tr>
<td>good vs. poor health</td>
<td>.659</td>
<td>.707</td>
</tr>
<tr>
<td>older vs. younger</td>
<td>.964</td>
<td>.946</td>
</tr>
</tbody>
</table>

The level of agreement across item estimates from the 2-parameter IRM is summarized by the intraclass correlations reported in Table 2. The invariance estimates from 2 parameter IRM item threshold values were highest in conditions where age groups were compared. Less agreement in threshold values was present among gender groups, and the lowest level of sample-invariance was obtained in the comparison of health groups.

#### Table 2: Agreement of 2-parameter IRM estimates

<table>
<thead>
<tr>
<th>Sample conditions</th>
<th>Original samples (intraclass correlation)</th>
<th>Cross-validation samples (intraclass correlation)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>discrimination</td>
<td>threshold</td>
</tr>
<tr>
<td>male vs. female</td>
<td>.878</td>
<td>.872</td>
</tr>
<tr>
<td>good vs. poor health</td>
<td>.892</td>
<td>.697</td>
</tr>
<tr>
<td>older vs. younger</td>
<td>.860</td>
<td>.931</td>
</tr>
</tbody>
</table>
Discussion

These results suggest that IRM item estimates for CES-D responses are stable across random replications, but that there is some lack of invariance related to underlying age, health and gender differences. The precision of scores from the CES-D may be improved via weighted IRM scoring to reflect different item reliabilities, however, any advantage must be examined against a valid criterion for the latent trait of depression. Item parameter estimates varied most across comparison samples; a formal study of differential item functioning would determine if depression scores were biased when gender, health or age-related subgroups are compared. In this study, results of harmonic factor analysis of CES-D responses suggest that aspects of depression, which appear to represent different dimensions, are highly correlated and perhaps indistinct for the purposes of IRM analyses. A unidimensional IRM for CES-D scores appears justified. This empirical study of the invariance of IRM item statistics across samples of real response data raises some additional issues that merit consideration and further exploration:

- The 1 parameter IRM in random samples led to very close threshold estimates, these were more similar than threshold estimates from the 2pl IRM. The addition of the item discrimination parameter in the model did not lead to a meaningful reduction in the overall RMSR, and standard errors around threshold values were similar for the 1 and 2 IRM. This finding suggests that while the more complex 2 parameter IRM resulted in closer statistical model-data fit, there was little advantage across random samples of real data when sample-invariance of item statistics was examined.

- Invariance estimates from the 2-parameter IRM in contrast conditions were less dependent on group types, all were in a closer range than invariance estimates from the 1-parameter IRM. This suggests that accounting for mean group performance would lead to greater sample-invariance of item parameter estimates. This may be accomplished more effectively via specification of the 2 parameter IRM, compared with the 1 parameter IRM.

- Possible interaction between contrast group findings and invariance findings may be attributed to the complex structure of these data. The impact of gender or health on responses may be an important factor for a sub-set of items (e.g. communication items, or relationship items). These CES-D items may merit review or elimination.

References


