January 2006

A multilevel Bayesian item response theory method for scaling

Henry May
University of Pennsylvania, hmay@gse.upenn.edu

Postprint version. Published in *Journal of Educational and Behavioral Statistics*, Volume 31, Issue 1, 2006, pages 63-79.
Publisher URL: http://jeb.sagepub.com/cgi/reprint/31/1/63

A multilevel Bayesian item response theory method for scaling

Abstract
A new method is presented and implemented for deriving a scale of socioeconomic status (SES) from international survey data using a multilevel Bayesian item response theory (IRT) model. The proposed model incorporates both international anchor items and nation-specific items and is able to (a) produce student family SES scores that are internationally comparable, (b) reduce the influence of irrelevant national differences in culture on the SES scores, and (c) effectively and efficiently deal with the problem of missing data in a manner similar to Rubin's (1987) multiple imputation approach. The results suggest that this model is superior to conventional models in terms of its fit to the data and its ability to use information collected via international surveys.

Keywords
Bayesian estimation, item response theory, missing data, multilevel modeling, socioeconomic status

Comments
Postprint version. Published in *Journal of Educational and Behavioral Statistics*, Volume 31, Issue 1, 2006, pages 63-79.
Publisher URL: http://jeb.sagepub.com/cgi/reprint/31/1/63

Running Head: SCALING SES FROM INTERNATIONAL SURVEYS

A Multilevel Bayesian IRT Method for Scaling Socioeconomic Status in International Studies of Education

Henry May

University of Pennsylvania

Henry May is a Researcher and Statistician at the Consortium for Policy Research in Education (CPRE) and a Research Assistant Professor at the Graduate School of Education of the University of Pennsylvania. His areas of specialization include multilevel modeling, Bayesian methods, and experimental design. Address correspondence to the author at CPRE, 3440 Market Street, Suite 560, Philadelphia, PA 19104; hmay@gse.upenn.edu.

This research was supported in part by a grant (Award Number R215U980021) from the Fund for Improvement in Education (CFDA Number: 84-215U), Office of Reform Assistance and Dissemination, Office of Education Research and Improvement, U.S. Department of Education; by a grant (Grant Number REC-9815112) from the Research on Education, Policy, and Practice Program (NSF 96-138), Division of Research, Evaluation, and Communication, Directorate for Education and Human Resources, National Science Foundation; and by the Center for Research and Evaluation in Social Policy (CRESP) at the Graduate School of Education of the University of Pennsylvania.

Abstract

A new method is presented and implemented for deriving a scale of Socioeconomic Status (SES) from international survey data using a multilevel Bayesian Item Response Theory (IRT) model. The proposed model incorporates both international anchor items and nation specific items, and is able to (a) produce student family SES scores that are internationally comparable, (b) reduce the influence of irrelevant national differences in culture on the SES scores, and (c) effectively and efficiently deal with the problem of missing data in a manner similar to Rubin's (1987) multiple imputation approach. The results suggest that this model is superior to conventional models in terms of its fit to the data and its ability to use information collected via international surveys.

A Multilevel Bayesian IRT Method for Scaling Socioeconomic Status in International Studies of Education

In the literature of education policy research and social science research in general, indicators of socioeconomic status (SES) typically represent any or all of three constructs: educational attainment, occupational status, and income or wealth (Buchmann, 2002; Powers, 1982). When education research involves surveys of students, the manners in which SES indicators are measured and combined vary from one study to the next, and indirect measures (e.g., the number of books in the home) are sometimes used as proxies when traditional information is unavailable. In international studies of education, composite scales of SES have often been derived by using simple averages, counts, or classifications into only a few categories based on survey item responses (Beaton, Martin et al. 1996; Beaton, Mullis et al. 1996; Comber & Keeves, 1973; Elley, 1994; Gorman, Purves, & Degenhart, 1988; Husén, 1967; Martin, Mullis, Gregory, Hoyle, & Shen, 2000; Martin et al. 2000; Mullis et al. 2000; OECD, 2002; Walker, 1976; Westbury & Travers, 1990; Wolf, 1992). The reasons for this probably stem from the difficulties in measuring SES in an international context and the unavailability of methods that deal with these problems. Some of these studies avoid the term SES by using labels such as "family educational resources" or "family wealth." Although this research aligns with the single-factor definition of SES, the methods presented herein are also relevant to these related scales. For the purposes of simplicity, all such scales will be heretofore referred to as scales of SES.

There are two primary problems encountered when measuring students' family SES in multiple nations: missing data and incomparable data. The problem of missing data is likely a result of students' inability or unwillingness to answer certain questions. When asked about their parents' education, occupation, or income, students may not know the answer, they may be

reluctant to answer, or they may be offended by the question (Bradlow, 1994; Bradlow & Zaslavsky, 1999; Rubin, 1987). Parents of students in international studies could be surveyed also, but the added cost would be enormous given the number of students typically sampled in such studies. As a result, the SES information collected in international education studies via student surveys is commonly plagued by a high rate of missing data (Keeves & Saha, 1992).

Whereas missing data are a problem regardless of whether an analysis involves only one nation or multiple nations, the problem of incomparable data is most likely to arise when the intent is a cross-national analysis. Differences in currency valuation, structures of the educational systems, and economic and social culture make it difficult to collect information about SES that represents the same thing in each nation. This is especially true for indirect measures of SES, such as the number of books in the home and other home possession questions. For example, while the majority of families in the United States have air conditioning in their homes, these appliances are uncommon in Europe, primarily as a result of a more temperate climate. A person in the United States is likely to have a greater probability of owning an air conditioner than a person in Europe with the same levels of education, wealth, and occupational status.

The methods presented in this article provide a means by which traditional methods for scaling SES can be extended to allow selected items (e.g., having an air conditioner) to operate differently across nations. The key characteristics of this method involve the use of (a) international “anchor items” that provide the same information in each nation, and (b) “nation-specific” items that operate differently across nations and can even provide information that is specific to a single nation. The new scaling model combines modern item response theory (IRT), multilevel modeling techniques, and Bayesian estimation techniques to produce a scale of SES that has three desirable characteristics. First, the resultant scores are internationally comparable;

that is, students with equivalent scores but different nationalities have the same family SES score relative to an international benchmark. Second, this new method is able to reduce the influence of national differences in cultures that affect survey responses. Lastly, it is able to effectively and efficiently deal with the problem of missing data in a manner similar to Rubin's (1987) multiple imputation approach.

The structure of this article is as follows. First, the survey data used in this research are described in detail in order to provide additional background to the problem. Next, the general method of scaling SES and the extension of the model using "anchor items" and "nation-specific items" is described. Next, the new model is implemented to create a scale of SES from an existing international database and its results are compared to those from a traditional scaling model applied to the same data. Lastly, the implications of the results of this research are discussed, and some limitations of this analysis are addressed.

Data and Sample

This analysis made use of data from the upper grade of Population 2 (i.e., the eighth grade in most nations) from the first Trends in International Mathematics and Science Study (TIMSS-1995). This grade level had the greatest number of participating nations (N = 42) and students (N = 147,505). Students were asked to complete a survey that included questions relevant to two components of socioeconomic status (i.e., educational attainment and wealth). After dropping three nations (Bulgaria, England, and Japan) that did not collect family SES data, and those students in remaining nations who did not answer any of the SES items (1,783 students), the sample used in these analyses consisted of 138,805 students in 39 nations.

The survey items from the student background questionnaire used as measures of components of family SES include a direct measure of parental educational attainment

(separately for mother and father) and indirect measures of family wealth via home possession items. The international version of the parental education question had six response categories from “finished primary school” to “finished university.” Another SES indicator used in every nation was “About how many books are there in your home?” The response categories ranged from "none or very few (0 - 10 books)" to "enough to fill three or more bookcases (more than 200)."

Other home possession items were presented in a consistent format under the question “Do you have any of these items at your home?” Students responded by selecting “yes” or “no” separately for each item in a list of up to 16 items. The first four of these (calculator, computer, study desk, and dictionary) were used in all nations, while the remaining items varied from nation to nation. These optional nation-specific items were selected by TIMSS coordinating groups in each nation in order to improve their relevance to the culture and economic standing of their country.[1]

Most nations elected to use all 12 optional spaces on the questionnaire. Four countries (France, Iran, Kuwait, and Scotland) did not use any nation-specific home possession items to supplement the first four home possession items. Even though the national coordinating groups could select any home possessions to use in this national option, many countries ended up choosing the same or similar items. When similar home possession questions are treated as one item, the pool of 372 nation-specific items reduces to only 113 items. Although many items were used in only one nation, 25 items were used in at least 5 nations, and 9 of these were used in at least 10 nations. The most common item across all nations was videocassette recorder, which was included in the student questionnaires from 25 nations.

Methods

Scaling Models

Item Response Theory (IRT) offers numerous statistical models for scaling data from items with discrete responses (see van der Linden & Hambleton, 1997, for a review of common IRT models). IRT models predict the probability of a particular response for each individual, and the relationship between that probability and the underlying trait is assumed to follow a logistic or probit curve. Using a maximum likelihood IRT model to produce a scale of family SES based upon parental education and home possession items would produce scores for each individual that maximized the probability of observing their particular pattern of items in the home and parental levels of education. Simply put, individuals with higher scores would have better educated parents and more items in the home, and those with lower scores would have less educated parents and fewer items in the home.

The traditional approach to scaling SES used in this analysis is based on a single-level IRT model, where all items are assumed to operate the same way in each nation. More specifically, this model is a standard graded response model (Samejima, 1997) with the threshold parameter (β_k) split into an overall threshold for each item and individual response category parameters (δ_{jk}). This model reduces to the standard 2-parameter logistic model (2PL) for items with only two categories. The mathematical form of the model is:

$$\ln\left(\frac{\Omega_{ijk}}{1-\Omega_{ijk}}\right) = 1.7\left[\alpha_k\left(\theta_i - \beta_k + \delta_{jk}\right)\right]$$

where Ω_{jk} is the cumulative probability of a response by student i in category j or higher on item k, θ_i is the family SES score for student i, α_k is the item discriminating power for item k, β_k is the overall threshold for item k, and δ_{jk} is the category parameter for response category j on item k.

For any one item, this model uses the same estimates of discrimination and threshold for each nation. Hence, this model does not allow for national level variations in item parameters for any single item (*k*) and will heretofore be referred to as the "constrained model."

A plot of the scaling function (with $P(X_{ik})$ on the ordinate and θ on the abscissa) for any one item is called an "item characteristic curve" (ICC). Each curve is "S" shaped and has a positive first derivative (assuming a positive relationship between $P(X_{ik})$ and the latent trait, θ, for any value of θ from $-\infty$ to $+\infty$, and horizontal asymptotes at $P(X_{ik})$ equal to 0 and 1. The α and β parameters determine the shape and location respectively for a particular ICC. The α parameters also indicate the degree to which item response varies with the latent trait (Lord & Novick, 1968) and are related to the item-total score correlations from a conventional item analysis (Lord, 1980, p. 33).

The threshold parameter (β) for a home possession item can be interpreted as the level of family SES required to have a 50% chance of having that item in a student's home. Consequently, items with low thresholds will be found in more students' homes. Items with high thresholds will be found in fewer students' homes. The discrimination parameter (i.e., α) from this IRT model indicates the slope of the ICC and the ability of the item to discriminate between individuals with SES scores just above and below the value of the threshold (β). Ideally, for an international model of family SES, the discrimination parameters should all be appreciable (e.g., $\alpha>.5$) indicating consistently high correlations between the individual items and the SES scale, and the threshold parameters should be evenly distributed throughout the range of SES, indicating consistent quality of measurement throughout the range of SES (e.g., $-3<\beta<3$ if SES is defined as N[0,1]).

The new model proposed here relaxes the constraint that all items operate the same way

in each nation, and, for any item not including seven anchor items (i.e, mothers education, fathers education, number of books in the home, calculator, computer, study desk, and dictionary), the discrimination and threshold parameters are specific to each nation. This "unconstrained model" is the multilevel approach proposed in this research and has the following mathematical form:

$$\ln\left(\frac{\Omega_{hijk}}{1-\Omega_{hijk}}\right) = 1.7\left[\alpha_{hk}\left(\theta_i - \beta_{hk} + \delta_{jk}\right)\right]$$

where Ω_{hijk} is the cumulative probability that student *i* from nation *h* responds in category *j* or higher on item *k*, $\theta_i$is the family SES score for student *i*, α_{hk} is the item discriminating power for item *k* in nation *h*, β_{hk} is the overall threshold for item *k* in nation *h*, δ_{jk} is the category parameter for response category *j* on item *k*.

Because the α and β parameters for the seven anchor items are constrained to be equal across nations, only the optional nation-specific items have nation-specific discrimination and threshold parameters. In effect, all national-level variance in the resultant SES scale is determined by the anchor items, and within nation (i.e., student-level) variation in SES scores is determined by both the anchor items and the nation-specific items. Therefore, this type of IRT model can be described as "multilevel," given its separation of items contributing to national-level and within-nation components of SES. Failure to use any anchor items would produce a model with no linkage between nations, which is equivalent to having a separate model for each nation. This would result in zero national-level variance in SES and equal national mean SES scores, thereby eliminating international comparability. This approach is similar to methods used to detect differential item functioning (DIF) (see Holland & Wainer, 1993); however, this unconstrained model estimates scores in all nations simultaneously, and comparisons of nation-

specific item parameters are to those from the constrained model (i.e., the international estimates), not nation-to-nation comparisons as is the case with DIF analysis.

In the application of this model to TIMSS data, the anchor items are the seven items asked in every nation. Ideally, each anchor item should be a direct measure of SES so that any differences in responses across nations reflect actual differences in SES, not differences in culture or geography. In this example, only the parent's educational attainment items are direct measures. The remaining items are indirect measures that may operate in different ways across nations. However, for the purposes of this illustration, the constraint of equivalent item parameters is relaxed for the national option home possession questions only.

Estimation

The constrained and unconstrained models were estimated via Bayesian Markov Chain Monte Carlo (MCMC) using Gibbs Sampling (Gelfand & Smith, 1990; Geman & Geman, 1984) as implemented in WinBUGS 1.3 (Spiegelhalter, Thomas, & Best, 2000). Bayesian estimation of IRT models using vague priors has been shown to produce results that are similar to those from traditional maximum likelihood estimation (Bradlow, 1994, Bradlow & Zaslavsky, 1999, Fox & Glas, 2001; Kim & Cohen, 1998). For both the constrained and unconstrained models, vague prior distributions for item parameters were defined as Normal(0,1000) with the discrimination parameters constrained to be positive.

Because estimation could not be performed using the full sample due to computing power constraints, estimation of both scaling models was carried out using a calibration sample of 250 students per nation that were selected with probability proportional to the student sampling weight.[2] This reduced the necessary computing resources and time, and eliminated the need for student sampling weights during model estimation. The final calibration sample consisted of

11,700 students in 39 nations.

The Bayesian models used here involve a full joint distribution on all quantities. As such, WinBUGS treats missing values for survey items in this model as additional parameters to be estimated. In other words, any missing responses for presented items (i.e., those that should have been answered) were imputed from the conditional distribution of response categories given the observed data and the relationships expressed by the model.[3] This stochastic imputation of missing values is similar to the multiple imputation technique described by Rubin (1987). Under the assumption that the missing data are missing at random (MAR; Rubin, 1987), the values of θ produced by this Bayesian model are unbiased, and each value of θ_i drawn by the Gibbs sampler is a plausible value[4] of θ_i given the observed pattern of responses for individual *i*. Handling the missing data in this manner improves the reliability of individual family SES scores for students with incomplete survey responses.

The necessary "burn-in" length for the MCMC chain was determined using the method proposed by Gelman and Rubin (1992). Monitoring of three independent chains with over-dispersed initial values showed that convergence occurred very quickly, in fewer than 200 iterations, for most parameters. The remaining parameters reached convergence in fewer than 500 iterations. After this 500 iteration burn-in, an additional 5000 iterations, retaining every fifth iteration to reduce autocorrelation, were carried out to define the sampling distributions of each of the parameters in the model.[5] In accordance with Newton & Raftery (1994), the likelihood for the each model is calculated as the harmonic mean of the marginal likelihood estimates for the 1000 retained iterations of the Gibbs Sampler after the burn-in period.

Density plots for most item parameters were symmetric and unimodal. Approximately 20% of the items were unimodal, but exhibited substantial skew. Therefore, the posterior mode

was preferred over the mean as point estimates for the item parameters. The mode of each posterior distribution for the item parameters was estimated using the Sheather-Jones Plug-In (SJPI) method of kernel density estimation (Sheather & Jones, 1991) as implemented in PROC KDE in SAS 9.1. These modal values are, by definition, the most likely values, and they are often used in Bayesian analysis as an approximation to the maximum likelihood estimates.

Expected A Posteriori (EAP) estimates of individual family SES scores and their standard errors were produced for the full sample of 138,805 students by submitting item parameter estimates produced under each model to scoring algorithms in PARSCALE (Muraki & Bock, 2002).

Estimating Variation in Item Parameters for Nation-Specific Items

For each nation-specific item asked in more than one country, the item characteristic curves (ICC) from the unconstrained model exhibit different degrees of variability across nations. It is helpful to compute a scalar metric of this variation so that comparisons of items can be made to determine the degree to which specific home possession items operate differently across nations. In this analysis, the degree of variation among the nation-specific ICC curves relative to the international ICC curve is calculated via the following formula:

$$D_k = \sqrt{\left[\sum_{n=1}^{N}\int_{-4}^{4}\left(ICC_{kn} - ICC_{k(\text{int})}\right)^2 P(\theta)\,\delta\theta\right] \Big/ N}$$

where ICC_{kn} is the ICC function for item k for nation n, $ICC_{k(\text{int})}$ is the international ICC function for item k, $P(\theta)$ is the probability density function for θ, N is the number of nations asking item k. The solution to each integral was obtained using the QUAD function in PROC IML in SAS 9.1.

This formula is similar to the basic formula for the standard deviation of a series of univariate data points; however, the integral allows calculation of the distance between two curves as opposed to single data points. Hence, D_k represents the "standard integrated distance" of the nation-specific ICC curves relative to the international ICC curve. Large values of D indicate large distances between the ICC curves within the relevant range of SES (i.e., the standard normal distribution). Values obtained using this function lie between 0, which signifies no variation, and 1, which signifies infinite variation. Distance values of 0 and 1 are practically impossible with real data, so a rule of thumb is necessary to identify small, medium and large distances. A simulation study suggested that integrated distances between 0 and .10 are small, those between .10 and .20 are moderate, and those above .20 are large. For example, a distance score between .20 and .30 would occur if the standard deviation of the threshold parameters for 30 parallel ICC curves was greater than 3. Although the maximum possible distance value is 1, the variation in ICC curves necessary to produce a distance score greater than .3 is much larger than the variation in the underlying SES scores, and therefore, would be very unlikely.

Results

Model Fit Comparison

The deviance statistic (i.e., $-2\ln(P(X|\theta))$) for the constrained model was 254,403. For the unconstrained model, the deviance statistic was 246,466. There is a difference of 518 parameters between the constrained and unconstrained models ($(113 - 372) \times 2$). A likelihood-ratio chi-square test comparing this model to the constrained model using the same sample of students yields a chi-square value of 7,937 on 518 degrees of freedom. The p-value for this test is less than 10^{-16}. This provides substantial evidence that the unconstrained model fits the data significantly better than the constrained model. As an alternative, the difference in the Bayesian

Information Criterion (BIC) statistics for the two models adjusts this test for sample size and the difference in the number of parameters estimated. Raftery (1995, equation 20) shows how the BIC difference can be calculated directly from the likelihood test above using the formula $\chi^2 - df(\ln(n))$. The difference in BIC statistics for these two models is 3,085. Raftery (1995) suggests a BIC larger than 10 "very strongly" favors the more complex model. In this case, the BIC difference is enormous, suggesting a far better fit with the unconstrained model.

Item Parameter Comparisons

Table 1 shows item parameter estimates for the seven anchor items from both the constrained and the unconstrained models. There is remarkable similarity in the parameter estimates for these anchor items. All anchor items show reasonably good discrimination (i.e., $\alpha>.50$) and their threshold parameters are distributed throughout the range of the SES distribution (e.g., $-3<\beta<3$). Discrimination parameters for the parental education items are considerably higher under the unconstrained model, suggesting a stronger relationship between those items and the family SES scores from the unconstrained model.

Table 2 shows item parameter estimates for nine home possession items asked in at least 10 nations. Comparing the parameter estimates from the two models reveals a large degree of variation across nations that is hidden by the single point estimate from the constrained model. Although standard errors are not shown in this table (see http://www.gse.upenn.edu/~hmay for an online appendix including point estimates and standard errors for all parameters), the sampling distributions of many of the nation specific estimates have little overlap with the point estimate from the constrained model. This suggests that the differences in point estimates between the two models cannot be attributed to sampling error. In other words, there is evidence that these items operate in different ways in different nations.

For the "Television" home possession item in Table 2, the variation among nations is quite extreme, and in many nations, this item operates relatively far out in the negative tail of the SES distribution (i.e., 8 of the 15 nation specific thresholds are less than -3). This suggests that having a television is too common in these nations to serve as an informative indicator of SES.

Variation in the Nation-Specific Item Characteristic Curves

The standard integrated distance scores computed using the formula presented previously are shown in Table 3 for the 54 nation-specific items that were asked in two or more nations. Fifteen of the items had small distance scores (i.e., less than .10), 27 items had medium distance scores (i.e., between .10 and .20), and 12 items had large distance scores (i.e., greater than .20). Because the nation-specific items in TIMSS were not randomly assigned to nations, the integrated distance scores for these TIMSS items are probably not unbiased. Nine of the twelve items with small distance scores were asked in only two nations, and it is not unreasonable to think that their small distances might be a result of similarities in the cultures for each pair of nations (which also led those nations to select that item in the first place). The fact that no items asked in more than 5 nations have small distances suggests that home possession items that have small variation in ICC curves for large groups of nations may be uncommon. The international and nation-specific ICC curves for four items asked in at least 10 nations are plotted in Figure 1.

Evidence of Scale Reliability and Validity

The median reliability of the individual family SES scores produced by the constrained model was .75. For the unconstrained model, the median reliability was .74. The average of the within-nation reliabilities was .58 for the constrained model and .62 for the unconstrained model. This suggests that while very little cross-national reliability is lost under the new model, there is a slight increase in the ability to differentiate between students with different values of family

SES within nations.

The correlation between national mean SES from the unconstrained model and national Gross Domestic Product from 1995 (GDP) is .64.[6] The correlation between national mean SES from the unconstrained model and expected educational attainment (1994-97) is .71.[7] This suggests that aggregating the individual family SES scores derived from TIMSS data using international anchor items and nation-specific items provides a reasonably reliable and valid indicator of SES. It also suggests that, at the national level, the relative influences of educational and economic factors on the SES scores are approximately equal.

Discussion

The new multilevel SES scaling model using international anchor items and nation-specific items has significant advantages over traditional SES scaling models. Consequentially, these advantages are both empirical and theoretical. The empirical advantage is straightforward: the new model fits actual data much better than a more traditional model. The theoretical advantage stems from the multilevel configuration of the new model and has two components. First, the separation of cross-nation and within-nation components of SES allows for the use of nation-specific items. These items need not be comparable across nations in terms of their psychometric characteristics, and can be selected so that the nation-specific items are tailored to the specific conditions and cultures of the nations in which they are used. Second, the use of carefully selected anchor items has the potential to reduce the contamination of the SES scale by non-SES factors at the national level. In other words, the new scaling model can produce true international comparability, whereas traditional scaling methods using home possession items are subject to influences resulting from cultural and situational factors outside of SES. For example, including the “VCR” item in a traditional scaling model may result in higher SES

scores for nations with a greater cultural affinity for videos – an affinity that may be unrelated to SES.

The results of this research show that responses on international student background surveys that include both direct measures of family educational status and indirect measures of wealth through home-possession items can be used to derive internationally comparable scales of family SES. However, this research also provides strong evidence that the assumption that home possession items have similar characteristics in each nation may be incorrect. Therefore, the implications of this research for the measurement of family SES using student questionnaire responses are two-fold.

First, anchor items are necessary to ensure international comparability and must have the same or very similar characteristics (i.e., item parameter estimates) in each nation. Therefore, any indirect measures of SES used as anchor items should have similar characteristics and relationships with SES in each nation. Any national-level variation in the responses to such items should be due only to national-level variation in SES. Any substantial influence of non-SES factors (e.g., cultural differences, climatological differences, etc.) on the national-level variation in students' responses is reason to exclude that item as an anchor item. Failing to do so would introduce national-level variation in the resultant scale that was not indicative of national differences in the three components of SES: educational attainment, occupational status, and income or wealth. Ideal anchor items are direct measures of family SES that can be used on student surveys including parents' years/type of schooling, parents' occupations, and direct measures of household income or wealth. Any national-variation in such direct measures of SES would be due to national differences in SES and should be reflected in an international scale of SES. Indirect measures of SES (e.g., home possession items) should be used as anchor items

only when they have been shown to have little variation in item parameter estimates across nations.

The second implication for the measurement of student family SES using surveys is that items which are not suitable as international anchors can still provide valuable information within nations. The new scaling model used in this research provides a means by which nation-specific items can be used in conjunction with international anchor items to improve the reliability and validity of student-level scores. These nation specific items can have any amount of national-level variation in their item parameters. In fact, these items can be completely different from nation to nation, as long they are valid indicators of SES within nations. This allows the selection of nation-specific items that are tailored to the cultures of each nation. Such items should be selected so that they maximize the ability to differentiate between students with different levels of family SES within the relevant range of SES for each nation. Hence, if a particular set of home possessions are known to be excellent indicators of SES within one nation, the nation-specific component of this model allows those items to be used on student surveys in that nation, while other, more appropriate items are used on student surveys in other nations. This approach has the potential to lead to very efficient and valid measurement of family SES in international studies of education.

It is important to note that although cultural and geographic differences that require the use of group-specific items are surely prevalent among nations, such differences may exist among other naturally occurring groups such as regions, states, or municipalities. Therefore, the scaling methods described here could also be used to improve the validity of SES scales in other contexts.

The most serious limitation of this study results from the fact that it relies on existing

international data and the use of specific international anchor items which may not be entirely suitable for this purpose. Yet the primary goal of this research is not to derive the perfect scale of SES from TIMSS data. It is to propose a new process by which international survey items relevant to SES and similar constructs are designed and then combined into a single scale. It is almost certain that the five indirect measures of SES used as anchor items in this study would behave differently across nations if the model allowed it. Unfortunately, this is not a viable option within this analysis. If the home possession anchor items were included as nation-specific items, then all national-level variation would be due only to national differences in educational attainment, and much of the national-level reliability would be lost.

An alternative application of this model could have utilized data from the 2000 PISA study, which collected information on parents' educational attainment, parents' occupational status, and several home possessions. However, the PISA study did not allow nations to select nationally tailored survey items. Therefore, a similar analysis using PISA data would serve to explore national variation in a fixed set of survey items, but it would not illustrate the feasibility of using nationally tailored survey questions as nation-specific items for scaling SES. The most significant value of using this scaling model with the PISA data would be to focus on identifying indirect measures of SES that operate similarly across different nations. Combining these with direct measures of SES and nation-specific items has the potential to improve the SES information provided by such international studies.

References

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years. IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA.: International Study Center, Boston College.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA.: International Study Center, Boston College.

Bradlow, E. T., (1994). Analysis of Ordinal Survey Data with 'No Answer' Responses (Doctoral dissertation, Harvard University, 1994). *Dissertation Abstracts International, 56*(01), AAT 9514762.

Bradlow, E. T., & Zaslavsky, A.. M. (1999). A Hierarchical Latent Variable Model for Ordinal Data From a Customer Satisfaction Survey with ‘No Answer’ Responses. *Journal of the American Statistical Association, 04*(445), 43-52.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A. C. Porter, & A. Gamoran (eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 150-197). Washington, D. C.: National Academy Press.

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries: An empirical study.* New York: Wiley.

Elley, W. B. (1994). *Achievement and Instruction in Thirty-two School Systems: The IEA Study of Reading Literacy.* Oxford, England: Pergamon Press.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Geman, S. & Geman, D. (1984). Stochastic relaxation Gibbs distributions, and the Bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*. 6(6) 721-741

Gonzalez, E. J., & Smith, T. A. (1997). *User guide for the TIMSS international database: Primary and middle school years*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Gorman, T. P., Purves, A., & Degenhart, R. E. (1988). *The International Writing Tasks and Scoring Scales: The IEA Study of Written Composition I.* Oxford, England: Pergamon Press.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Husen, T. (1967). *International study of achievemnt in mathematics.* Stockholm, Sweden: Almqvist and Wiksell.

Keeves, J. P., & Saha, L. J. (1992). Home background factors and educational influences. In J. P. Keeves (ed.), *The IEA study of science III: Changes in science education and achievement: 1970-1984* (pp. 165-186). Oxford, England: Pergamon Press.

Kim, S., & Cohen, A. S. (1998). *An evaluation of a Markov Chain Monte Carlo method for the two-parameter logistic model*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J: L. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A., & O'Connor, K. M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Science and Science Study at the Eighth Grade*. Chestnut Hill, MA.: International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Gregory, K. D., Hoyle, C., & Shen, C. (2000). *Effective Schools in Science and Mathematics*. Chestnut Hill, MA.: International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Science and Science Study at the Eighth Grade*. Chestnut Hill, MA.: International Study Center, Boston College.

Muraki, E., & Bock, R. D. (2002). PARSCALE (Version 4.1) [Computer Software] Illinois: Scientific Software International.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B., 56*(1), 3-48.

Organisation for Economic Co-operation and Development (OECD). (2002). *PISA 2000 Technical Report*. Paris: Organisation for Economic Co-operation and Development.

Powers, M. G. (1982). Measures of socioeconomic status: An Introduction. In M. G. Powers (ed.), *Measures of socioeconomic status: Current issues* (pp. 1-28). Boulder, CO: Westview Press, American Association for the Advancement of Science.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology, 25*, 111-163.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden, & R. K. Hambleton (eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B., 53*(3), 683-690.

Spiegelhalter, D., Thomas, A., & Best, N. (2000). WinBUGS (Version 1.3) [Computer Software] Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health; London, UK: Department of Epidemiology and Public Health, Imperial College School of Medicine at St. Mary's Hospital. [On-line] Available: http://www.mrc-bsu.cam.ac.uk/bugs

United Nations Organization for Education, Science and Culture (UNESCO). (2002). *UNESCO Education and Literacy Database*. [electronic database] Available from the UNESCO Institute for Statistics: http://www.uis.unesco.org/en/stats/stats0.htm

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden, & R. K. Hambleton (eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

Walker, D. A. (1976). *The IEA six subject survey: An empirical study of education in twenty-one countries.* Stockholm, Sweden: Almqvist and Wiksell.

Westbury, I., & Travers, K. J. (1990). *Second International Mathematics Study.* Urbana, IL: University of Illinois.

Wolf, R. M. (1992). The Second International Science Study. *International Journal of Educational Research, 17*(3-4), 227-397.

World Bank. (2002). *World Development Indicators Database*. [electronic database] Available from: http://www.worldbank.org/data/wdi2002/

Footnotes

[1] The contents of the nation-specific items and their positions in the questionnaire are documented in Gonzalez and Smith (1997).

[2] This same technique for creating a calibration sample was used by TIMSS analysts for the purposes of estimating item parameters for the IRT model of student achievement in TIMSS.

[3] The two parental education variables had up to 25% missing data. The home possession items all had less than 2% missing data.

[4] A plausible value is any value drawn at random from the posterior distribution of a parameter.

[5] Due to constraints on computing resources, monitoring of parameter estimates could only be carried out for a single chain beyond the burn-in period. Each model ran in just over 12 hours on a 2.53 GHz computer running WinBUGS 1.3 under Windows XP Professional.

[6] The source for national GDP is the World Development Indicators Database from the World Bank (2002).

[7] The source for national educational attainment statistics is the Education and Literacy Database from the United Nations Educational, Scientific and Cultural Organization (UNESCO, 2002).

Table 1

Point estimates of item parameters for the seven anchor items estimated using the constrained and unconstrained models of SES.

	Constrained Model		Unconstrained Model	
Item	Threshold [a] (β–δ)	Slope (α)	Threshold [a] (β–δ)	Slope (α)
Father's Education				
Finished primary school	rc	rc	rc	rc
Finished some secondary school	-1.67	0.73	-1.42	0.95
Finished secondary school	-0.54	0.73	-0.45	0.95
Some Technical/Vocational Ed.	0.63	0.73	0.55	0.95
Some University	1.21	0.73	1.05	0.95
Finished University	1.53	0.73	1.33	0.95
Mother's Education				
Finished primary school	rc	rc	rc	rc
Finished some secondary school	-1.35	0.74	-1.14	0.99
Finished secondary school	-0.37	0.74	-0.32	0.99
Some Technical/Vocational Ed.	0.83	0.74	0.71	0.99
Some University	1.34	0.74	1.15	0.99
Finished University	1.73	0.74	1.46	0.99
Number of Books in the Home				
None or very few (0 - 10 books)	rc	rc	rc	rc
one shelf (11-25 books)	-2.30	0.69	-2.22	0.73
one bookcase (26-100 books)	-1.09	0.69	-1.06	0.73
two bookcases (101 - 200 books)	0.30	0.69	0.29	0.73
Three or more bookcases (more than 200books)	1.21	0.69	1.16	0.73
Calculator	-2.10	1.09	-2.19	1.00
Computer	0.12	0.77	0.12	0.62
Study Desk	-1.83	0.90	-1.85	0.88
Dictionary	-2.17	0.87	-2.26	0.81

Note. rc = reference category

[a] The threshold parameters for items with multiple categories are calculated as the difference between the overall item threshold (β) and the category (δ) parameters as shown in the equation for the constrained model.

Table 2

Point estimates of item parameters for home possession items asked in at least 10 nations estimated using the constrained and unconstrained models of SES.

	Constrained Model		Unconstrained Model [a]	
Item	Threshold β	Discrimination α	Threshold β_{min} , β_{max}	Discrimination α_{min} , α_{max}
Video camera	1.40	0.51	0.48 , 2.39	0.09 , 0.79
Dishwasher	0.29	1.09	-0.68 , 3.13	0.36 , 1.26
Microwave oven	-0.31	0.81	-4.19 , 1.37	0.11 , 1.17
Car	-0.50	0.80	-2.13 , 0.74	0.30 , 1.38
CD player	-0.58	0.88	-4.26 , 2.98	0.08 , 5.61
VCR	-0.67	0.75	-4.37 , 1.37	0.07 , 7.99
Encyclopedia	-0.92	0.82	-1.78 , 0.14	0.46 , 1.23
Own room/bedroom	-1.49	0.46	-4.32 , -0.04	0.08 , 0.76
Television	-3.33	0.58	-8.36 , -1.43	0.08 , 16.79

[a] The minimum and maximum point estimates across all nations are shown for the unconstrained model.

Table 3

Standard integrated distance scores for nation-specific items asked in two or more nations.

Item	Total Nations	Std. Distance	Item	Total Nations	Std. Distance
Telephone	6	.27	Video games	6	.12
Cable/satellite TV[a]	6	.25	Own books	3	.12
Washing machine	6	.25	Two+ bathrooms	4	.12
Microwave oven	12	.24	Newspaper/magazines[d]	5	.11
Motorcycle	2	.24	Two+ televisions	4	.11
Air conditioner	6	.24	Educational computer pgm[e]	2	.11
VCR[b]	25	.23	Camera	4	.11
Own bicycle	2	.23	Television	15	.11
CD player	14	.22	Laboratory instruments	2	.10
Two+ cars	8	.22	Piano/organ (or violin)	6	.10
Garden	5	.22	Four+ bedrooms	2	.10
Telescope or binoculars	5	.21	Cordless telephone	4	.10
Stereo/audio system	7	.19	Electronic gameboard	2	.09
Car	14	.18	Domestic help/servants	2	.09
Dishwasher	11	.18	Bicycle	5	.09
Aquarium or pets	4	.18	Lawn mower	3	.08
Central heating	4	.17	Own television	2	.08
Cassette player	9	.16	Microscope	4	.08
Boat or Cabin	4	.16	Swimming pool	2	.07
Atlas (or globe)[c]	6	.15	Own CD or video player	3	.07
Own room/bedroom	10	.14	Refrigerator	4	.06
Encyclopedia	13	.14	Classical music	3	.06
Clothes dryer	6	.14	Radio	5	.05
Summer/weekend house	3	.13	FAX or faxmodem	2	.05
Portable CD player	2	.12	Study corner	2	.03
Video camera	15	.12	Gas stove	2	.03
Musical instruments	4	.12	House	2	.01

[a] Iceland used both “cable TV” and “satellite dish” on their survey. The “satellite dish” item from Iceland is not used in this distance calculation.
[b] Portugal used both “VCR” and “video cassettes” on their survey. The “video cassettes” item from Portugal is not used in this distance calculation.
[c] Norway used both “atlas” and “globe” on their survey. The “globe” item from Norway is not used in this distance calculation.
[d] Latvia and the Netherlands used both “newspaper” and “magazine” on their surveys. The “magazine” items from Latvia and Netherlands are not used in this distance calculation.
[e] Iceland used both “mathematics computer program” and “science computer program” on their survey. The “science computer program” item from Iceland is not used in this distance calculation.

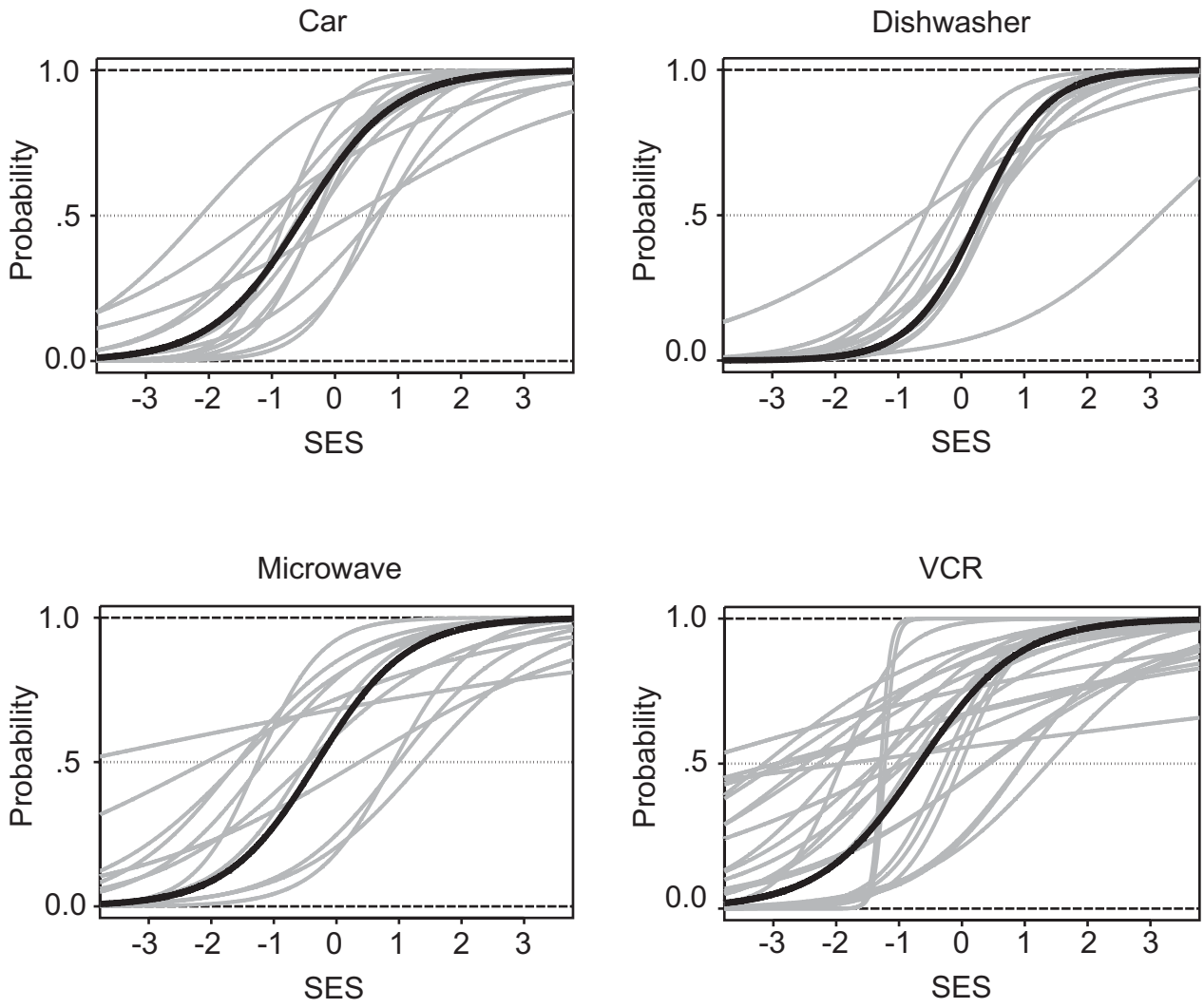

Figure 1. International (bold) and nation-specific (gray) item characteristic curves (ICC) for Four items.