

APPLYING ITEM RESPONSE THEORY METHODS TO COMPLEX SURVEY DATA

D. Roland Thomas and Andre Cyr ¹

ABSTRACT

Item response theory (IRT) offers many advantages to researchers who need to quantify children's reading and writing abilities, and for this reason, IRT methods have been adopted in Statistics Canada's National Longitudinal Survey for Children and Youth. IRT methods have a long history in the field of psychometrics, and provide a model based method for characterizing both test items and subject abilities, and for generating predictions of individual abilities. For the most part, IRT methods implicitly assume independent and identically distributed (i.i.d.) observations, so that the application of these methods to complex surveys raises a number of issues. The paper will review basic IRT theory and provide a rationale for its use in complex surveys. NLSCY data will be used to illustrate various IRT issues, including point and variance estimates of item parameters, the potential for bias due to ignoring survey weights, biases in the distribution of ability predictors, and the dependence of this bias on test length.

KEY WORDS: Item response theory; IRT; complex surveys; pseudo likelihood; ability predictions; EAP estimates; latent distributions; attenuation.

RÉSUMÉ

La théorie des éléments de réponse (IRT) offre beaucoup d'avantages aux chercheurs qui doivent mesurer les capacités de lecture et d'écriture des enfants, et pour cette raison, les méthodes relatives à IRT ont été adoptées dans l'enquête longitudinale sur les enfants et la jeunesse de Statistique Canada (NLSCY). Les méthodes d'IRT ont une longue histoire dans le domaine de la psychométrie, et fournissent une méthode basée sur un modèle pour caractériser des éléments de test et des capacités du sujet à l'étude, et pour produire des prévisions concernant les capacités individuelles. La plupart du temps, les méthodes d'IRT assument implicitement des observations i.i.d., de sorte que l'application de ces méthodes aux enquêtes complexes soulève un certain nombre de questions. Le but de cet article est de passer en revue la théorie de base de l'IRT et de fournir les raisons pour son utilisation dans le contexte des études complexes. Les données provenant de NLSCY seront utilisées pour illustrer les différents problèmes reliés à l'IRT, tels que l'estimation de paramètres et de leur variance, le potentiel de biais lorsqu'on ne tient pas compte des poids d'enquête, le biais dû à la distribution des prédicteurs des compétences ainsi que la dépendance de ce biais sur le temps de complétude des évaluations.

MOTS CLÉ: Enquêtes complexes, estimations EAP, IRT, la théorie des éléments de réponse, prédictions des compétences.

1. INTRODUCTION

Item response theory (IRT) provides a set of techniques for predicting the value of a trait or ability for one or more individuals. The traits or abilities in question cannot be directly measured, but must be inferred from each individual's responses to a set of questions, or test items. IRT methods allow for both the calibration of the test items and the prediction of individual abilities, and are now used in Statistics Canada's National Longitudinal Survey of Children and Youth (NLSCY). One of the goals of the NLSCY is to track children's educational development, and in the most recent administration of the survey (Cycle 3, 1998-99) both reading ability and mathematical ability measures for each

child were directly predicted using IRT methods. These IRT "scores" were subsequently added to the NLSCY datafile for subsequent aggregation and analysis.

IRT methods have a long history dating back to Lord (1952) and earlier, and are now well established in the field of psychometrics, providing an alternative to the "classical test theory" (CTT) which was used in the first NLSCY cycle to construct tests of mathematical ability. IRT is a model-based approach that provides additional tools for measuring traits and abilities by clearly separating test items, characterized by individual item parameters, from the characteristics of examinees. The purpose of this paper is to review the practical and theoretical issues arising from the application to complex surveys of IRT methods that were developed

¹ D. Roland Thomas, Sprott School of Business, Carleton University, Ottawa, Ontario, Canada K1S 5B6; Andre Cyr, Social Surveys Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

primarily under the assumption of independent observations. It is hoped that the paper will be of interest to survey statisticians who may not be familiar with IRT methods, and to analysts who may be familiar with IRT methods, but less familiar with the sampling literature. Specific issues will be illustrated using data from the NLSCY.

1.1 Complex Survey Applications of IRT and Associated Literature

Probably the earliest large scale survey to use IRT methods is the National Assessment of Education Progress (NAEP), a Congressionally mandated national assessment program in the U.S., which has been in operation for nearly 30 years. The methodology used in the NAEP differs in many important respects from that used in the NLSCY. Rather than using ability "scores" for subsequent aggregation, the NAEP obtains aggregate measures of ability using the "plausible values" (PV) methodology, an adaptation of Rubin's (1987) multiple imputation technique for missing data (the missing data here being the latent ability scores). As will be seen later, avoiding the calculation of individual ability scores can be of considerable advantage. Plausible values methodology has now become widespread in the field of large-scale educational testing. For example, it was used in the Third International Mathematics and Science Study (TIMSS), a large international survey with participation from 38 countries, conducted most recently by the International Association for the Evaluation of Educational Achievement (IEA) in 1995. Another example is the Program for International Student Assessment (PISA), designed to assess the reading, math and science skills of 15 year olds, in which Statistics Canada participates.

In the NAEP literature, in particular the special issue of the *Journal of Educational Statistics* (1992) and the most recent technical report describing NAEP methodology (Allen, Carlson and Donoghue, 2001), there is no mention of the survey design in connection with either item parameter estimation or plausible value generation. The survey design is mentioned only in conjunction with the variance of aggregate statistics computed from the plausible values, in which case the sampling component of the total variance is estimated using design-based methods. A review of NAEP methodology by Patz (1996) does provide an outline of the use of survey weights during the plausible values generation stage, though not at the stage of item parameter estimation. Documentation and other information on TIMSS and PISA are likewise silent on the issues of design complexity. The IRT methodology used in the National Education Longitudinal Study (NELS), sponsored by the U.S. National Centre for Education Statistics, was very similar to that of the NLSCY. However, the NELS technical report by Rock, Pollack and Quinn (1995) again fails to mention the survey

design in connection with item parameter estimation or ability prediction.

It is clear from these examples that, at the very least, the issue of survey design and IRT parameter estimation for large scale surveys is poorly documented. This paper will attempt to remedy that deficiency.

1.2 Organization of the Paper

Section 2 of the paper will provide an overview of IRT models and corresponding estimation techniques designed for samples of independent and identically distributed (i.i.d.) observations. The focus will be on estimation of fixed parameters, i.e., the item parameters and the parameters of the ability distribution. A rationale for the design-based (i.e., weighted) estimation of these parameters will be described in Section 3, with particular reference to the software used in the NLSCY. Section 4 discusses the lack of consensus in the educational testing community regarding the need to take account of the survey design when estimating IRT parameters, and provides an illustration of the need for design weighted point estimation. Section 5 focuses on the methods used in the NLSCY for predicting individual abilities, and demonstrates the critical importance of test length using NLSCY data. A summary is given in Section 6.

2. IRT MODELS AND ESTIMATION TECHNIQUES

The key to IRT is a model that links the characteristics of a given item, and the latent ability of an individual subject, to the probability that the subject will respond correctly to that test item. IRT models for both binary and polytomous test items have been developed, but since all test items used in the NLSCY are binary, this discussion will be confined to models of the binary logistic type.

2.1 The Three-Parameter Logistic Model

Consider a set of n binary test items, administered to a sample of N individuals, with the aim of measuring the latent ability of interest. Under the three-parameter logistic response model, the probability of a correct response by the i 'th individual to the j 'th binary item is modeled as

$$P_j(\theta_i) = c_j + (1 - c_j) \frac{\exp\{\alpha_j(\theta_i - \beta_j)\}}{1 + \exp\{\alpha_j(\theta_i - \beta_j)\}} \quad (2.1)$$

for $i = 1, \dots, N; j = 1, \dots, n$, where θ_i is the value of the ability of individual i . For specific values of α_j , β_j and c_j ,

$P_j(\theta_i)$ is referred to as an item response function. The α_j parameter is referred to as the item discrimination parameter, the location parameter β_j is considered to represent item difficulty, while the third parameter, c_j , usually referred to as the guessing parameter, represents the observation that no item ever has zero probability of a correct response. It is clear from equation (2.1) that without some additional restrictions, the item parameters and the abilities are not identified. The usual way method of dealing with the linear indeterminacy is to arbitrarily fix the location and scale of the θ distribution.

3.2 Item Parameter Estimation

The most commonly adopted method for estimating IRT parameters is marginal maximum likelihood (MML), based on the work of Bock and Lieberman (1970) and Bock and Aitkin (1981). Other estimation methods have been reviewed by Baker (1992), and an MCMC approach has been described by Patz and Junker (1999).

For MML estimation with i.i.d. observations, the marginal likelihood is

$$L = \prod_{i=1}^N \int P(\mathbf{U}_i | \theta, \xi) g(\theta | \boldsymbol{\tau}) d\theta \quad (2.2)$$

where \mathbf{U}_i represents the data, consisting of binary test responses u_{ij} , $j = 1, \dots, n$, ξ denotes the set of item parameters $\{\alpha_j, \beta_j, c_j, j = 1, \dots, n\}$, and $\boldsymbol{\tau}$ is a vector of parameters of the assumed distribution, g , of θ . Under local independence of item responses, the first term under the integral is given by

$$P(\mathbf{U}_i | \theta, \xi) = \prod_{j=1}^n P_j(\theta_i)^{u_{ij}} \{1 - P_j(\theta_i)\}^{(1-u_{ij})}. \quad (2.3)$$

All versions of the MML algorithm for i.i.d. observations begin by obtaining the derivatives of the marginal likelihood (2.2) with respect to the set of item parameters $\xi = \{\alpha_j, \beta_j, c_j, j = 1, \dots, n\}$, with the IRT model defined by equations (2.1) and (2.3). In the interests of brevity, the following summary will be restricted to the two parameter model, i.e., the model with all guessing parameters, c_j , set to zero. After some algebra, the likelihood equation for the j 'th discrimination parameter α_j becomes:

$$\frac{\partial \log L}{\partial \alpha_j} = \sum_{i=1}^N \int [u_{ij} - P_j(\theta)] (\theta - \beta_j) [P(\theta | \mathbf{U}_i, \xi, \boldsymbol{\tau})] d\theta \quad (2.4)$$

$$= 0,$$

where $P(\theta | \mathbf{U}_i, \xi, \boldsymbol{\tau})$ denotes the posterior distribution of θ , namely

$$P(\theta | \mathbf{U}_i, \xi, \boldsymbol{\tau}) = \frac{P(\theta | \mathbf{U}_i, \xi) g(\theta | \boldsymbol{\tau})}{\int P(\theta | \mathbf{U}_i, \xi) g(\theta | \boldsymbol{\tau}) d\theta}. \quad (2.5)$$

For the j 'th difficulty parameter β_j , the corresponding likelihood equation is

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^N \int [u_{ij} - P_j(\theta)] [P(\theta | \mathbf{U}_i, \xi, \boldsymbol{\tau})] d\theta \quad (2.6)$$

$$= 0.$$

Most implementations of MML, e.g., the program BILOG-MG used in the NLSCY, are variations on the approach of Bock and Aitken (1981). Following Mislevy and Stocking (1989), the θ distribution is represented in histogram form, using equally spaced θ nodes denoted X_k , $k = 1, \dots, q$, with nodal weights A_k . The A_k are chosen to represent a normal distribution in the first iteration, but if the assumption of normality is not tenable, the normality assumption can be relaxed in which case the nodal weights, A_k , become the elements of the unknown parameter vector $\boldsymbol{\tau}$, to be estimated along with the item parameters ξ . In quadrature form, equations (2.5) and (2.6) reduce to:

$$\sum_{k=1}^q (X_k - \beta_j) [\bar{r}_{jk} - \bar{f}_k P_j(X_k)] = 0 \text{ and} \quad (2.7)$$

$$\sum_{k=1}^q [\bar{r}_{jk} - \bar{f}_k P_j(X_k)] = 0,$$

where $\bar{f}_k = \sum_{i=1}^N P(X_k | \mathbf{U}_i, \xi, \boldsymbol{\tau})$ represents the expected frequencies of examinees out of a sample of size N expected to have ability $\theta = X_k$, and where $\bar{r}_{jk} = \sum_{i=1}^N u_{ij} P(X_k | \mathbf{U}_i, \xi, \boldsymbol{\tau})$ represents the expected frequency of examinees having ability X_k who correctly answer item j . The posterior probability $P(X_k | \mathbf{U}_i, \xi, \boldsymbol{\tau})$ featured in the expressions \bar{f}_k and \bar{r}_{jk} can be written as

$$P(X_k | \mathbf{U}_i, \xi, \boldsymbol{\tau}) = \frac{L_i(X_k) A_k}{\sum_{k=1}^q L_i(X_k) A_k}, \quad (2.8)$$

where $L_i(X_k) A_k = \prod_{j=1}^n P_j(X_k)^{u_{ij}} \{1 - P_j(X_k)\}^{(1-u_{ij})}$. Though the likelihood equations (2.7) resemble those for logistic regression, the "frequencies" \bar{f}_k and \bar{r}_{jk} are functions of the unknown parameters ξ . However, these expected frequencies can be regarded as the E-step of an EM

algorithm, with the solution of the estimating equations (2.7) comprising the M-step, yielding an easily implemented iterative solution. For non-normal θ distributions, the quadrature weights, A_k , are modified at each iteration of the EM algorithm using the scheme

$$A_k^{(t-1)} = \frac{1}{N} \sum_{i=1}^N \left[\frac{L_i(X_k) A_k^{(1)}}{\sum_{k=1}^q L_i(X_k) A_k^{(1)}} \right]. \quad (2.9)$$

3. A PSEUDO-LIKELIHOOD MML-EM ALGORITHM FOR COMPLEX SURVEY DATA

This section describes a rationale for the design-weighted MML estimation used in the NLSCY, and implemented in BILOG-MG, which is a straightforward application of the approach of Binder (1983), Fuller (1975) and others. Details are provided because many IRT analysts are unfamiliar with complex survey data analysis, and since the rationale appears not to have been presented in the psychometric literature.

3.1 The Pseudo-Likelihood Approach

The pseudo-likelihood approach (see Binder, 1983) can be applied to the current problem by considering a marginal likelihood of the form (2.2) written for a finite population of N^P units. This finite population will in turn be considered a random realization from some super-population for which the IRT model is defined. If the test responses for all respondents in the finite population were known, then equations (2.4) and (2.6), with N replaced by N^P , would comprise likelihood equations for the super-population. Solutions to these ‘‘census’’ equations define finite population parameters termed ‘‘corresponding descriptive population quantities’’ (CDPQs) by Pfeifferman (1993). Of course, only a sample of units from the finite population is available. However, it can be seen from equations (2.4) and (2.6), with N replaced by N^P , that each census equation features a population sum that can be consistently estimated using the sample values and corresponding weights w_i . Thus if the summations over N^P in the census equations are replaced by weighted sums over the complex sample of size N , estimates of the census likelihood equations are given by

$$\sum_{i=1}^N w_i \int [u_{ij} - P_j(\theta)] (\theta - \beta_j) [P(\theta | \mathbf{U}_i, \boldsymbol{\xi}, \boldsymbol{\tau})] d\theta = 0, \quad (3.1)$$

$$\sum_{i=1}^N w_i \int [u_{ij} - P_j(\theta)] [P(\theta | \mathbf{U}_i, \boldsymbol{\xi}, \boldsymbol{\tau})] d\theta = 0. \quad (3.2)$$

Equations (3.1) and (3.2) comprise pseudo-likelihood equations, or estimating equations, whose solutions will

provide design-consistent estimates of the finite population analogues (the CDPQ’s) of the item and ability distribution parameters. When written in a form suitable for application of the EM algorithm, the estimating equations for the item parameters α_j and β_j are again given by equation (2.7), with the posterior expectations \bar{f}_k and \bar{r}_{jk} replaced by

$$\bar{f}_k = \sum_{i=1}^N w_i P(X_k | \mathbf{U}_i, \boldsymbol{\xi}, \boldsymbol{\tau})$$

and

$$\bar{r}_{jk} = \sum_{i=1}^N w_i u_{ij} P(X_k | \mathbf{U}_i, \boldsymbol{\xi}, \boldsymbol{\tau}),$$

respectively. The iteration scheme (2.9) for the ‘‘pseudo-MLEs’’ of the parameters of the ability distribution can be expressed similarly in terms of a weighted sum over the units in the sample.

Finally, it will be assumed that the super-population and its finite sample realization are structured so that the design-based estimator of a finite population item parameter is also a consistent estimator of the corresponding super-population parameter (see, for example, Pfeifferman, 1993). This provides the final step in the rationale for applying IRT to complex surveys.

3.2 Estimation in the NLSCY

All IRT estimation in the NLSCY is carried out using the program BILOG-MG (Zimowski, Muraki, Mislevy and Bock, 1996). BILOG-MG allows for the specification of design weights, and implements design weighted point estimation of parameters consistent with the pseudo-likelihood approach described above (R.D. Bock, private communication). Experience with the NLSCY has shown that weighted estimates of item parameters and abilities may differ considerably from unweighted estimates, and weighted estimation has therefore been adopted as the NLSCY standard. Thus all item and latent distribution parameters in the NLSCY are consistently estimated by a survey weighted BILOG-MG analysis. It is important to note, however, that BILOG-MG does not generate design-consistent variance estimates when the weighted estimation option is selected.

Some preliminary efforts have been made to obtain bootstrap variance estimates of the item difficulty parameters, using the bootstrap weights available on the NLSCY main datafile. Results for the Cycle 2 mathematics test for Grade 5 boys are shown in Table 1 for a selection of test items. For the full test, the design effects, namely the ratio of the bootstrap variances to the BILOG-MG variances based on the assumption of

Table 1 - Bootstrap variance estimates of NLSCY item parameters (Cycle 2, Grade 5 Mathematics Test, Boys)

Item	Item Difficulty	BILOG SE	B'strap SE	Deffs
3	-0.738	0.13	0.22	2.86
6	-0.225	0.126	0.209	2.75
9	-1.178	0.203	0.346	2.91
12	0.209	0.13	0.233	3.21
15	-0.273	0.089	0.162	3.31
18	0.147	0.115	0.216	3.53

independent observations, range from about two to four, indicating a substantial clustering effect. This illustrates an important point. Even if unweighted and weighted estimation yields similar point estimates, the survey design still cannot be ignored for purposes of inference.

4. WEIGHTED VERSUS UNWEIGHTED ESTIMATION OF IRT PARAMETERS

Though survey statisticians would likely agree that the design-based strategy described above is the logical way to handle parameter estimation in the NLSCY and similar complex surveys, some members of the educational testing community still claim that survey weights can be ignored and that IRT parameters can be estimated from complex survey data using standard unweighted methods. One source of this belief is the IRT concept of invariance (see, for example, Hambleton and Swaminathan, 1985) which states tautologically that, if the IRT model is correct, the item parameters will be invariant to (or consistent with) whatever subset of examinees that is selected, and examinee abilities will be invariant to (or consistent with) whatever subset of pre-calibrated items are included in the test. Some of the conclusions in Mislevy and Sheehan's (1989) examination of the use of collateral information in IRT estimation may also contribute to this belief in unweighted analyses. These authors used a model-based likelihood argument to show (among other things) that unequal stratum probabilities in a stratified SRS sample can be ignored without biasing item parameter estimates. The key assumption in both these examples is that the IRT model is correct. However, it is a truism that no model ever fits perfectly. The issue is always whether departures from the model are sufficiently serious to overcome the benefits of a parsimonious description of the data, a principle that applies to all data modeling tasks including IRT. In the NAEP context, for example, Mislevy, Johnson and Muraki (1992) stated that "The IRT summary is

expected not to capture all meaningful variation in item response data but to reflect distributions of overall proficiency . . .". In other words, the NAEP analysts did not assume that the model was correct, only that it was realistic enough to provide a good overall summary of the main features of interest. Thus any discussion of the importance of the survey design in IRT estimation should consider the possibility of model mis-specification, and its effect on unweighted analyses. It will be argued below that model mis-specification can lead to severely biased parameter estimates if the design weights are ignored and that the most practical protection against local model failure is provided by design weighted point estimation.

When IRT models are used with large complex surveys, one model failure that is likely to be encountered is *differential item functioning* (DIF). DIF involves the IRT analysis of two or more sub-populations, and means that one (or more than one item) is more (or less) difficult for one of them. In other words, separate IRT analyses within the two sub-populations or domains would yield different item difficulty parameter values, $\beta_j^{(1)}$ and $\beta_j^{(2)}$ for the offending variable (the j 'th, say). In the context of a large national sample such as the NLSCY, it is not practical to search for and remove items exhibiting DIF in all possible sub-populations that may subsequently be of interest to researchers. Thus it is likely that DIF effects are hidden in the NLSCY data, and the IRT model $P(U_i | \theta, \xi)$ must therefore be regarded as a typically fallible model that provides a good overall description of test outcomes and student abilities. Further, if any of the sub-populations or domains exhibiting DIF effects have different average weights, it is clear that the estimates of the item difficulty parameters exhibiting DIF will differ depending on whether or not weights are used in the analysis. As noted earlier, point estimates of item parameters in the NLSCY have been found to differ considerably depending on the use or otherwise of survey-weighted estimation, particularly when the sample includes rare domains featuring large weights due to non-response adjustments. A likely explanation of this observation is that the basic IRT model assumption has been violated by DIF, with its impact magnified by interaction with specific domains having large weights.

This explanation has been reinforced by a numerical experiment involving the NLSCY Cycle 2 mathematics test for Grade 5 students. Separate difficulty parameters were obtained for boys and girls using the BILOG-MG DIF procedure, and those items exhibiting DIF were then identified by testing parameter differences against zero, using bootstrapped standard errors. The potential effect of an interaction between DIF and sub-population weights on item difficulty parameters was then simulated by comparing the item difficulties for the combined sample of boys and girls to

a modified sample of boys and girls formed by first multiplying the weights of the girls by a factor of 10. For the three test items exhibiting DIF, it was predicted that the increased weighting for the girls would cause a shift in the difficulty parameters based on the modified sample towards the values estimated for the girls alone, and this prediction was borne out. The shifts were in the predicted direction, and statistically significant. This confirms the earlier suggestion that differential item functioning can result in weight dependent item parameter estimates when the sub-groups exhibiting DIF have different average weights. Besides hidden DIF, other potential model violations include DIF-like effects for the discrimination and guessing parameters of the standard three-parameter model (Bock, 1993), as well as violations of the assumption of *essential unidimensionality* (Stout, 1987), which again might have specific subgroup interactions. Finally, it should be noted that the conclusions of studies such as Mislevy and Shehan’s (1989) should be extrapolated to complex surveys with considerable caution, because simplified likelihood analyses do not generally allow for the dependence among observations caused by clustering, or for other complex design features such as post-stratification and non-response weight adjustments.

As noted by Pfeffermann (1993), a primary benefit of design-based estimation is that it provides protection against misspecification of the model, in the sense that it yields design consistent estimates of descriptive population quantities that are clearly defined. In the IRT context, the pseudo-likelihood approach provides design-consistent estimates of the finite population IRT parameters defined by the census likelihood equations, and hence provides protection in a population averaged sense against the model violations discussed. Thus analyses that ignore the survey weights cannot be justified when IRT models are applied to data from large complex surveys. Design-based parameter estimation using the pseudo-likelihood approach of Section 3 should be used, together with appropriate design-based variance estimation.

5. ABILITY PREDICTION IN THE NLSCY

A variety of methods have been proposed for predicting the value of the latent ability, θ_i , for a given subject, given the subject’s test outcomes $\mathbf{U}_i = \mathbf{u}_i$. These methods include maximum likelihood estimates (MLE’s), obtained by treating θ_i as a fixed parameter and maximizing the “likelihood” (2.3) in which the locally independent items play the role of independent observations. The weighted likelihood estimator (WLE; Warm, 1989) is closely related to the MLE but has better conditional bias properties as $n \rightarrow \infty$, where n is the number of test items. Estimators can also be obtained from the posterior distribution of θ , given by equations (2.5) and (2.8). For example, the posterior mean has been adopted as

the ability predictor in the NLSCY, while in the NELS survey mentioned earlier, ability predictions were based on the posterior mode (Rock et al., 1995). It is standard practice to first estimate the item and ability parameters ξ and τ , and then treat them as fixed for purposes of ability estimation. For complex surveys, these parameters should be estimated using the design-consistent methods described in Section 3. The posterior mean used in the NLSCY can thus be regarded as an empirical Bayes predictor, $\hat{\theta}^{EB}$, expressed as

$$\begin{aligned} \hat{\theta}^{EB} &= \int \theta P(\theta | \mathbf{U}_i, \hat{\xi}, \hat{\tau}) d\theta \\ &\approx \frac{\sum_{k=1}^q X_k L_i(X_k) A_k}{\sum_{k=1}^q L_i(X_k) A_k} = \hat{\theta}^{EBD}, \end{aligned} \quad (5.1)$$

where $\hat{\theta}^{EBD}$ denotes the discrete approximation to $\hat{\theta}^{EB}$ provided by BILOG-MG. Note that psychometricians often refer to these predictors as “expected a posteriori” (EAP) estimators.

5.1 Characteristics of the Empirical Bayes Predictor

The large sample performance of the empirical Bayes predictor as a function of test length, n , will be similar to that of the Bayes predictor corresponding to known item and ability parameters ξ and τ . In this section, these will be considered super-population (rather than finite population) parameters. It is well known that the sample mean of the Bayes predictor is an unbiased estimator of the mean of the ability distribution, $\bar{\theta} = \int \theta g(\theta | \tau) d\theta$. It is also well known that the unconditional variance of the Bayes predictor, $\sigma^2(\hat{\theta}^B)$, is an underestimate of the unconditional variance of the latent ability distribution, $\sigma^2(\theta)$. This can be easily demonstrated by re-expressing $\sigma^2(\theta)$ in an obvious notation as

$$\begin{aligned} \sigma^2(\theta) &= V_{\mathbf{U}, \theta} = E_{\mathbf{U}} [V_{\theta}(\theta | \mathbf{U})] + V_{\mathbf{U}} [E_{\theta}(\theta | \mathbf{U})] \\ &= MSE(\hat{\theta}^B) + \sigma^2(\hat{\theta}^B). \end{aligned} \quad (5.2)$$

The variance of θ is thus underestimated by an amount equal to the mean squared error of the Bayes predictor, which is of order $O(n^{-1})$, n being the number of test items. The extent of this bias in the variance of the population distribution of examinee predictions will be investigated below using a subset of NLSCY data.

5.2 Under-Estimation of Ability Variance in the NLSCY

The distribution of abilities can be obtained without evaluating individual predictors, simply by aggregating individual posterior distributions over the population, i.e.,

$$\int P(\theta | \mathbf{U}, \xi, \boldsymbol{\tau}) h(\mathbf{U}) d\mathbf{U} = \int P(\mathbf{U} | \theta, \xi) g(\theta | \boldsymbol{\tau}) d\theta = g(\theta | \boldsymbol{\tau}), \quad (5.3)$$

where $h(\mathbf{U}) = \int P(\mathbf{U} | \theta, \xi) g(\theta | \boldsymbol{\tau}) d\theta$ is the marginal distribution of an individual test response. For i.i.d. samples, the corresponding estimate of the ability distribution is obtained by averaging the posteriors over the sample, in practice using the discrete form (2.8). For a complex survey such as the NLSCY, the estimation will involve the design weights, and will generate a consistent estimate of the discrete version of the population ability distribution given by $g(X_k | \boldsymbol{\tau} = \{A_k\})$, provided there are sufficient cases in each of the ability categories, X_k . Recall that we are assuming that a design-consistent estimate of a finite population parameter will also be a consistent estimate of the corresponding super-population parameter. Formally, from equations (2.8) and (5.3), we get

$$g(X_k | \boldsymbol{\tau} = \{A_k\}) = \sum_{i=1}^N w_i \frac{L(X_k) A_k}{\sum_{k=1}^q L(X_k) A_k}. \quad (5.4)$$

This estimate of the latent ability distribution is automatically generated by BILOG-MG. It depends only on the item and ability parameter estimates and is thus free of any bias arising from prediction of individual abilities. Therefore, comparing the histogram (5.4) to a histogram of individual $\hat{\theta}^{EBD}$ predictions, again obtained using design weighted aggregation, will yield a direct assessment of the bias in the population distribution of the ability predictions. The comparison of the two histograms will be based on their means and their variance.

The following empirical results relate to the 20 item reading test administered to Grade 4 students in Cycle 2 of the NLSCY. The indeterminacy in the ability measure was resolved by setting the mean and the variance corresponding to the estimated latent distribution (5.4) to 225 and 625, respectively. Table 5.1 displays the mean (*emean*) and standard deviation (*estd*) of the distribution of the empirical Bayes predictions of individual ability, for the full test and for two subsets of test items containing 13 and 6 items, respectively.

Table 2 -A Comparison of the Latent Distribution to the Distribution of Ability Estimates

Items (<i>n</i>)	emean($\hat{\theta}$)	estd($\hat{\theta}$)	φ
20	225.63	21.53	0.74
13	225.09	19.73	0.62
6	225.76	17.2	0.47

Note: $\bar{\theta} = 225; \sigma(\theta) = 25$

It can be seen from column two of the table that the mean of the empirical distribution is close to that of the latent distribution (i.e., 225) irrespective of the number of items, as expected. It can also be seen that the standard error of the distribution of ability predictors is an underestimate of the latent distribution standard error (i.e., 25), again as predicted. The extent of the underestimate is summarized in the last column of Table 2 in terms of a prediction effect, φ , which is similar in concept to the familiar design effect (*deff*) used to summarize the variance distortion arising from design complexity. For these data it is defined as $\varphi = e \text{var}(\hat{\theta}) / 625$. It can be seen that the prediction effect increases (i.e., the degree of under-estimation increases) as the total number of items per test decreases, again as expected. For the full test of 20 items, the ratio of the empirical distribution variance to the latent variance is 0.74, a modest distortion. However, for a short test of 6 items, a larger distortion is observed, namely $\varphi = 0.47$. The result for 13 items is intermediate between these extremes.

It is not immediately clear what impact the above latent variance distortion will have when the EAP scores are used directly in subsequent analyses. Some preliminary results based on an i.i.d. simulation (Thomas, Lu and Zumbo, 2003) suggest that a degree of distortion consistent with the 20 item scales used in the NLSCY will result in parameter attenuation of approximately 20% for a latent regression featuring one explanatory variable.

6. SUMMARY AND CONCLUSIONS

This paper has provided a summary of item response theory and considered its application in large complex surveys. In particular, a rationale for design-based estimation has been described. It has been shown that ignoring survey features, the weights in particular, can cause serious biases in point estimates of item parameters when the IRT model is misspecified. Even if unweighted estimates are free of bias, their estimated variances will be seriously distorted unless design features such as clustering are accounted for. Preliminary

bootstrap estimates of parameter variances suggest that variance estimates obtained using standard software may underestimate item parameter variances by a factor of between two and four. Techniques for predicting individual abilities have also been described, and the divergence between the true latent distribution and the distribution of predicted abilities has been estimated using NLSCY data. For 20 item NLSCY tests, the true latent variance is underestimated by about 25%. More recent work (Thomas et al., 2003) suggests that this degree of distortion may result in regression parameter attenuation of approximately 20%. Clearly, test length is a critical issue whenever ability estimates are to be directly used in later analyses. The plausible values methodology used in the NAEP survey avoids this source of bias, but at the cost of a major investment in data processing at the data production stage. Further research is needed to identify methods for analyzing latent variables in complex surveys that avoid both the bias inherent in using predicted scores and the heavy data processing requirements of the plausible values approach.

ACKNOWLEDGMENTS

The authors wish to thank Kathleen MacEachern for assistance with BILOG-MG computation and Catalan Dochitioiu for generating the bootstrap estimates. The first author also gratefully acknowledges financial support from Statistics Canada, and research grant support from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- Allen, N.C., Carlson, J.E. and Donoghue, J.R. (2001). Overview of Part II: The analysis of 1998 NAEP data. In *The NAEP 1998 Technical Report* (Donoghue and Schoeps, eds.), National Centre for Education Statistics. <http://nces.ed.gov/pubsearch/pubinfo.asp>
- Baker, F.B. (1992). *Item Response Theory*, New York: Marcel Dekker.
- Bock, R.D. (1993). Different DIFs. In P. Holland and H. Wainer (eds), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 115-122.
- Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, **46**, 443-459.
- B Bock, R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, **35**, 179-197.
- Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.
- Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph no. 7*. Psychometric Society.
- Mislevy, R.J., Johnson, E.G. and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, **17**, 131-154.
- Mislevy, R.J. and Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, **54**, 661-679.
- Mislevy, R.J. and Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, **13**, 57075.
- Patz, R.J. (1996). *Ph.D. Thesis*, Dept. of Statistics, Carnegie Mellon University.
- Patz, R.J. and Junker, B.W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, **24**, 146-178.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **62**, 317-337.
- Rock, D.A., Pollack, J.M. and Quinn, P. (1995). *Psychometric Report for the NELS:88 Base Year Through Second Follow-Up* (NCES 95-382). Washington DC: US Department of Education, National Center for Education Statistics.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
- Stout, W.F. (1987). A non-parametric approach for assessing latent trait dimensionality. *Psychometrika*, **52**, 589-617.
- Thomas, D.R., Lu, I.R.R. and Zumbo, B.D. (2003). Embedding IRT In Structural Equation Models: A Comparison With Regression Based On IRT Scores. Proceedings of Statistics Canada Symposium 2002, *Modelling Survey Data for Social and Economic Research*. Ottawa: Statistics Canada, in press.
- Warm, A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, **54**, 427-451.

Zimowski, M.F., Muraki, E., Mislevy, R.J. and Bock, R.D.
(1996). *BILOG-MG: Multiple -group IRT Analysis and
Test Measurement for Binary Items*. Chicago: Scientific

Software International.

