Item Response Theory and Latent variable modeling for surveys with complex sampling design
The case of the National Longitudinal Survey of Children and Youth in Canada

André Cyr and Alexander Davies

Social Survey Methods Division, Statistics Canada
Main Building, Room 2500, Tunney's Pasture
Ottawa Ontario, K1A 0T6
Andre.cyr@statcan.ca
Alexander.davies@statcan.ca

Background

For the purposes of social science research, a scale is a type of composite measure consisting of several items (questions) that share an underlying empirical or logical structure. A scale is created by assigning scores to patterns of responses resulting from a group of questions. Researchers use scales to draw inference upon non measurable characteristics (latent concepts) such as latent traits or abilities. They are often employed because they facilitate the efficient reduction of large amounts of data into manageable and meaningful sources of information. Moreover, scales have higher reliability than individual items do and can demonstrate the intensity of the relationships that may exist within a concept and between concepts.

Applying latent trait analysis or Item Response Theory (IRT) to survey data is relatively new at Statistics Canada. Many surveys at Statistics Canada use a multi-stage sample design that make use of stratification and/or clustering of population units before sampling. This type of sampling will be referred generally as a complex survey design. Surveys such as the Youth in Transition Survey (YITS), Program for International Student Assessment (PISA), the Adult Literacy and Life-skills Survey (ALLS) and its international counterpart IALS, the International Math and Science Survey (TIMSS) as well as the National Longitudinal Survey of Children and Youth (NLSCY) employ this methodology in various ways in order to meet different analytical goals.

IRT is a model-based approach for estimating latent scores and their associated standard errors. Some surveys focus upon making one time measurements while others try to assess the effects of changes of these measurements over time. Some focus on domain estimates while others necessitate individual measurements to assess growth over repeated measurements. The type of latent trait studied also varies from survey to survey. For example, NLSCY uses IRT estimated latent scores in order to measure student achievement in the areas of mathematics, reading and cognitive abilities while YITS uses IRT mainly to gauge students' responses to attitude-based questions. There are also many valid ways to use IRT within either a repeated or a longitudinal survey with regards to parameter estimation. While some surveys have decided to benchmark the estimated parameters used to describe the items used to assess ability and the model used to measure ability (the score function) to a reference population, other surveys may allow these parameters to change based upon each sampled population independently.

The goal of this Statistics Canada presentation on *Latent Variable Measurement Methods and Techniques* is neither to focus upon all these methods, nor even upon how they are used within the agency. Instead, the objective of this presentation is to highlight some unresolved challenges that have arisen when applying this methodology within the framework of a complex survey design. While these issues are common to the aforementioned surveys, they will be discussed contextually in terms of the NLSCY.

Latent traits

A latent trait is a variable that is not directly observable but does have a measurable impact on observable characteristics. Through observation of these characteristics, it is possible to make inferences about the presence or magnitude of a latent trait. Latent traits are assumed to have a continuous but unknown distribution, although for computational reasons this distribution is typically specified to be standard normal. For example, a child's mathematical ability is typically assessed by measuring their answers to solving mathematical problems or questions. The more questions we ask of the child and the wider the breadth of questions is included in the assessment, the more our understanding of that child's ability will be accurate. The process of applying a standard metric to a latent trait is referred to as "scaling." The term "scale" may refer to both the set of observations upon which the inferences are made as well as the metric upon which values of the latent trait are expressed.

The relationship between these items and the latent trait is assumed to be direct, and the items are assumed to be conditionally independent. In other words, responses to the items should be governed entirely by the latent trait, any covariance among the items is due to their common dependence on the assumed latent trait and there should be no covariance between item responses along any other latent dimension. For example, a scale on a questionnaire intended to measure attitudes towards mathematics should include items that are directly influenced by the latent attitudes towards mathematics, and should not include items that are also influenced by their attitude towards writing or sciences.

The relationship between responses to items and the latent trait has two main conceptualizations. The first, typically referred to as Item Response Theory (IRT) is probabilistic. The fundamental assumption of IRT is that the probability of producing a certain response on a specific item (e.g., producing a correct answer on a skill assessment, or strongly disagreeing with a statement describing a particular point of view) is a function of the latent trait and characteristics unique to the item. For example, some responses may have a low probability of occurring except for those respondents with extremely high latent trait values. Conversely, other responses may only have a high probability of occurring for those with low latent trait values. The *location* or *difficulty* of the item describes the region or regions of the latent trait distribution where the probability of producing a specific response changes from low to high. A simple example is a difficult mathematics test item – for most examinees, the probability of answering correctly is low, but is usually close to one for an examinee with extremely high mathematical ability. Typically, the models used to relate probability of response to the latent trait distribution are combinations or transformations of the logit or normal ogive functions.

The second conceptualization, which is known as factor analysis, or more generally as Structural Equation Modeling (SEM), assumes that each response represents a threshold on the continuum of the latent trait. For example, an incorrect response for a mathematics test item may be associated with respondents with a standard normal score below -1, whereas all respondents with a score above -1 would be expected to respond correctly. The value of -1 defines the threshold between the two response categories. If the thresholds between categorical responses are specified as non-linear (i.e., some respondents with scores below -1 may be expected to respond correctly, and some with scores above -1 may respond incorrectly), then the mathematical model describing the SEM conceptualization of individual item response looks very similar to the IRT model. These thresholds are typically estimated to maximize the correlations of the items with each other and with other related variables. The latent trait is estimated as a structural factor which explains the observed covariation between the scale items.

For both approaches, items may also vary in the degree to which they discriminate between individuals with different latent trait scores. From the IRT perspective, an item is poorly discriminating if the probability of producing a certain response does not vary that much across the latent trait distribution. In contrast, the probability of a response to a highly discriminating item varies a large amount over narrow regions of the latent distribution. From a SEM perspective, an item's discrimination is described by its correlation with the empirical factor (i.e., its *factor loading*). Highly discriminating items have high correlations with the factor, and poorly discriminating items have low correlations.

The main difference between IRT and SEM is in how the parameters describing the relationship between item response and the latent trait are estimated. IRT methods use all information in the pattern of responses

for the estimation of all item parameters and are sometimes referred to as *full-information methods*. In contrast, SEM methods estimate the relationship of items to the latent trait using only sufficient statistics – the item correlation matrix. For this reason, SEM methods are sometimes referred to as *partial information* methods.

The National Longitudinal Survey of Children and Youth (NLSCY)

A unique study of Canadians from birth to adulthood, the National Longitudinal Survey of Children and Youth (NLSCY) provides a single source of data for the examination of child development in context, including the diverse life paths of normal development. The NLSCY is designed to follow a representative sample of Canadian children from 0 to 25 years of age, with data collection occurring at two-year intervals. The current sample of NLSCY children is large enough for analysis by cohorts, sub-populations and provinces. Starting in 1994, the first year of data collection, the sample of about 20,000 children then aged 0 to 11, were selected and information about their home and school environment is being collected.

There are a number of milestone measurements that were included in this survey to assess children at different stages in their lives. While some measurements relied on existing classically scaled instruments (or tests), some were new and without a normative benchmark. Many of the direct measures in the NLSCY were also intended to follow development of children through time. The use of IRT seemed ideally suited to meet the many analytical needs being addressed by this survey.

IRT methods were adopted in the NLSCY to track children's educational and cognitive development, and tests were designed to measure their latent reading, mathematical and cognitive abilities administered in each two year survey cycle. It is intended that IRT-based estimates of students' abilities will be made publicly available for secondary analysis by interested researchers. The investigation of the theoretical and practical issues arising from the use of IRT methods in the NLSCY were placed under scrutiny when the effect of complex design on IRT score modeling was not addressed in the literature.

The Effect of the Complex Survey Design on Parameter Estimates

In IRT, the relationship between the "observable" and the "unobservable" quantities is described by a mathematical function. The item response models are then mathematical models, which are based on specific assumptions about the test or questionnaire data. Different models (IRT models), are formed through specifying the assumptions that one is willing to make about the responses given in the assessment. With complex survey data, the problem is like the one met in linear regression. One knows that, from answers observed of variables, parameters that characterize the model (the regressive coefficients) must be estimated. The literature for these methods is usually in the context of simple random samples or a specific domain. However, Statistics Canada needs to adapt these techniques to the complex survey design by incorporating survey weights and complexity into the analysis. Statistics Canada research into the effect of complex survey designs on item parameter estimation and their ancillary error measurements has concluded that survey weights must be incorporated into the estimation of the parameters and that the design effect was not negligible when estimating the sampling error.

In the literature there are references to the impact of the survey design when test subjects are not selected with equal probability. Lord in 1959 demonstrated that cluster sampling in a norming study for a standardized test might require testing as many as 12 to 30 times as many examinees to achieve estimates comparable in accuracy to those based on a simple random sample. Unfortunately many do not use the standard error produced by the software to evaluate the effectiveness of items used to estimate ability. Each estimate is subject to error. The range of that error should serve as an indication of the reliability of the estimates produced.

There are different models that are better suited for the type of data being studied. As well, there a few models being used and are programmed in different commercial software available for researchers. We use a logistic model which is appropriate for dichotomous data, which is true/false, or correct/incorrect data and the unidimensionality hypothesis. Of these the Rasch model is a 1 parameter model (difficulty parameter),

and the Birnbaum model is a 2 parameter model (difficulty and discrimination parameters) and the more general 3 parameter model can be expressed as follows:

$$P(\theta) = \frac{c_i + (1-c_i)e^{D a_i (\theta - b_i)}}{1 + e^{D a_i (\theta - b_i)}}$$

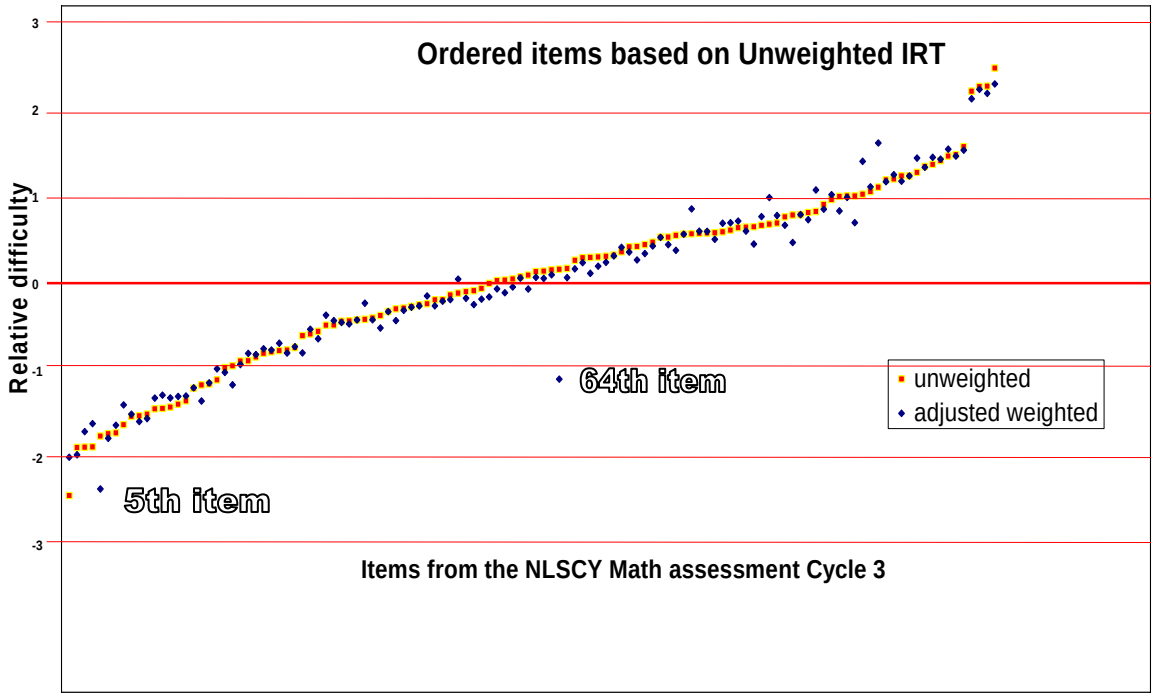
where θ *is the latent trait (ability) &*
b_i *(item difficulty)*
a_i *(item discrimination)*
c_i *(pseudo guessing)*

The value of D is a constant which can arbitrarily be set, however, usually D is set at 1.7 because then $P(\theta)$ for the normal and logistic ogives will not differ by more than 0.01 for any value of θ, the latent trait score.

Areas Researched and findings at Statistics Canada

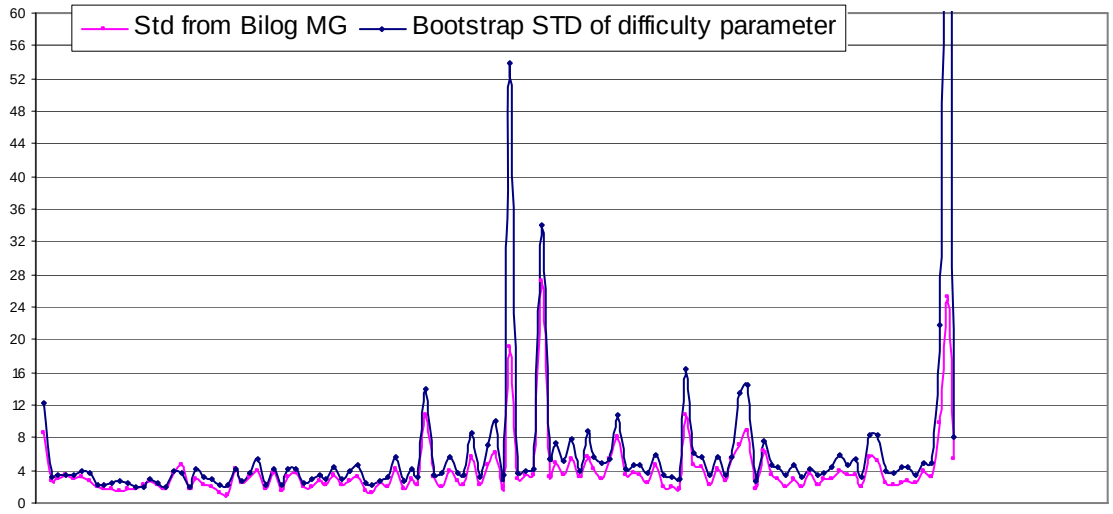
IRT methods have a long history and are now more widely used (in the field of psychometrics and educational testing) providing an alternative to the "classical test theory" (CTT) which, incidentally, was used in the first NLSCY cycle to construct tests of mathematical ability. However, very little attention has been paid to the use of IRT methods on data obtained from complex sample surveys, and there exists a major gap between IRT theory on the one hand and the analysis of complex survey data on the other hand. No general prescriptions exist for applying IRT methods to complex surveys, and the approaches taken in various major surveys differ widely. In some cases, design-based estimation is carried out (perhaps incompletely) and not comprehensively documented, and in other cases it is ignored. Irrespective of current practice, no plausible and convincing arguments exist for ignoring the survey design at the item estimation stage of IRT in complex surveys

The chart bellow shows the impact of survey weights on the estimated difficulty parameter. The data is from the NLSCY mathematical assessment of children from grades 2 to 10 administered in 2000-01. The original estimates of difficulty of each item were ranked and plotted in yellow; they were ranked from easiest item to most difficult for all mathematical items found on all the tests. The alternate (or true) estimates of item difficulty, using design information, are shown in blue; highlighting the rank differences once the sampling and response biases have been corrected.



Not surprisingly, the correlation between the estimated parameters remains high. However, the effects on parameter estimation can be substantial especially when one considers the likely link between design stratum (provinces) and school curriculum, known to have an affect on performance. Of particular note are two items that differ substantially from the originally estimated difficulty parameter (the 5th item near the beginning of the ordered items and the 64th item near the center of the ordered sequence). These items show a much reduced estimated difficulty; probably because they were easily answered in provinces where the sample weights were quite large. Similar findings were noted with the other parameters of the logistic model.

Our investigations then focussed on the variance estimates of these parameters. Using a series of 300 bootstrap weights, the model was calibrated with each sub-sample and an estimate of the variance was produced for the estimated difficulty parameter for all the items of the math assessment of the NLSCY. As noted in the graph below, the design effect on the variance was quite pronounced for some items, reaching as high as 4 times the standard error as calculated using a simple random sample.



Integrating Errors from Estimating Item Parameters into Score Estimation

Despite being able to compensate for the complex survey design, the question of how to ensure proper measurements remains problematic because of the limited flexibility of most commercially available software in terms of incorporating the survey design. Although normalizing the weights helps in accounting for the parameter variability in that it at least makes sure that the testing criteria are evaluated with respect to the proper sample size and that the estimates of the score obtained are design-consistent, this approach is not using the true variability of the sample associated with the complex survey design.

The estimation of latent trait scores operates under the assumption that the item parameters are known. Therefore, the variance of an individual's latent trait estimate is only a function of his or her response pattern and the item characteristics, which are assumed to be free of error. In practice, item parameters are usually estimated using the same sample for which scores are required. Thus, it may be an incorrect practice to ignore the variability resulting from the estimation of the item parameters. Scores produced for these examinees may be less accurate than assumed by the analyst. If this is the case, inferences made using this data may be questionable.

Effects of Differential Item Functioning

A further point of research has been focused on the interpretation of statistics resulting from the IRT. One such analysis of the resulting statistics is Differential Item Functioning (DIF). Distinct groups within a population may have higher or lower estimated scores on the latent trait that the scale is attempting to measure. This means that the scale is able to fulfill its purpose and discriminate between these groups on

the basis of their estimated level of the underlying trait. However, sometimes items are found to behave differently in distinct groups such as gender or language (such as loading on different dimensions in a multi-dimensional factor analysis, or having largely different mean item scores). In other words, two examinees with the same latent trait value but differing in other characteristics may have differing probabilities of response. This tendency is referred to as DIF. In a complex sample, one might reasonably expect to find variation in parameter estimation across strata or sample clusters. For example, some items in a reading test may favour males, or residents in rural communities may respond differently to particular items on an attitudinal scale.

Assuming that DIF is present in an item or a series of items, the question then becomes how one can take its effects into account without introducing error into the measurement of the theorized latent trait. One option to control for it is to remove those items that exhibit DIF from a scale. When this happens, the effects on the latent variable due to the dropped item are not taken into account. Even though the item still contains potentially valuable response information to the analyst that information would not be used in predicting the latent trait. There does not seem to be any valid mathematical reason for forcing parameter invariance across groups. An alternate option is to treat the item as being different for each group. In this case, while the item exhibits a connection to the latent trait, its meaning within each group is to be interpreted differently. However, given the small sample sizes within strata or sample clusters, it may not be possible to accurately estimate the parameters for each group defined by the sample complexity. Furthermore, it may problematic to determine when parameters should be allowed to vary and by how much.

Item Drift

When repeated measurements of responses to items are recorded and analyzed for an assumed equivalent population, one would expect that the estimate of the item parameters derived over time would be relatively stable. However, it has been found that in some cases, the nature of the item, as illustrated by the value of its item parameters, is changing. While some of these differences can be attributed to simple sampling error, when item drift occurs, the changes in item characteristics are monotonic and often linear. For instance, some items would become easier with each testing period, or else they would become less discriminating over time. One needs to have a valid method to be able to adequately discriminate between unstable item functioning and true population changes.

Issues for Discussion on DIF and Drift

The problem with this definition is that other variables besides item “bias” contribute to these differences. Item-group interaction has to be featured into the description of item “bias”. When is a measure of DIF the result of an important population characteristic and when is it an indication of a poorly constructed item and how can one determine this?

Our understanding of DIF and the related DRIFT of item functioning through time has increased through the various discussions and we are able to refine the issues and clarify the process. The literature seems to suggest that item functioning is measured as a constant and one compares equivalent populations to measure dif. The reality is that analysis tends to focus on differences in sub-populations and the risk of poor item design is particularly important if it invalidates the findings. While group differential are natural expectations of analysis, differential item functioning is only in question when concomitant information among items is not present. In other words, and put more simply, while we can expect two groups to perform differently in measuring the latent trait, it is when an item bucks the trend set by most items that we can identify true dif or drift. This information, used in conjunction with proof of the unidimensionality assumption (if not respected, it may also create a similar effect) should be sufficient to identify true dif or drift.

Being able to identify DIF or DRIFT doesn’t resolve the issue of what to do in the presence of poor item design. Keeping the item will usually introduce instability and introduce error in the estimation function; however a loss of an item may still result in a reduced reliability. Keeping the item and treating it as different item by group may be a way to retain the information without destabilizing the estimation function. Ultimately a trade-off between reliability (error free estimates of the latent trait) and stability (variance of estimated parameters) will prevail in determining the right course of action.

Conclusions

It has been shown that some key aspects of the implementation of IRT methods in the NLSCY can be theoretically justified. First, the point estimates of item parameters and of the parameters of the ability distribution obtained using the survey weight option provided in the software package BILOG-MG are design-consistent. Second, the method used in the NLSCY to estimate individual abilities can be justified in the sense that it yields empirical Bayes predictors of latent ability. These predictors of ability can therefore be used to construct properly weighted estimates of aggregate measures of student ability. There are some caveats, however. First, the estimated variances of the point estimates of the fixed parameters are incorrect. Second individual Bayes predictors of latent ability are subject to a bias that depends upon the number of binary items included in each test. This defect results in a moderate degree of bias at the population aggregate level.

Many inferential problems remain, some of which stem from the interaction between IRT methods and the complex design of the NLSCY, while others are characteristic of IRT methods generally. The research focused upon a general investigation of these issues, but the specific research required on how to properly integrate IRT methods with complex survey analysis remains to be done. Some require mainly software development and have been initiated as part of the ongoing research. Others require more fundamental theoretical work and are of a longer term nature.

The research led to the many steps taken by Statistics Canada to apply psychometric techniques in a manner consistent with the complex survey design scenario, to derive appropriately measured items parameters in the construction of assessment tools and to estimate relevant latent ability scores for children to support longitudinal analyses.

While it is certain that IRT will continue to be used in surveys at Statistics Canada it is also clear that commercial software development has some catching up to do for measurements done in a complex survey design setting. While methods to compensate for the errors can easily be derived, they are difficult and laborious to implement in a production setting. At recent workshops and conferences, positive development has been shown that development in this area is forging ahead.

Statistics Canada will continue to engage its analytical and research community to explore issues related to IRT when it is applied to surveys with a complex design and in the longitudinal context. The NLSCY has already begun to investigate methods to expand the concept of latent measurement of ability through time. By combining theory of item analysis with re-scaling and standardization of measurements, new techniques and issues immerge in the interpretation and analysis of latent variable constructs. It is clear that IRT in complex surveys and longitudinal surveys is in development.

References

Ronald Hambleton and Hariharan Swaminathan's book *"Item Response Theory – Principles and Applications"*, 1985, Kluwer * Nijhoff Publishing.

Linda Crocker and James Algina, "Introduction to Classical & Modern Test Theory", 1986, Holt, Rinehart and Winston, Inc.

Mislevy, R.J. 1985. Estimating latent group effects. *Journal of the American Statistical Association, 80,* 993-997.

Mislevy, R.J. 1991. Randomization-based inferences about latent traits from complex samples. *Psychometrika, 56,* 177-196.

D. Roland Thomas, Ph.D, 2001, Item Response Theory and its Application to the National Longitudinal Survey of Children and Youth – Final Report, Internal Document, Statistics Canada

D. Roland Thomas, Carleton University, André Cyr, Statistics Canada, 2001 Applying Item Response Theory Methods to Complex Survey Data, SSC 2001 proceedings

D. Roland Thomas, Bruno D. Zumbo and Irene R.R. Lu, Carleton University, University of British Columbia, Canada, 2002, Modelling Survey Data for Social and Economic Research, Using IRT and Factor Scores in Regression and Other Analyses: A Review, SYMPOSIUM 2002 PROCEEDING, Catalogue no. 11-522-XCB