OECD*publishing*

Please cite this paper as:

Cartwright, F. (2012), "Technical Feasibility of Reporting YITS 2010 Skill Assessment Results on the PISA 2000 Reading Scale", *OECD Education Working Papers*, No. 69, OECD Publishing.
http://dx.doi.org/10.1787/5k9fhndspvf1-en

OECD Education Working Papers
No. 69

Technical Feasibility of Reporting YITS 2010 Skill Assessment Results on the PISA 2000 Reading Scale

Fernando Cartwright

OECD

Unclassified **EDU/WKP(2012)6**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **14-Feb-2012**

English - Or. English

DIRECTORATE FOR EDUCATION

EDU/WKP(2012)6
Unclassified

TECHNICAL FEASIBILITY OF REPORTING YITS 2010 SKILL ASSESSMENT RESULTS ON THE PISA 2000 READING SCALE

OECD Education Working Paper number 69

This paper was written by Fernando Cartwright of Statistics Canada.

Contact: Andreas Scheicher; E-mail: andreas.schleicher@oecd.org; Tel: +33 (0)1 45 24 93 66

JT03316025

Complete document available on OLIS in its original format

English - Or. English

OECD DIRECTORATE FOR EDUCATION

OECD EDUCATION WORKING PAPERS SERIES

TABLE OF CONTENTS

Tables

ABSTRACT

This study examines the feasibility of reporting scores of a test based on the Programme for International Student Assessment (PISA) 2000 instrument that was administered to a sample of 25-year-old Youth in Transition Survey (YITS) respondents on the PISA scale. Each of these respondents also participated in PISA 2000. The study examines the considerations for estimating proficiency estimates for the YITS 2010 sample and describes the methods recommended for analyzing the data. The results indicate that, despite much higher performance, there is no ceiling effect in the YITS 2010 sample for the PISA items. Although the estimated scores for the YITS 2010 sample should not be misconstrued as true 'PISA results,' there is no technical impediment to reporting them on the PISA scale and examining the differences between these results and the PISA 2000 results.

RÉSUMÉ

La présente étude examine la possibilité de rapporter sur l'échelle PISA les scores d'un test basé sur les items du cycle 2000 du Programme international pour le suivi des acquis des élèves (PISA) et administré à un échantillon de jeunes de 25 ans participant à l'Enquête auprès des jeunes en transition (EJET). Chacun de ces participants avait auparavant pris part à l'enquête PISA 2000. Cette étude examine les possibilités d'estimation du niveau de compétence de l'échantillon du cycle 2010 de l'EJET et décrit les méthodes recommandées pour l'analyse des données. Les résultats indiquent que, en dépit d'un niveau de performance bien plus élevé, il n'existe pas d'effet de plafonnement pour les items PISA dans l'échantillon du cycle 2010 de l'EJET. Bien que les scores estimés pour l'échantillon du cycle 2010 de l'EJET ne doivent pas être interprétés à tort comme de véritables « résultats du PISA », il n'existe pas d'obstacle technique à les rapporter sur l'échelle PISA et à les examiner en regard de ceux du cycle 2000 de l'enquête PISA.

TECHNICAL FEASIBILITY OF REPORTING YITS 2010 SKILL ASSESSMENT RESULTS ON THE PISA 2000 READING SCALE

In 2000, Canada participated in the first cycle of the Program for International Student Assessment (PISA). This survey was merged with the first cycle of the Canadian Youth in Transition Survey (YITS), a longitudinal survey intended to monitor the trajectories of youth as they make the transition from mandatory secondary education into post-secondary education and the workforce. Beginning in 2000, the survey participants are interviewed every two years, updating information about their personal, education and work history. In the most recent cycle of this survey, conducted in 2010, the survey participants were also administered a subset of the items from the PISA reading assessment that have also been used to link the PISA 2009 reading results to the PISA 2000 reading results. The goal of this reassessment was to provide information about skill development during this period on a scale comparable to that of the PISA 2000 reading assessment.

There are several challenges in describing reading proficiency of the now-25 year old participants on the reading scale established by PISA 2000. The first issue is the appropriateness of the items for the current population. Previous studies using PISA data have suggested that the effect of schooling on reading proficiency is approximately 33-35 PISA points per grade level (Willms, 2004; Fuchs & Woessmann, 2004). Although it is unlikely that this estimate will continue to be linear beyond the grades sampled by PISA 2000, after an additional two and a half years of public schooling and, for most students, additional post-secondary education, it is reasonable to expect the average performance of these participants to rise substantially from the value of 535 in 2000 in the subsequent decade. The major limitation of the PISA test items with respect to the measurement of reading skills of 24-year-olds in Canada is that they are intended to measure student proficiency across the wide range of the PISA scale. As a result, there are few items providing accurate measurement at the high extremes of the score range, where the expected scores of the YITS participants are expected to be at age 25. If the items selected for the reassessment are too easy for the 25-year-old population, respondents will have uniformly correct responses to items, and variations in estimated scores will be more influenced by random error than individual proficiency.

The second major issue is the consistency of item performance between the population of 15-year-olds in 2000 and 25-year-olds in 2010. Although all items generally measure a common dimension of reading proficiency, to some extent each item measures a specific manifestation of reading proficiency and requires a subtly distinct set of skills. For example, some items make use of different primary documents, from narrative prose to data entry forms, and use different information processing skills. The risk in applying a set of items to different populations is that comparability assumes that the skills required by all the items have the same relative presence or absence in the two populations, with respect to the general dimension of reading proficiency. If this is not the case, then the relative item difficulty will vary between the two populations and the common dimension of reading proficiency will have a different interpretation between the two populations.

Although the initial work in developing the PISA items and tests across diverse contexts of culture, language, and levels of reading proficiency suggests that the items used in the PISA reassessment will be robust to these issues, the population of the YITS 2010 survey is 10 years older than any individual previously assessed with the PISA items, and experiences beyond secondary education may introduce new factors affecting the relative difficulty of items.

The current paper details the procedures used to scale the item responses from the YITS PISA reassessment onto the PISA 2000 reading proficiency scale. The first section of analyses examines the robustness of items to the differences in population between the PISA 2000 administration and the YITS

2010 administration. The second section replicates the PISA 2000 scaling, conditioning and plausible value generation for the YITS participants. In the discussion of the results, I present a computationally simple method for examining the change in reading proficiency between the PISA 2000 and YITS 2010 assessments and the challenges with using change scores in secondary analysis.

Data

Respondents

In PISA/YITS 2000, the sample consisted of ~30,000 students selected from a two-stage nationally representative sample. The sample design first selected ~1,000 schools from a stratified frame that was stratified to represent each educational jurisdiction in Canada, as well as minority language and rural schools. Within each school, approximately 35 15-year-olds were randomly selected and assigned test booklets. Students and school administrators completed paper-and pencil questionnaires, and interviewers from Statistics Canada completed background questionnaires from parents. Over the course of the intervening 5 cycles of data collection, approximately 18,000 participants were dropped or lost from the longitudinal sample, resulting in approximately 11,000 continuous participants in Cycle 5. The YITS Cycle 5 responding sample was stratified into 12 strata according to gender, PISA reading level and education status and a random sample was selected within each stratum. In total approximately 2,000 Cycle 5 respondents were selected to participate in the YITS PISA reassessment. Of the selected respondents, 1,297 actually completed the assessment.

Population weights were calculated by Statistics Canada to adjust the representation of the current sample to the population represented by the PISA/YITS 2000 sample. The starting point for the creation of weights for YITS PISA reassessment was the final weight from YITS Cycle 5. To derive a final weight for PISA-R, the following adjustments were applied to initial weights of the individual records on the YITS PISA reassessment:

- Adjustment for Sub sampling of the Cycle 5 YITS Sample

  To select the YITS PISA reassessment sample, the YITS Cycle 5 responding sample was stratified into 12 strata and a random sample was selected within each strata. To adjust the initial weights to account for this sampling, the initial weight of each sampled unit in stratum h was multiplied by a factor equal to the number of Cycle 5 units in stratum h (N_h) divided by the number of units selected for the Reading Skills Reassessment sample in stratum h (n_h).

- Adjustment for Non-response

  To adjust the weights for non-response to the YITS PISA reassessment, logistic regression was used to estimate the expected probability of response for each sample unit. Modelling was done within region (Atlantic, Quebec, Ontario, Prairies, and British Columbia). To form response groups within which weight adjustments were to be made, the sample file was sorted by the estimated probability of response within each region. It was then divided into deciles, giving ten response adjustment groups for each region. Within each response adjustment group, the non-response adjustment factor was computed as the ratio of the sum of the weights for all units selected in the PISA-R sample to the sum of the weights for all responding sample units.

- Adjustment for Calibration to Cycle 5 Gender Totals

  To bring estimates for the YITS PISA reassessment in line with YITS Cycle 5 estimates a final calibration adjustment was made. Non-response adjusted weights were adjusted such that the survey weights sum to the same totals, by gender, as the Cycle 5 weights for all units in the Cycle 5 sample.

The final YITS PISA reassessment weight is the product of the initial weight multiplied by (1) the adjustment for subsampling of the YITS Cycle 5 sample, (2) the non-response adjustment, and (3) the calibration adjustment.

Although the participants in the YITS PISA reassessment tended to have more advantageous socio-economic conditions than the original PISA 2000 sample, population weights were calculated by Statistics Canada to adjust the representation of the current sample to the population represented by the PISA/YITS 2000 sample. These weights are used for all analyses presented in this study.

Items

In order to assess a wide breadth of content, the PISA 2000 assessment used a balanced incomplete block (BIB) design, which balances different item content across different test booklets, then randomly assigns different test booklets to different students. With the BIB design, items are grouped into blocks, and each block is repeated in more than one test booklet such that the content is balanced across test booklets.

The PISA/YITS reading assessment consisted of 28 items from PISA 2000, the same items used to link the PISA 2003, 2006 and 2009 reading results to the PISA 2000 scale. These items include 22 items that were assigned scores of 0 or 1 and six items that were assigned scores of 0, 1 or 2.

Table 1 summarises the characteristics of the PISA link items in terms of their item format, situational context, text type, text format, reading process (aspect) and PISA level.

In terms of the item format, the assessment consisted of the following: multiple choice (9 items), complex multiple choice (1 item), open constructed responses, (10 items), closed constructed responses (4 items), and short responses (4 items). These items include 22 items that were assigned scores of 0 (incorrect) or 1 (correct) and six items that were assigned scores of 0 (incorrect) or 1 (partially correct) or 2 (correct). Of the possible 34 items responses (28 items with correct responses, 6 item with partially correct responses), the distribution of responses across the 5 PISA reading levels defined in 2000 is shown in Table 2.

Table 1. Characteristics of PISA - R Items

Unit Item Code	Unit Name	Submitted By	Item Format	Situation	Text Type	Text Format	Reading Process	PISA Reading Level
R055Q01	Drugged Spiders	CITO	Multiple Choice	Public	Expository	Continuous	Interpreting	2
R055Q02	Drugged Spiders	CITO	Open Constructed Response	Public	Expository	Continuous	Reflecting and evaluating	3
R055Q03	Drugged Spiders	CITO	Open Constructed Response	Public	Expository	Continuous	Interpreting	3
R055Q05	Drugged Spiders	CITO	Open Constructed Response	Public	Expository	Continuous	Interpreting	2
R067Q01	Aesop	Greece	Multiple Choice	Personal	Narrative	Continuous	Interpreting	1
R067Q04	Aesop	Greece	Open Constructed Response	Personal	Narrative	Continuous	Reflecting and evaluating	2 (code 1): 4 (code 2)
R067Q05	Aesop	Greece	Open Constructed Response	Personal	Narrative	Continuous	Reflecting and evaluating	2 (code 1): 3 (code 2)
R102Q04A	Shirts	CITO	Open Constructed Response	Personal	Expository	Continuous	Interpreting	4
R102Q05	Shirts	CITO	Closed Constructed Response	Personal	Table	Non-continuous	Interpreting	4
R102Q07	Shirts	CITO	Multiple Choice	Personal	Expository	Continuous	Interpreting	1

Unit Item Code	Unit Name	Submitted By	Item Format	Situation	Text Type	Text Format	Reading Process	PISA Reading Level
R104Q01	Telephone	New	Closed Constructed Response	Public	Table	Non-continuous	Retrieving information	1
R104Q02	Telephone	New	Closed Constructed Response	Public	Table	Non-continuous	Retrieving information	4
R104Q05	Telephone	New	Short Response	Public	Table	Non-continuous	Retrieving information	4 (code 1): 6 (code 2)
R111Q01	Exchange	Finland	Multiple Choice	Educational	Expository	Continuous	Interpreting	2
R111Q02B	Exchange	Finland	Open Constructed Response	Educational	Expository	Continuous	Reflecting and evaluating	3 (code 1): 5 (code 2)
R111Q06B	Exchange	Finland	Open Constructed Response	Educational	Expository	Continuous	Reflecting and evaluating	3 (code 1): 4 (code 2)
R219Q01T	Employment	IALS	Closed Constructed Response	Occupational	Form	Non-continuous	Retrieving information	3
R219Q01E	Employment	IALS	Short Response	Occupational	Form	Non-continuous	Interpreting	2
R219Q02	Employment	IALS	Open Constructed Response	Occupational	Form	Non-continuous	Reflecting and evaluating	1
R220Q01	South Pole	France	Short Response	Educational	Map	Non-continuous	Reflecting and evaluating	4
R220Q02B	South Pole	France	Multiple Choice	Educational	Chart/Graph	Non-continuous	Interpreting	3
R220Q04	South Pole	France	Multiple Choice	Educational	Expository	Continuous	Interpreting	3

Unit Item Code	Unit Name	Submitted By	Item Format	Situation	Text Type	Text Format	Reading Process	PISA Reading Level
R220Q05	South Pole	France	Multiple Choice	Educational	Expository	Continuous	Interpreting	1
R220Q06	South Pole	France	Multiple Choice	Educational	Expository	Continuous	Interpreting	2
R227Q01	Optician	Switzerland	Multiple Choice	Occupational	Descriptive	Continuous	Retrieving information	3
R227Q02	Optician	Switzerland	Complex Multiple Choice	Occupational	Descriptive	Continuous	Retrieving information	2 (code 1): 4 (code 2)
R227Q03	Optician	Switzerland	Open Constructed Response	Occupational	Descriptive	Continuous	Reflecting and evaluating	3
R227Q06	Optician	Switzerland	Short Response	Occupational	Chart/Graph	Non-continuous	Retrieving information	2

Table 2. Distribution of PISA - R items by PISA Reading Level

PISA Reading Level	Number of item responses
Level 1	5
Level 2	8
Level 3	11
Level 4	8
Level 5	2

Unlike the original PISA test design, the PISA/YITS skill reassessment consisted of one booklet comprised of all 28 items. Items were grouped into two clusters (cluster 1 followed by cluster 2) and within each cluster items appeared in the same order as they appeared as linked clusters in the PISA 2009 design. Respondents had 60 minutes to complete the booklet.

Missing responses to items were given distinct codes and treated differently, depending on their position within each individual's response vector. An individual's response vector was defined as the contiguous items from the first item in the booklet to the first item following the last item with a valid response. Consistent with the practice from PISA, missing responses that are internal to the set of items with valid responses were treated as incorrect. Items that were outside of the set of valid item responses were coded as not-reached. The item response model parameters and percent correct statistics for these items from PISA 2000 are presented in Table 3.

Table 3. Statistical properties of PISA 2000 link items

Test position	ID	Percent Correct	Difficulty (δ)	τ_1	τ_2
1	R227Q01	57.65	0.196	0	0
2	R227Q02T	59.58	0.045	-1.008	1.008
3	R227Q03	55.58	0.295	0	0
4	R227Q06	74.29	-0.916	0	0
5	R111Q01	63.87	-0.053	0	0
6	R111Q02B	34.14	1.365	-0.554	0.554
7	R111Q06B	44.42	0.808	0.828	-0.828
8	R055Q01	83.79	-1.377	0	0
9	R055Q02	52.93	0.496	0	0
10	R055Q03	60.57	0.067	0	0
11	R055Q05	77.45	-0.877	0	0
12	R104Q01	82.63	-1.235	0	0
13	R104Q02	41.3	1.105	0	0
14	R104Q05	28.89	1.875	-0.914	0.914
15	R219Q01T	57.37	0.278	0	0
16	R219Q01E	69.94	-0.55	0	0
17	R219Q02	76.24	-0.917	0	0
18	R067Q01	88.35	-1.726	0	0
19	R067Q04	54.31	0.516	-0.456	0.456
20	R067Q05	62.48	0.182	0.482	-0.482
21	R102Q04A	36.00	1.206	0	0
22	R102Q05	41.8	0.905	0	0
23	R102Q07	85.23	-1.566	0	0
24	R220Q01	46.03	0.785	0	0
25	R220Q02B	64.49	-0.144	0	0
26	R220Q04	60.67	0.163	0	0
27	R220Q05	84.88	-1.599	0	0
28	R220Q06	65.54	-0.172	0	0

Conditioning Variables

As detailed in the PISA 2000 Technical report (Adams and Wu, 2002), the likelihood functions based on individual response vectors are conditioned by the predicted values of a regression function that models the proficiency outcome with the set of principal component scores representing the majority of the variance in the questionnaire items and sample stratification. For the current study, the conditioning model used by PISA 2000 could not be perfectly replicated for several reasons, notably the inapplicability of the regression model to the current population and the substantially limited number of corollary variables available from the questionnaire. As a result, the conditioning procedure was replicated in the current study using a smaller set of variables. From the YITS reassessment, a smaller set of conditioning variables were included in the conditioning process. These variables, which were used to derive principal component scores, following the methodology established for PISA, are described in Table 4.

Missing data were replaced by the sample mean for the purpose of estimating principal component scores.

Table 4. Conditioning variables from YITS 2010

ID	Variable	Level of measurement
ACENG	academic engagement	Interval (Scale)
BIRTH	country of birth	Interval (dichotomous)
FBIRTH	father's country of birth	Interval (dichotomous)
FED	father's tertiary education status	Interval (dichotomous)
GRADE	approximate mark in current education	Interval (Scale)
LANGDIFF	Home language!=school language	Interval (dichotomous)
LANGHOME	language at home	Interval (dichotomous)
MBIRTH	mother's country of birth	Interval (dichotomous)
MED	mother's tertiary education status	Interval (dichotomous)
PROG	school program	Quasi-interval (ordinal)
PROVLANG	province and school language	Nominal
PSED	post-secondary education	Nominal
SES	parental socio-economic status	Interval (scale)
SEX	sex	Interval (dichotomous)

Methods

Exploratory analysis

The first stage of data treatment involved the scoring of item responses. Selected-response items were scored separately from open-ended responses. Open-ended responses were scored by trained raters. All items were scored consistently with the item scoring manual provided by the PISA consortium (see Appendix A). All item scorers received the same training and were managed by the same team responsible for scoring the PISA assessments.

The analysis of response data begins with an evaluation of the ability of items and the test overall to meaningfully discriminate between individuals with different levels of reading proficiency. Due to the idiosyncratic nature of the current sample, this analysis used the classical test statistics of item facility and point-biserial correlation to evaluate item suitability. These statistics are used due to the robustness with which they are estimated in small samples with non-normally distributed sample distributions compared to item response model parameters. Item facility is simply the weighted proportion of respondents who correctly responded to each item. The point-biserial correlation is the Pearson product moment correlation between the score of individuals on each item with the raw test score, calculated from all items except for the item currently under analysis.

Another measure of the appropriateness of the test administration is the degree to which respondent fatigue plays a role in the item responses. Respondent fatigue is indicated by order effects in either item difficulty or the frequency of missing item responses. If items that appear later in the assessment have lower item facility or are more likely to have missing responses, respondent fatigue may reduce the validity of the results. If respondent fatigue plays a role in the missingness of item responses, not-reached responses must be scored as randomly missing in order to apply the PISA 2000 item parameters to calculate test scores.

To evaluate the impact of respondent fatigue on the item responses, the item facilities were calculated twice: first by treating the not-reached items as incorrect, and second by treating them as randomly missing. To detect order effects, the item facilities from each treatment were regressed first on the item facilities from PISA 2000 and then on item position in the YITS assessment. If item fatigue is affecting item responses, difficult items and items appearing later in the test would be more difficult than expected. These exploratory analyses will determine the treatment of not-reached items for the remaining analyses.

Estimation of Reading Proficiency

Following the methodology established in PISA 2000 for producing proficiency estimates (Adams, 2002), posterior likelihood functions are estimated by the cumulative product of the likelihood functions based on each item response function and the prior distribution estimated based on the item facilities. The item response model used in PISA expressed in equation (1) describes the probability, P_{xi}, of a respondent, n, producing a specific score, $x = \{0, 1, 2\}$, for item i, given the threshold parameters of each score, τ_{xi}, and the overall item difficulty, δ_i, as a function of the respondent's proficiency, β_n.

$$P_{xi}\left(\beta_n\right)=\frac{\exp\sum_{j=0}^{x_i}\left(\beta_n-\left(\delta_{i\bullet}-\tau_{ij}\right)\right)}{\sum_{h=0}^{2}\exp\sum_{j=0}^{h}\left(\beta_n-\left(\delta_{i\bullet}-\tau_{ij}\right)\right)},\quad x_i=0,1,2 \qquad (1)$$

Given an observed response, x, the equation (1) describes the likelihood of the response being produced at any β value. The product of all likelihood functions corresponding to an individual's set of item responses, describes the likelihood of any particular level of proficiency having produced the observed set of responses. The likelihood function for an individual can be combined with information about the distribution of proficiency from which the individual is randomly sampled, a *prior distribution*, to produce a *posterior distribution* that describes the probability of the individual's proficiency. Plausible values are then drawn from the posterior distribution following the PISA 2000 methodology. The accuracy with which an individual's proficiency is described depends on the variance of the posterior distribution. This variance, in turn, depends on the variance of the number of test items and the appropriateness of the prior distribution.

Prior distribution. Accurate estimates of the prior distribution are particularly critical in the current study, because there are relatively few items, and the available items were designed for a substantially different population. As a result, the information available from the items is much more limited than the information available from the PISA 2000 assessment. A standard practice in IRT scaling is to use the marginal likelihood function of the full sample as the prior proficiency distribution during the final scaling. However, if the items provide more accuracy at the lower end of the sample proficiency distribution than the higher end of the distribution, both the individual likelihood functions and the marginal likelihood, will be asymmetric towards the upper end of the scale. This asymmetry means that plausible values will be more likely to be drawn from the higher range of proficiency due to the limited selection of test items rather than the proficiency of the individual. I use a symmetric distribution as the prior to minimise the effects of this bias. The mean and standard deviation of the prior are estimated from the item facilities, rather than the item-response-based likelihood functions, which avoids the influence of bias resulting from asymmetric item information.

There are two stages in the calculation of the prior distribution parameters. First, using the estimated item facilities from the full PISA 2000 administration and the YITS 2010 sample, I calculated interim estimates of the first two moments of the latent scale distribution of the YITS sample on the PISA 2000 reading proficiency scale. Second, interim proficiency estimates are used to estimate a conditioning model. The conditioning model produces a prior distribution that is unique to each respondent's vector of corollary information. This unique prior distribution is used for the final plausible value generation.

Calculation of the first stage of estimates followed a procedure to maximise the fit between the observed YITS 2010 item facilities and estimates of item facility based on the PISA 2000 item parameters. Equation (2) estimates the facility of an item, $P_{ij,}*$, for a given population, j, by integrating across the β distribution of that population, $P(\beta_j)$.

$$P_{ij}* = \int \sum_{k=0}^{K} kP_{ik} P_j d\beta \qquad (2)$$

In order to estimate the first two moments of the YITS 2010 distribution, I assume that $P(\beta_j)$ in both the PISA and YITS samples follows a parametric Gaussian function, with mean, μ_j, and standard deviation, σ_j.

$$P_j\left(\beta\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} Exp\left(-\frac{\left(\beta-\mu_j\right)^2}{2\sigma_j^2}\right) \qquad (3)$$

Estimates of the mean and standard deviation of the YITS 2010 population can be estimated by minimising the following equation across μ_j* and σ_j* using iterated numerical integration:

$$f\left(\mu_j^*, \sigma_j^*\right) = \sum_i \left(P_{ij}^* - P_{ij}\right)^2. \qquad (4)$$

Equation (4) is the squared difference between the observed item facility in the YITS 2010 sample and the item facility that is estimated by solving Equation (2), summed across all items. The variables μ_j* and σ_j* that solve the minimisation of Equation (4) represent the central moments of the YITS 2000 β distribution. The resulting prior distribution is used as a common prior distribution for the first stage of estimates.

For validation, this method is replicated on the PISA 2000 sample. Although the PISA reading scale has been arbitrarily set to have a population mean of 500 and standard deviation of 100, the parameters of the β distribution were produced in the PISA 2000 sample using an item-centered logit scale. The PISA scale was established by transforming the logit scale through two steps. First, the logit scale used to calibrate the items is normalised to a population-based standard normal scale (z-scale) by subtracting the calibration population mean and then dividing by the calibration population standard deviation. The z-scale is then linearly transformed through multiplying by 100 and adding 500. The equation used to perform this transformation for any item or person location on the logit scale, β, to its corresponding value on the PISA scale is

$$PISA = (((\beta - (0.5076))/1.1002)*100)+500 \qquad (5)$$

where the term, 0.5076 is the population mean and 1.1002 is the population standard deviation on the original logit scale (OECD, 2001). These parameters are used to produce item facility estimates for PISA 2000 and compared to the observed item facilities.

Plausible value estimation. The conditioning model was estimated by drawing a single plausible value from the first-stage posterior distribution based on the common prior. This value is used as the outcome in a regression model, where the predictors are the principal component scores derived from the variables in Table 4. The scaling procedure produces new posterior distributions by applying unique priors generated from the predicted value of the regression model for each respondent and, finally, drawing plausible values from the resulting posterior density functions (Adams, 2002).

Compared to the 141 reading items used to estimate the integrals of parameter distributions for PISA 2000, the YITS 2010 assessment uses only 28 items. To accommodate the greater measurement errors in the current sample that result from fewer items; misfit between the distribution of proficiency around which the assessment was designed and the distribution in the current sample; and the smaller and more idiosyncratic sample, I draw ten plausible values for each respondent, rather than the five values used for PISA 2000. Although the larger number of plausible values cannot correct for any potential model misfit, they provide increased accuracy in estimating the numerical integrals of the parameter distributions for secondary analyses (Rubin, 1987), and the increased computational cost in dealing with them is negligible. Efficiency is asymptotic, so an even larger number of plausible values would not result in noticeably larger increases in accuracy.

Results

Exploratory Analysis

Table 5 lists the results of the exploratory analyses produced by minimising equation (4). In Table 5, each row represents a single item. The columns under the Percent Correct heading describe the observed percentage of items in the YITS 2010 sample that were answered correctly. The Not Applicable column describes the percent-correct scores that are calculated by treating the items with missing responses as missing-at-random so they do not affect the calculations. The Incorrect column describes the percent-correct scores that are calculated by treating the missing responses as incorrect. The columns under the Quadratic Residual heading describe the residual of the estimated percent-correct values for each item from a quadratic line-of-best-fit between the PISA 2000 and YITS 2010 percent-correct statistics (illustrated in Figure 1), and the columns under the Estimated Percent Correct heading contain the estimated values themselves. The final column, Not Reached, simply describes the percentage of respondents who did not respond to each item.

Table 5. Item statistics for assessment items

ID	Percent Correct		Quadratic Residual		Estimated Percent Correct		Not reached
	Not Applicable	Incorrect	Not Applicable	Incorrect	PISA2000*	YITS2010*	
R227Q01	69.237	69.237	8.214	6.479	56.215	77.919	0.000
R227Q02T	71.049	71.049	8.052	6.259	59.177	80.162	0.000
R227Q03	78.951	78.951	-3.338	-5.011	54.250	76.361	0.000
R227Q06	85.197	85.197	4.473	2.237	76.062	90.715	0.000
R111Q01	86.662	86.662	-4.112	-6.034	61.070	81.530	0.000
R111Q02B	60.100	60.100	-7.653	-8.680	33.388	55.839	0.000
R111Q06B	72.745	72.745	-8.249	-9.586	44.004	67.251	0.000
R055Q01	94.680	94.680	-0.071	-2.593	82.409	93.791	0.000
R055Q02	70.779	70.779	2.380	0.787	50.231	72.990	0.000
R055Q03	83.423	83.423	-3.500	-5.323	58.749	79.845	0.000
R055Q05	92.059	92.059	-0.582	-2.913	75.466	90.402	0.000
R104Q01	94.517	94.372	-0.432	-2.773	80.592	92.957	0.154
R104Q02	52.664	52.583	8.358	7.196	38.224	61.310	0.154
R104Q05	42.813	42.714	2.813	2.042	24.797	44.843	0.231
R219Q01T	89.164	88.820	-11.957	-13.341	54.589	76.633	0.386
R219Q01E	87.539	87.201	-0.625	-2.393	70.140	87.422	0.386
R219Q02	95.430	94.988	-4.626	-6.479	76.077	90.722	0.463
R067Q01	96.034	95.220	0.420	-1.425	86.351	95.476	0.848
R067Q04	75.039	74.287	-0.588	-1.470	49.831	72.640	1.002
R067Q05	86.135	84.541	-4.670	-4.956	56.492	78.134	1.850
R102Q04A	52.174	49.961	2.582	3.711	36.316	59.208	4.241
R102Q05	63.900	61.141	-2.310	-0.809	42.094	65.355	4.318
R102Q07	97.417	93.061	-2.189	-0.398	84.635	94.763	4.472
R220Q01	65.112	60.293	1.113	4.547	44.460	67.693	7.402
R220Q02B	87.130	79.337	-4.106	1.745	62.802	82.739	8.944
R220Q04	72.184	65.227	7.822	12.952	56.866	78.423	9.638
R220Q05	93.745	84.348	1.337	8.178	85.002	94.918	10.023
R220Q06	72.405	64.534	11.406	17.304	63.330	83.099	10.871

The item facilities (expressed in percentage) from the YITS 2010 sample, treating not-reached items as not applicable and incorrect, are plotted against the PISA 2000 item facilities in Figure 1. The expected relationship between the YITS estimates and the PISA estimates is nonlinear, due to the different proficiency distributions and the upper and lower bounds of 0 and 1. This nonlinearity is indicated by the lines-of-best fit in Figure 1 (the dashed line represents the best fit for the incorrect not-reached treatment).

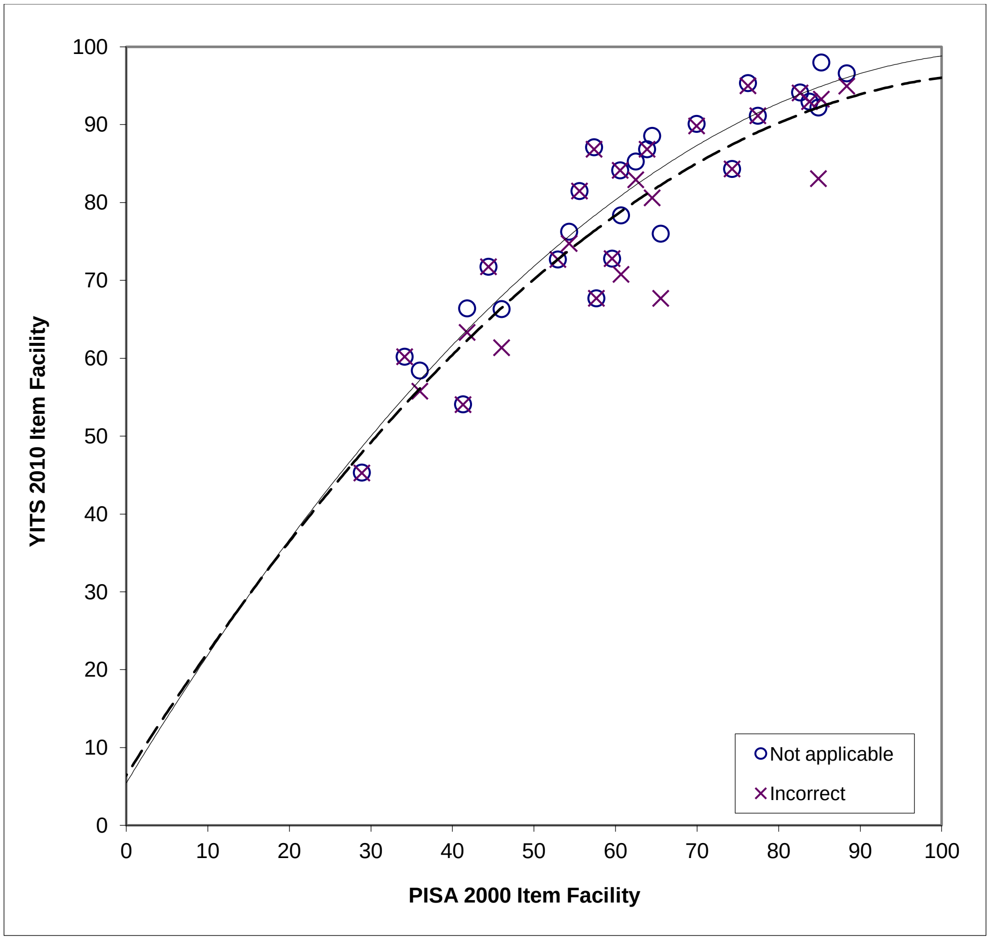

Figure 1. Scatterplot of PISA 2000 item facilities against YITS 2010 item facilities with different not-reach treatments

The residuals, $P_i - \hat{P}_i$, of the nonlinear relationships in Figure 1 are plotted against the PISA 2000 item facilities (expressed in percentage) in Figure 2. Positive values indicate higher facilities in PISA, and negative values indicate lower facilities in PISA. In general, the 'not applicable' treatment provides a better fit to the PISA 2000 item facilities than the 'incorrect' treatment. Even accounting for the generally lower average, three items (R220Q04, R22Q05 and R220Q06) were notably more divergent with the incorrect treatment. Item R219Q01T was distinctly less difficult than other items. Further examination of the content of this item suggests that these differences are the result of explainable differences in the populations (R219Q01T relates to filling out an employment application, with which 25-year-olds will be more familiar than 15-year-olds). Even though the item facilities are much higher with the YITS sample, the lack of bias at the upper end suggests restricted range is not artefactually diminishing the variances of item scores. In other words, there is no ceiling effect for the YITS 2010 sample.

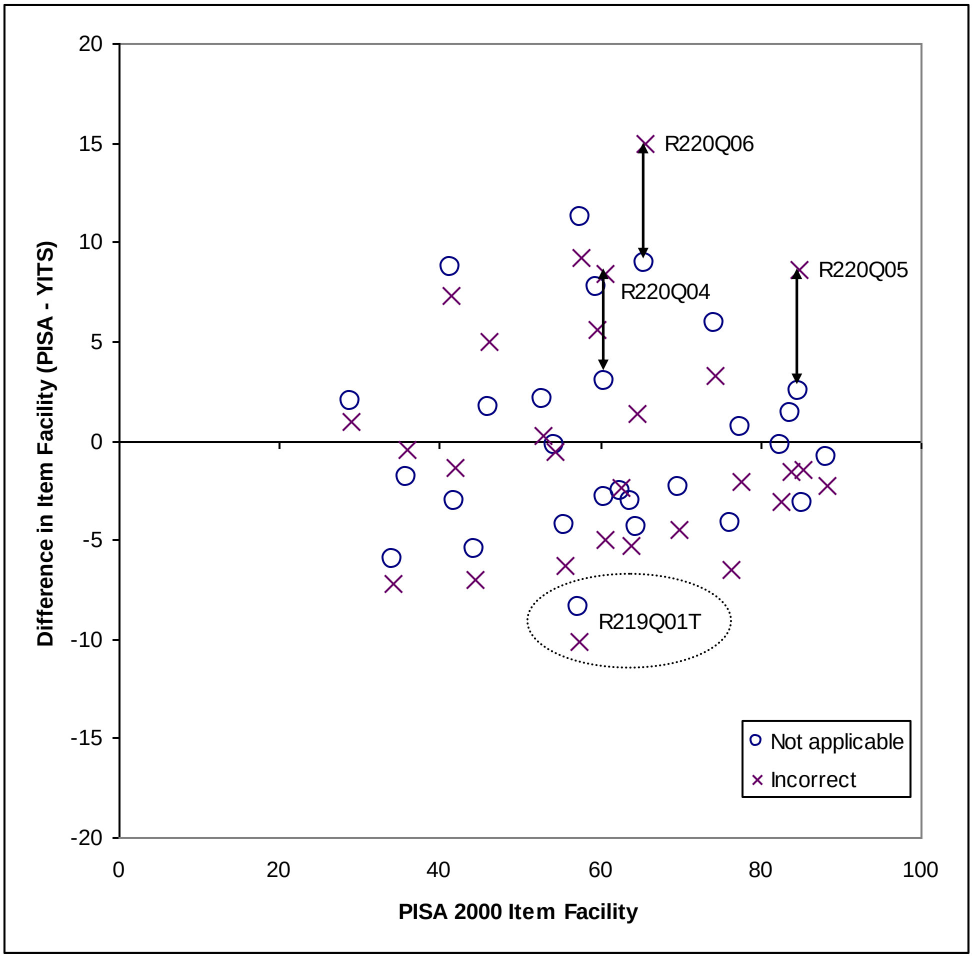

Figure 2. Residuals in item facility plotted against PISA 2000 item facility.

Many of the larger residuals can also be explained by item position, which is illustrated in Figure 3. Although there is a slight order effect with both assessments, this effect is more pronounced for the incorrect treatment (the dashed line in Figure 1). The three discrepant items are the last three items on the test form, and last item has the largest discrepancy (and is the only item which is notably more difficult for the YITS 2010). These results suggest that not-reached items are more likely a function of testing conditions rather than individuals' proficiency and should be treated as missing at random so that they do not bias the estimates of proficiency.

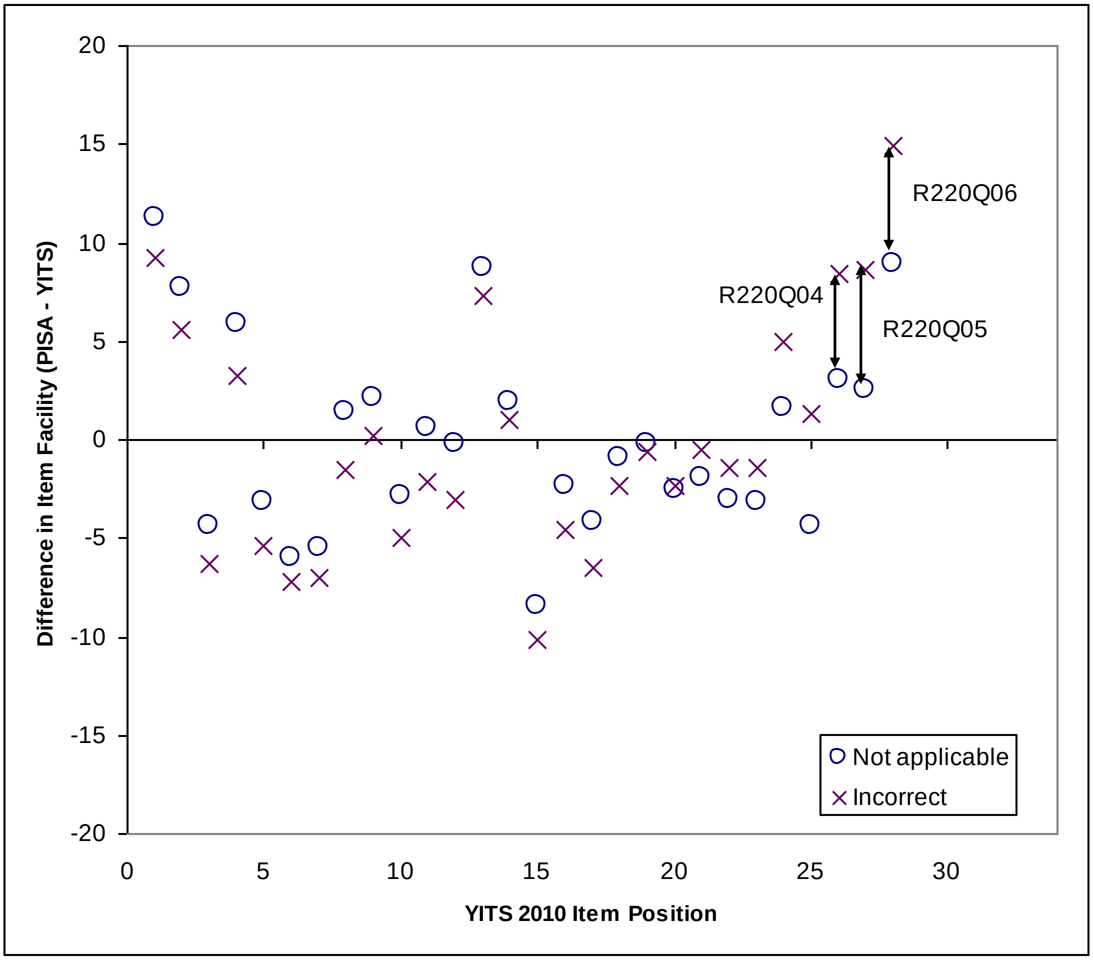

Figure 3. Residuals in item facility plotted against their position in the YITS assessment.

These exploratory results suggest that the PISA items function consistently between the PISA 2000 administrations and the YITS 2010 administrations if the not-reached items are scored as not applicable. Although the difficulty of R220Q06 may be the result of test-specific order effects, the relative cost of using sample-specific item parameters from the idiosyncratic YITS sample is greater than the cost of possible misfit of the PISA 2000 item parameter, given the goal of expressing reading on the PISA 2000 scale with so few items. Therefore, estimation of reading proficiency for YITS 2010 uses the item parameters from PISA 2000, treating not-reached items as not applicable.

Estimation of Prior Distributions

The initial estimates of the YITS reading proficiency mean and standard deviation based on minimising equation (4) are

$\mu^*_{YITS} = 1.678$, and

$\sigma^*_{YITS} = 0.805$.

These results are consistent with the expectation that, relative to the international PISA 2000 population, the YITS 2010 population would have a higher average and smaller variance. To validate the methodology, the procedure is reversed using the known central moments of the

PISA 2000 population to estimate item facilities for PISA 2000. The comparison of these estimates to the observed PISA 2000 item facilities, as well as the estimates for YITS 2010, is illustrated in Figure 4. The items are distributed evenly around the 1:1 line, indicating that there is no bias in the recovery of the item facilities from the procedure. As expected, there is larger random error in the YITS 2010 estimates, due to the smaller sample. However, the correlation between the observed and expected item facilities is still very high (0.93). This correlation is similar in magnitude to the correlation between facilities for these items as they are used to link results of adjacent PISA cycles (OECD, 2009, p 238-239). The replication of this methodology with known parameters of PISA 2000 indicates an almost perfect recovery of the population item facilities, with a correlation greater than 0.99.

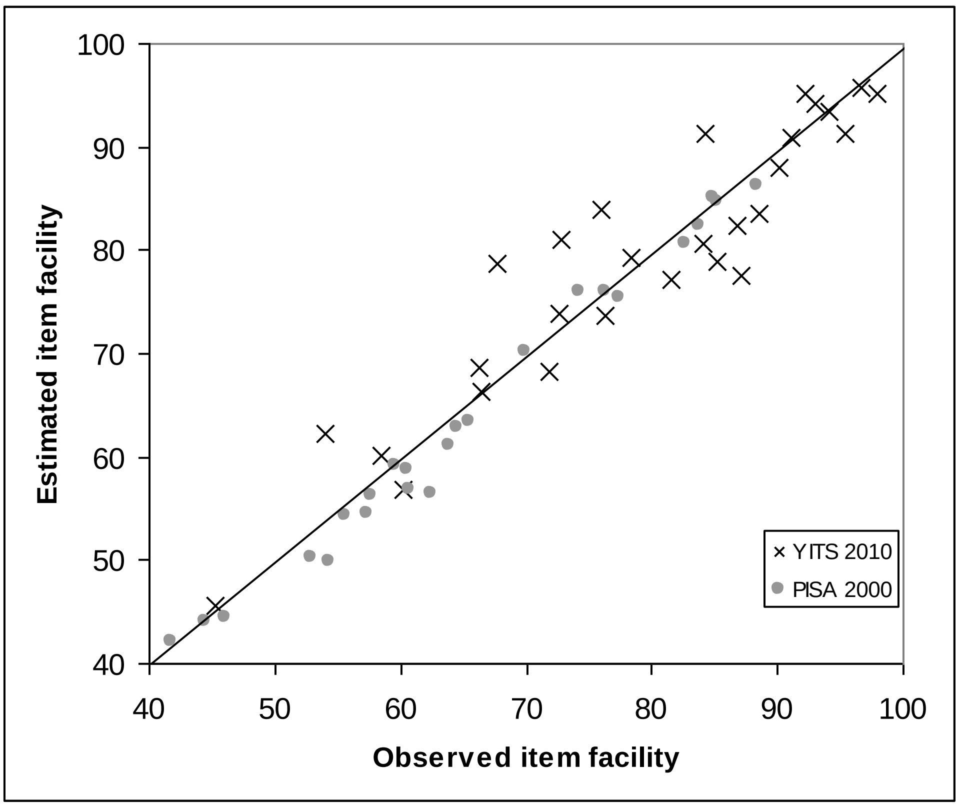

Figure 4. Estimated item facility plotted against observed item facility.

Figure 5 illustrates the prior distribution (solid line) based on the estimates of μ_j and σ_j and the marginal distribution (dashed line) estimated by the item response-based likelihood functions alone. The positive bias in the marginal distribution is produced by the asymmetric measurement of the test items relative to the YITS 2010 population. This bias is largely removed in the marginal posterior distribution (dotted line). The individual posterior distributions are the source of the interim plausible values used in the estimation of the conditioning model.

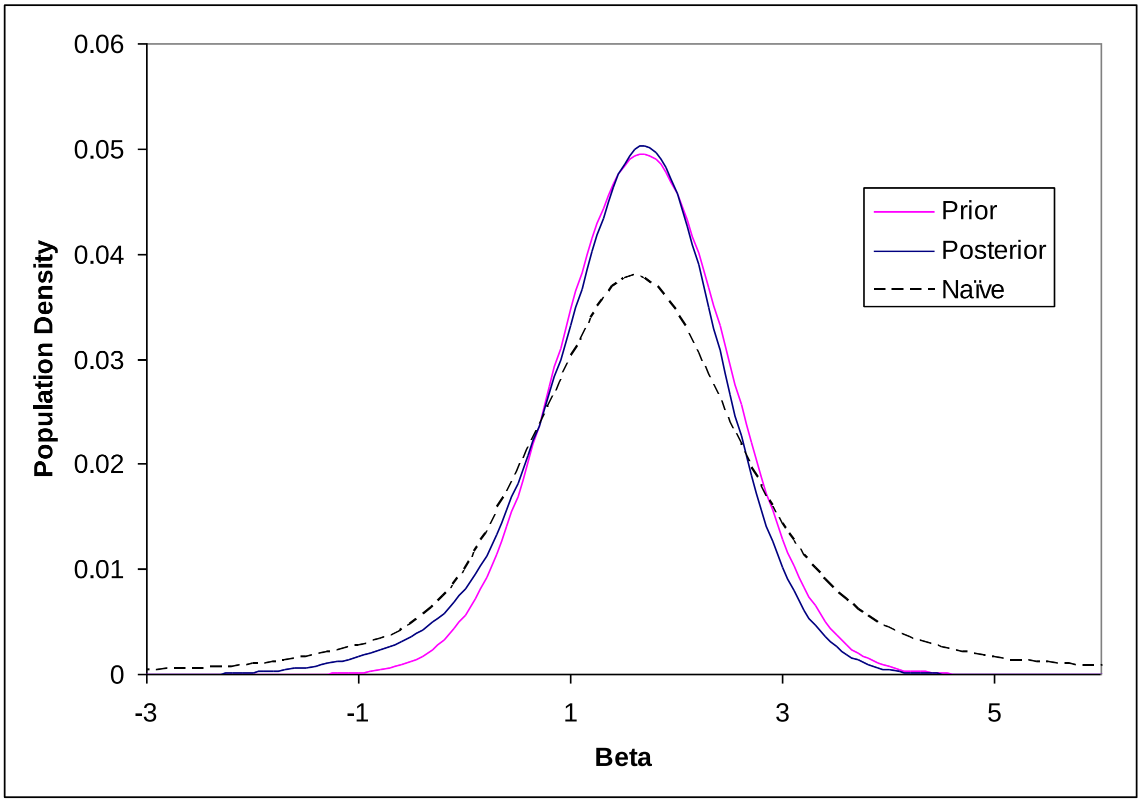

Figure 5. Estimated population distributions for YITS 2010 reading proficiency.

Conditioning Model

The first stage of analysis uses the interim plausible values, one for each respondent, as the outcome in a regression model. The predictors in the regression model are the principal component scores estimated summarising the set of variables in Table 4 as well as the interim plausible value (pv1). Sets of dummy-coded variables that represent a single categorical variable have the same variable name and are subscripted with the dummy code test value.

Table 6. Descriptive statistics for conditioning variables

Variable	N	n	Mean	Standard Deviation
ACENG	344989	1297	0.160	1.601
BIRTH	338343	1277	1.115	0.319
FBIRTH	337226	1272	1.294	0.456
FED	323846	1229	1.480	0.500
GRADE	312958	1168	2.895	1.155
LANGDIFF	344989	1297	0.107	0.309
LANGHOME	335983	1270	1.128	0.334
MBIRTH	338855	1280	1.266	0.442
MED	329172	1248	1.469	0.499
PROG	306436	1146	3.913	1.103
PROVLANG13	344989	1297	0.005	0.068
PROVLANG21	344989	1297	0.039	0.195
PROVLANG22	344989	1297	0.001	0.024
PROVLANG3	344989	1297	0.024	0.153
PROVLANG31	344989	1297	0.017	0.129
PROVLANG32	344989	1297	0.009	0.094
PROVLANG41	344989	1297	0.020	0.138
PROVLANG42	344989	1297	0.201	0.401
PROVLANG51	344989	1297	0.350	0.477
PROVLANG52	344989	1297	0.022	0.146
PROVLANG61	344989	1297	0.040	0.197
PROVLANG62	344989	1297	0.001	0.038
PROVLANG73	344989	1297	0.038	0.192
PROVLANG83	344989	1297	0.104	0.305
PROVLANG93	344989	1297	0.129	0.335
PSED10	344989	1297	0.007	0.081
PSED11	344989	1297	0.009	0.092
PSED12	344989	1297	0.016	0.125
PSED13	344989	1297	0.000	0.016
PSED20	344989	1297	0.007	0.081
PSED23	344989	1297	0.003	0.051
PSED3	344989	1297	0.023	0.151
PSED4	344989	1297	0.005	0.071
PSED5	344989	1297	0.269	0.443
PSED6	344989	1297	0.015	0.121
PSED7	344989	1297	0.005	0.068
PSED8	344989	1297	0.023	0.150
PSED9	344989	1297	0.394	0.489
PSED96	344989	1297	0.210	0.407
PSED99	344989	1297	0.016	0.126
SES	331135	1264	0.249	0.842
SEX	340957	1289	1.505	0.500
pv1	344989	1297	1.669	0.866

In consideration of the smaller number of variables available in the current study, the criterion for principal component extraction is 99% of the total variance, which allows for retention of variance from relatively unique variables in the set. This is greater than the criterion of 90% used by PISA 2000. Figure 6 displays the eigenvalues for each principal component extracted from the variable matrix.

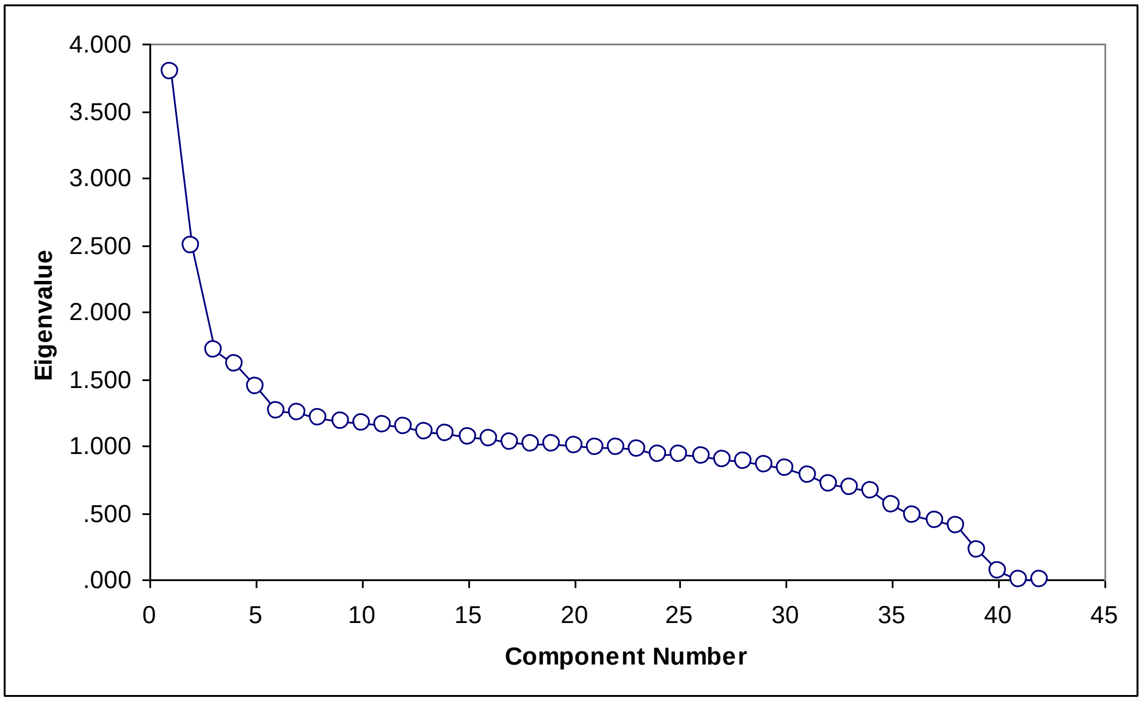

Figure 6. Eigenvalue plot of components extracted from conditioning variables.

I used the Anderson-Rubin method to compute the principal component scores in order to simplify the specification of the regression model. This method produces standardised orthogonal scores, which implies that any possible interaction term between predictors would have a null co-efficient and all co-efficients are expressible using the same number of significant figures to reduce the computational errors of rounding with 32-bit floating point arithmetic. Table 7 summarises the parameter estimates of the resulting regression model. This model explained 25.3% of the variation in reading proficiency (represented by the single plausible value). These model parameters were used to calculate a proficiency estimate for each respondent, based on the values of their corollary information, rather than their item responses. These estimates defined the means of the prior distribution used to condition the item-response-based likelihood functions for the final score estimation. Each prior used the same standard deviation as the uniform prior.

Table 7. Regressiong co-efficients for conditioning model

Co-efficient	B	(Standard Error)
Conditioned mean	1.624	(0.001)
Component score 1	-0.033	(0.001)
Component score 2	0.011	(0.001)
Component score 3	0.111	(0.001)
Component score 4	-0.203	(0.001)
Component score 5	0.011	(0.001)
Component score 6	0.094	(0.001)
Component score 7	0.023	(0.001)
Component score 8	-0.024	(0.001)
Component score 9	0.034	(0.001)
Component score 10	-0.018	(0.001)
Component score 11	-0.002	(0.001)
Component score 12	-0.041	(0.001)
Component score 13	-0.018	(0.001)
Component score 14	-0.012	(0.001)
Component score 15	0.010	(0.001)
Component score 16	0.047	(0.001)
Component score 17	-0.064	(0.001)
Component score 18	-0.041	(0.001)
Component score 19	0.012	(0.001)
Component score 20	0.010	(0.001)
Component score 21	0.049	(0.001)
Component score 22	0.048	(0.001)
Component score 23	-0.022	(0.001)
Component score 24	-0.015	(0.001)
Component score 25	0.024	(0.001)
Component score 26	0.000	(0.001)
Component score 27	-0.019	(0.001)
Component score 28	-0.032	(0.001)
Component score 29	0.002	(0.001)
Component score 30	-0.231	(0.001)
Component score 31	-0.058	(0.001)
Component score 32	0.009	(0.001)
Component score 33	0.007	(0.001)
Component score 34	-0.089	(0.001)
Component score 35	0.007	(0.001)
Component score 36	0.007	(0.001)
Component score 37	0.163	(0.001)
Component score 38	-0.034	(0.001)

I draw a total of ten plausible values for each respondent from the posterior distributions of each respondent. Applying equation (5) to these estimates converts the results to the PISA 2000 reading scale. Table 8 describes the distributional properties of each of the sets of values. Due to the smaller sample size and greater measurement error in the YITS 2010 assessment, the variance between the sets for any statistic is greater than was the case for the same population in PISA 2000.

Table 8. Distribution of plausible values

Plausible Value	Minimum	Maximum	Mean	Standard Deviation
pv1	135.744	863.214	596.102	83.717
pv2	226.668	846.619	595.989	86.305
pv3	154.454	845.147	598.051	83.151
pv4	135.977	873.094	598.930	84.669
pv5	135.754	963.058	598.525	88.609
pv6	254.082	854.252	596.883	82.513
pv7	135.921	864.795	597.015	87.026
pv8	254.089	863.878	600.008	81.924
pv9	199.404	890.209	599.067	85.833
pv10	144.833	899.523	597.176	85.164

Discussion

This study focused on the evaluation of the PISA 2000 items in the context of the YITS 2010 reading assessment. Based on the results of this analysis, there is no impediment to using the PISA 2000 item parameters to calculate scores for the YITS 2010 respondents and express the reading proficiency estimates on the PISA 2000 scale. In fact, the test items performed much better than expected, with no observed ceiling effect or evidence of differential item functioning between the two populations. This stability of the instrument over time is most likely the result of the intensive international vetting of items during the creation of PISA 2000. If items perform consistently across the diversity of cultural contexts and proficiency range in the OECD, it appears that they are equally stable across equivalent ranges of proficiency across age groups and school experience.

However, there are several limitations in the interpretation of these estimates. Firstly, any interpretation of results should consider that the YITS subset of items does not constitute the complete PISA assessment. The set of items used in PISA 2000 provides more accurate and broader information about reading proficiency than the subset of items used in YITS. Although the items used in the YITS 2010 assessment behave similarly in both YITS and PISA, the same may not be true for all items from PISA 2000. Thus, any analysis should make the distinction that the results of YITS are expressed on the same scale as PISA, but are not true 'PISA results.'

The measurement variance tends to be greater for YITS reading estimates than for PISA estimates, due to the shorter test. Any statistical comparisons between PISA plausible values and YITS plausible values should consider the heterogeneity of error variance between the two sets. Also, because each plausible value is drawn independently, each of the sets of plausible values within each survey (PISA or YITS) are independent. However, the entire set of plausible values from one sample is dependent on the set of plausible values from the other survey. Thus, to calculate variances in change of proficiency between PISA 2000 and YITS 2010 for an individual, the complete permutation of plausible values should be used to calculate the variance for difference in proficiency between surveys (*i.e.*, 5 x 10 = 50 difference estimates). In equations

(6) to (9), i=1,2,...,I and j = 1,2,…,J are used to index the individual plausible values from the set of plausible values from each survey.

$$\Delta_{ijn} = pv_{PISA_i n} - pv_{YITS_j n} \quad (6)$$

$$s.e.(\Delta_n) = \frac{\sum_i^I \sum_j^J \left(\Delta_{ijn} - \overline{\Delta}_n\right)^2}{IJ - 1} \quad (7)$$

Note that

$$\overline{\Delta} = \frac{\sum_i^I \sum_j^J pv_{PISA_i} - pv_{YITS_j}}{IJ} \equiv \frac{\sum_i^I pv_{PISA_i}}{I} - \frac{\sum_j^J pv_{YITS_j}}{J} . \quad (8)$$

So if only the mean difference is required, it may be computationally simpler to use the latter form. Matching plausible values one-to-one in the manner of a paired-sample estimation of mean difference will produce inaccurate results of variance (not to mention being inconsistent, as there are more plausible values for YITS 2010 than for PISA 2000). With modern computers, it is usually easiest to script the necessary statistical procedures to handle multiple imputations following the algorithms described in Rubin (1987).

Most alternate approaches, including matching a single plausible value from one survey to all plausible values in the other survey or averaging across plausible values within cases, will produce underestimates of variance. Because the plausible values are independent, the variance of an individual PISA-YITS difference is simply the sum of the variances from each set.

$$\sigma(\Delta_n) = \frac{\sum_i^I \left(pv_{PISA_{in}} - \overline{pv}_{PISA_n}\right)^2}{I - 1} + \frac{\sum_j^J \left(pv_{YITS_{jn}} - \overline{pv}_{PISA_n}\right)^2}{J - 1} \quad (9)$$

However, this error variance estimate is not easily aggregated to the population level, because the YITS 2010 survey employs a complex sample. Thus, even simple sum-based statistics would require estimates for covariances between cases that are difficult to compute. Ignoring these covariances will lead to inaccuracies, particularly with the highly-clustered YITS sample.

Another limitation of the YITS results is the lack of reporting for any reading subscales. Although the items were balanced across reading content areas, there are not enough items available in any subdomain to produce sufficiently accurate results.

REFERENCES

Adams, R.J. (2002). Scaling PISA Cognitive Data In Adams, R.J. & Wu, M.L. (editors) *PISA 2000 Technical Report*. OECD Publications: Paris.

Adams, R.J. & Wu, M.L. (editors) (2002). *PISA 2000 Technical Report.* OECD Publications : Paris.

Fuchs, T. & Woessmann, L. (2004). What accounts for international differences in student performance? A re-examination using PISA data, CESifo Working Paper Series 1235, CESifo Group: Munich.

OECD (2009). *PISA 2006 Technical Report.* OECD: Paris.

OECD PISA Project Consortium (June 5, 2001). PISA international item parameter thresholds for a response probability of 0.62. OECD: Paris.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons: New York.

Willms, J. D. (2004). Reading achievement in Canada and the United States: findings from the OECD Programme for International Student Assessment. Human Resources and Skills Development Canada: Ottawa.

APPENDIX A
READING MARKING GUIDE

GENERAL INTRODUCTION

The PISA Reading instrument comprises several different types of items including multiple-choice items, short constructed response items and items which require more extended responses.

Responses to multiple-choice items and some short constructed response items will be directly entered into the data-entry software.

Responses to most short-response and to all extended-response items need to be coded by expert markers. This Marking Guide contains directions for marking all such Reading items.

In a second stage, the codes determined by the markers, together with codes for the rest of the items, will be entered into the data entry software and converted electronically into scores for each student.

Layout of the marking guide

Each item in this Marking Guide begins at the top of a page. The **question heading**, which includes the name of the unit, appears top left, with the **item identifier** and the available **codes** at top right. This line is followed by the question **stem** as it appears in the Student Booklet, and then the **question intent**, which gives a general description of what the question is intended to assess. A description of how to code the item follows. The **coding** for each item in the guide consists of:

- **credit labels**. The section describing the highest score for each item is headed with the label “full credit”. The section describing unacceptable responses and omissions (Code 0 and Code 9) is headed with the label “no credit”. Those items for which there are intermediate codes include a section headed “partial credit”.

- the **numerical code** (*e.g.* Code 1) for each category of response;

- a general **description** of the type of response for each code; and

- **examples** of responses for each coding category, listed as dot-points below the description, and often followed by an explanation in italics. The dot-pointed examples are offered as some possible answers; they are not a complete list of possible answers.

General Principles for Coding

Spelling and grammar

Spelling and grammar mistakes should be ignored unless they seriously obscure meaning. This is not a test of written expression.

Exercising judgement

Although the coding descriptions and examples are intended to minimise subjectivity, markers will inevitably have to exercise some judgement in determining the boundaries between codes and the limitations of what constitutes a particular code. As a general principle, judgement should be based on the marker's best assessment of whether the student is able to answer the question. Markers should avoid applying a deficit model, that is, deducting points for anything that falls short of a perfect answer.

Note that the terms "full credit", "partial credit" and "no credit" are used instead of "correct" and "incorrect" responses. There are two main reasons for adopting these terms. First, some questions do not have "correct" answers. Rather, responses are graded based on the degree to which students demonstrate an understanding of the text or topic in question. Secondly, "full credit" responses do not necessarily include only fully correct, or perfect, responses. In general, "full credit", "partial credit" and "no credit" sections divide the students' responses into three groups in terms of the extent to which the student demonstrates ability to answer the question.

When to consult the supervisor

If a marker is unable to determine which code a response should receive, or if the student response clearly demonstrates understanding of the text and the question but does not fit into any given score category, the marker should consult the supervisor for the marking of Reading who will either make the judgement, or, if unable to do so, pass the question on to the National Project Manager (or the person in charge of marking). Cases which cannot be resolved at the National Centre must be referred to the Consortium through the email marker query service.

Some Common Problems

Response is given in a format other than the one asked for

In these cases the marker should consider whether the student has understood the substance of the question and met the purpose of the question in the response. Some examples of such responses and how to deal with them are outlined below.

> *Numbered lines are provided for separate parts of a response, but the student writes more than one (correct) element on one line.*

In this case, the marker should ignore the arrangement of the information. Each element of the response should be considered separately, regardless of how it is positioned on the lines.

> *The question calls for a written response, but the student responds by circling or underlining part of the text.*

The response should be coded according to how well the indicated section satisfies the response called for. For example, if the item asks the student to "list three things in the text", and the student draws arrows from the stem to the appropriate part of the text or underlines or circles the three things in the text, the answer should be given full credit. On the other hand, if the answer asks for a response "in your own words", underlining in the text should be given no credit.

The question calls for underlining part of the text, but the student gives a written response.

The response should be coded according to how well it matches the substance of the section which should have been underlined.

Response contains elements of two or more different codes

First, consider whether the elements in the response contradict one another or not.

Contradictory elements

If the elements of the responses are contradictory, choose Code 0. For example, if a numerical answer is required, a response which provides two different numbers is considered self-contradictory and should therefore be coded 0.

Non-contradictory elements

If the elements of the response do not contradict one another, consider whether the elements are drawn from a restricted or a wide pool of possibilities.

Restricted pool of possible responses

In some cases a very limited pool of possible answers is available, for example when students are asked to choose one of only four or five factors mentioned in the passage. Here, a response containing two or more elements, one of which is incorrect or irrelevant, should be coded 0. (This is considered equivalent to checking more than one alternative in a multiple-choice item.)

Wide or unrestricted pool of possible responses

In some cases an extensive range of possible answers is available either within the text or outside it. Here, codes should be assigned to give the student credit for the acceptable part of the response unless the incorrect element of the response conflicts with the text.

Superfluous underlined text

Where underlining is called for, if more than the required section is underlined apply the following rules:

- If a section of the text is underlined which does not relate to the item or to another item on the same text, choose Code 0.

- If a section of the text is underlined which does not relate to the item but DOES relate to another item on the same text, choose Code 1.

“No Credit” Codes

Code 0

This code is used for responses where there is any evidence that the student has made an unsuccessful attempt to answer the question. Use as specified for each item. Some common specifications are “insufficient or vague”, “inaccurate comprehension”, and “implausible or irrelevant”.

Code 0 should also be used for the following:

- An answer such as “I don’t know”, “this question is too hard”, a question mark or a dash (—);
- An answer which has been written and then erased or crossed out, whether it is legible or not; and
- An answer which is clearly not a serious attempt. Examples of this kind of response include jokes, swearing, names of pop stars and negative comments about the test.

Code 9

This code is labelled “Missing” in the body of the Marking Guide. It is used for those cases where a student has apparently not attempted a question. A blank space or words indicating no attempt (*e.g.* “Ran out of time”) should receive a Code 9

Not Applicable code

This code is used if a question was misprinted so that it was not possible for the student to answer it. For example, there may be a photocopy or printing error so that the question is not legible. In this case, please write “n” as the Not Applicable Code next to the item. We expect that the Not Applicable Code will only be used on rare occasions, if at all.

(Item-specific instructions are omitted for confidentiality)

THE OECD EDUCATION WORKING PAPERS SERIES ON LINE

The OECD Education Working Papers Series may be found at:

- The OECD Directorate for Education website: www.oecd.org/edu/workingpapers
- The OECD's online library, SourceOECD: www.sourceoecd.org
- The Research Papers in Economics (RePEc) website: www.repec.org

If you wish to be informed about the release of new OECD Education working papers, please:

- Go to www.oecd.org
- Click on "My OECD"
- Sign up and create an account with "My OECD"
- Select "Education" as one of your favourite themes
- Choose "OECD Education Working Papers" as one of the newsletters you would like to receive

For further information on the OECD Education Working Papers Series, please write to: edu.contact@oecd.org.