Linking the British Columbia English Examination to the OECD Combined Reading Scale

Prepared for the British Columbia Ministry of Education

by

Fernando Cartwright, Polymetrika Inc.

February, 2012

Application of equipercentile equating methods to link provincial grade 10 assessment in British Columbia with the OECD PISA scale. Estimation of linking error makes use of BRR and plausible value methodology. Results include validation using subsamples and an exploration of an alternate methodology for creating statistical projections to link assessments.

Contents

Introduction

In 2009, a sample of 2367 15-year-olds enrolled in British Columbia schools participated in the Organisation for Economic Cooperation and Development's (OECD) Programme for International Student Assessment (PISA). The main assessment of PISA 2009 was the reading component, meaning that all students who participated in PISA 2009 completed a reading assessment (in addition to a minor domain assessment). Of these participating students in British Columbia, 2195 were administered the English version of the PISA assessment and also completed a grade 10 English language examination (BC). The purpose of the study described in this report is to establish a statistical linkage between the two assessments such that the results of each assessment may be expressed on the scale of the other.

The scales of the two assessments differ in both their manner of reporting as well as their order of magnitude. The PISA assessment does not produce a single point estimate for each student; rather, each student receives a set of five plausible values that represent the posterior distribution of reading proficiency corresponding to the student's item responses and personal characteristics. The variability between these five plausible values represents the error of measurement for the student. These plausible values are reported on a scale that was initially established in PISA 2000 to have an international mean of 500 and standard deviation of 100 (based on participating OECD countries). In subsequent cycles of PISA, the observed mean and standard deviation of OECD countries tend to vary from these initial values due to the use of equating methods; the differences reflect both changes in student performance within participating countries as well as changes in the participation of different OECD countries. However, the values remain of similar magnitude (for more information, see http://www.pisa.oecd.org). In contrast, the BC assessment produces a single point estimate for each student, representing the mean of the posterior distribution, which is itself based on the student's item responses and the empirical distribution of proficiency for all students completing the assessment.

Each score estimate is accompanied by an estimate of the standard error, calculated as the standard deviation of the posterior distribution function. The scale of the BC assessment score estimates is the unadjusted theta scale, which has a latent mean of 0 and standard deviation of 1.

This report details the methodology used to establish the linkage and provides a summary of the results. There are five sections to this report following this introduction. The first section describes the scope of the linkage project and provides the conceptual foundations and rationale for performing the linkage. The second section describes the statistical methodology used to estimate the linking function, as well as the methods used to validate the linkage and produce estimates of data quality for each linked score. The third section presents the results of the linkage. The fourth section includes an interpretation of the results as well as a discussion of their limitations. The fifth section provides some recommendations for future linking studies, including an alternative to statistical equating methods.

1. Conceptual Foundations

1.1. Content review

As a precursor to estimating the statistical linkage, the first stage of the process involved a review of the content from each of the two assessments. The review was conducted by content specialists with expert knowledge of the British Columbia English curriculum who are also familiar with the PISA reading assessment framework. The review involved two parts: first, an allocation of each item from each assessment onto the specifications framework of the alternate assessment, and, second, an holistic comparison of the cognitive demands of the items from the two assessments. The report summarizing the findings of the second part can be found in Annex B of this report. The salient finding of the content review is that “[o]verall, both test instruments measure approximately the same reading skills” (see Annex B).

The data from the second part of the content review, excluding the writing item (which is outside the scope of the PISA reading framework)[1], are summarized in Table 1.1. In Table 1.1 (and also noted in the holistic review) there are several differences in the content allocation between the two assessments. The PISA assessment used an approximately equal number of multiple choice (48) and short written response (53) types, whereas the BC assessment used almost exclusively multiple choice items (29) with two extended written response item. However, the complexity of written responses on these two items was much greater than the complexity of any written responses required by the PISA assessment.

In general, there is a greater focus in the BC Exam on literature and literary conventions. On the other hand, PISA placed greater emphasis on cognitive operations typically considered as "higher order", with greater proportion of items in the 'Reflect and evaluate' subscale from the PISA framework and a greater proportion of items in the 'Analyze texts' question type from the BC framework. Notwithstanding the overall conclusions of the content review, these discrepancies should be taken into consideration when interpreting the results of the linkage.

[1] The BC score that is used to link is the total IRT score produced using all items, including the writing components. The justification for using the total score is the strength of the correlation with the PISA scale, compared to the other scores (see section 1.3).

Table 1.1. Item allocations according to framework in PISA reading and B.C. English reading.

	PISA Items (total=101)			British Columbia Items (total=29)			
Row Labels	Info	Lit prose	Total	Info	Lit prose	Poetry	Total
PISA Framework							
Access and retrieve	21.8	1.0	22.8	6.9	6.9	3.4	17.2
Integrate and interpret	39.6	12.9	52.5	17.2	27.6	20.7	65.5
Reflect and evaluate	22.8	2.0	24.8	3.4	0.0	3.4	6.9
(Not classifiable)	0.0	0.0	0.0	3.4	3.4	3.4	10.3
British Columbia Specifications							
Analyze texts (AT)	14.9	3.0	17.8	0.0	6.9	0.0	6.9
Interpret texts (IT)	32.7	9.9	42.6	10.3	10.3	10.3	31.0
Retrieve information (RI)	21.8	3.0	24.8	10.3	10.3	10.3	31.0
Recognize meaning (RM)	14.9	0.0	14.9	10.3	10.3	10.3	31.0

Note.
Values in this table are expressed as percentages of the total number of items.
In addition to these reading items, the B.C. English examination includes two writing items, one which requires students to analyze texts (AT) and write a response to a prompt and another extended free writing task. These items are scored on a seven-point scale.

In general, despite having different allocations of items, there is a strong association between the frameworks themselves. Table 1.2 contains the pairwise percentages in the cross-classification of items from the two assessments on both frameworks. The columns on the left half describe the cross-classifications of PISA items, and the columns on the right describe the cross-classification of BC items. The correlation between classifications is 0.68 using the PISA items and 0.80 using the BC items.

Table 1.2. Concordance between cognitive framework categories in PISA reading and B.C. English reading.

	PISA Items (Percent of total)					British Columbia Items (Percent of total)				
	AT	IT	RI	RM	Total	AT	IT	RI	RM	Total
Access and retrieve	0.0	0.0	20.8	2.0	22.8	0.0	0.0	5.0	0.0	5.0
Integrate and interpret	3.0	33.7	4.0	11.9	52.5	2.0	7.9	4.0	5.0	18.8
Reflect and evaluate	14.9	8.9	0.0	1.0	24.8	0.0	1.0	0.0	1.0	2.0
(Not classifiable)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	3.0
Total	17.8	42.6	24.8	14.9	100.0	6.9	31.0	31.0	31.0	100.0

Note.
This table does not include the two items requiring written responses on the B.C. English examination.

1.2. Equity

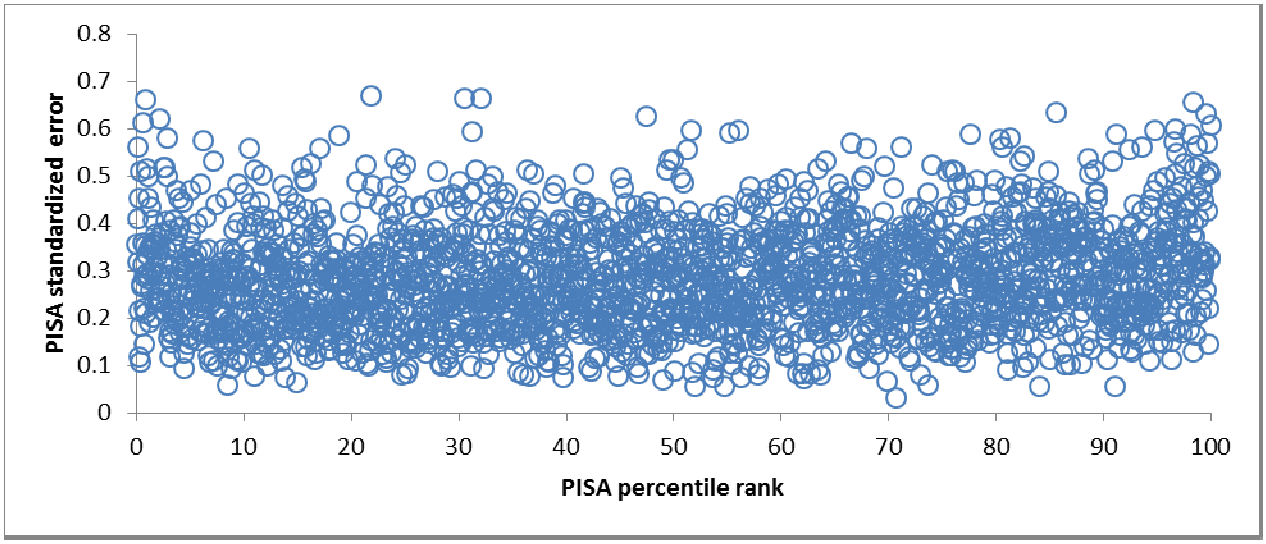
One of the principles underlying a statistical linkage is ‘equity.’ Equity exists between two assessments when they share the same test specifications and degree of statistical accuracy. While it is uncommon that two independently-developed assessments would satisfy these conditions, the degree to which these conditions are not satisfied communicates the extent to which the results of the statistical linkage may be interpreted. Where two assessments satisfy the condition of strong equity, there are no limits to the interpretation of the linkage – the linked results from one test may be used interchangeably with the results of the other. As the equity becomes weaker, the valid uses of the results become more restricted (Mislevy, 1992; Linn, 1993). In particular, when the both assessments are constructed to different specifications, but still measure a common construct, and their levels of accuracy also differ, the validity of different interpretations of the linkage may depend on the context or subpopulation to which the interpretation is applied and it may require additional information to support specific inferences.

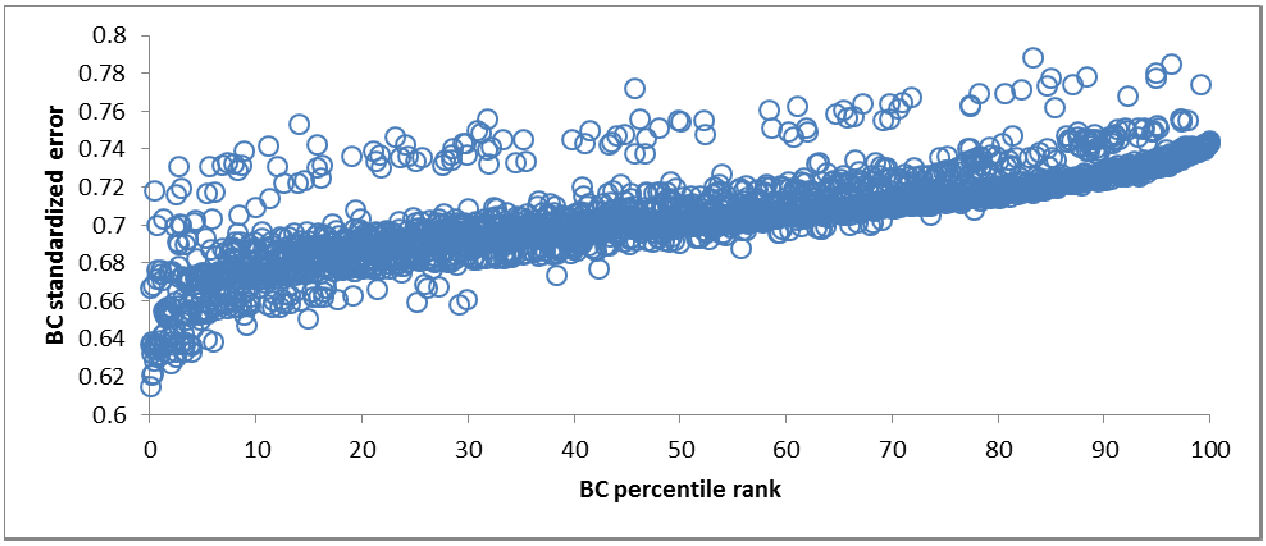
The findings of the content review indicate that, although the tests measure the same general concept of ‘reading’, the specifications of the two assessments are different.

These differences reflect the purposes for which they were produced. The PISA assessment is intended to provide information about how students' reading skills are preparing them for general post-secondary education and the labour force, consistent with the OECD's economic focus. On the other hand, the specifications of the BC assessment reflect the role of curriculum in determining content, and the greater specificity to the goals of the curriculum beyond economic outcomes.

Comparing the statistical accuracy of the two tests is slightly complicated when the two tests have different scales. The difficulty arises because the standard error of each score on an assessment is expressed on the same scale as the score itself. In order to make a fair comparison, the standard errors must first be re-standardized by dividing each by the standard deviation of the sample. Furthermore, because the standard errors often vary across different scores, even on the same scale, the standard errors must be compared on an equivalent interim scale. In this case, because a common sample of respondents is used for both assessments, the natural equivalent scale is the percentile scale; on both assessments, percentile rank denotes the same interpretation. Accordingly, the re-standardized errors, matched by percentile rank, are plotted in Figure 1.1.

Figure 1.1. Comparison of standardized errors, PISA reading and B.C. English reading.





As illustrated in Figure 1.1, the standard errors of the assessment scales are not equivalent. The magnitude of error is relatively random in PISA with respect to the percentile rank and is generally around a third the magnitude of the standard deviation (~0.3). In contrast, the BC standard errors tend to increase linearly with proficiency, indicating that fewer items are present on the test to adequately discriminate between higher proficiency students. The BC standardized errors are also larger (over twice that of the PISA results), which is most likely the result of the shorter test and the lack of the complex score conditioning used by PISA.

1.3. Score interchangeability

Regardless of any conceptual differences between two assessments, an argument could be made for the de facto equivalence of two scores if the distributional properties of the scores are approximately the same and there is a strong correlation between the two. However, the distributional characteristics are different between the BC and PISA assessments (see Table 1.3). The PISA scale is greater in both magnitude and variability. Also, the PISA scale has greater negative skewness than the BC scale, and the BC scale has greater negative kurtosis. While a simple linear transformation would suffice to equate the first two central moments, it will not remedy the differences in skewness and kurtosis. From a practical perspective, this means that students at the lower end of the distribution appear to lag further behind

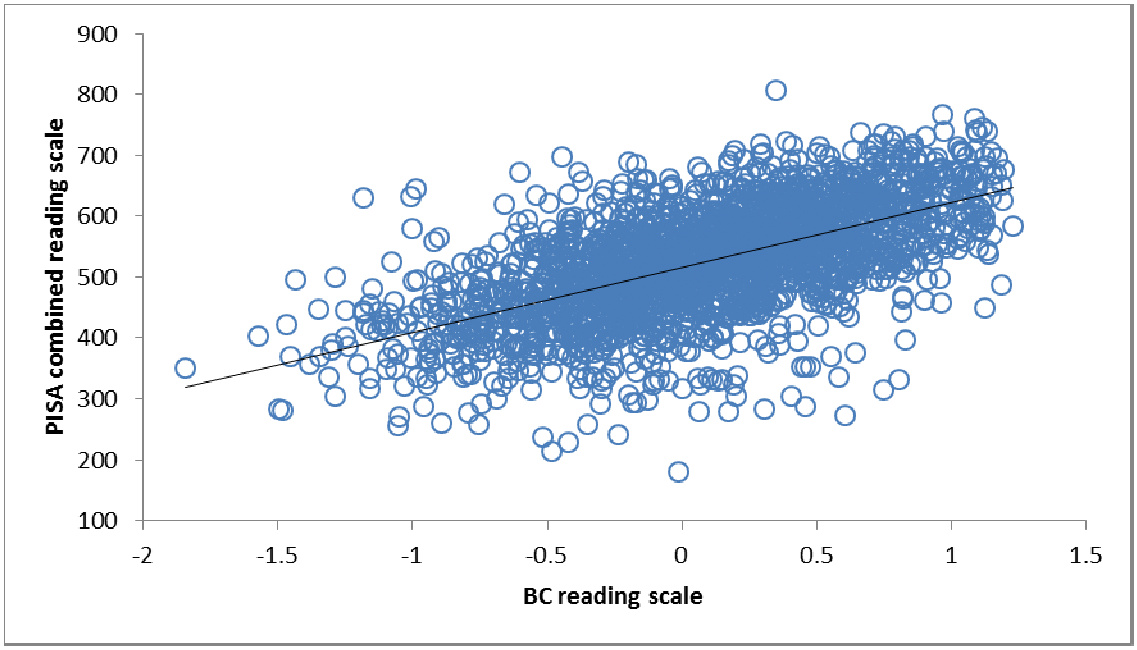
other students when examined on the BC scale compared to the PISA scale, regardless of the magnitude and variability of the scale.

Table 1.3. Distributional characteristics of PISA Reading and British Columbia English, British Columbia, 2009.

	PISA	BC
Mean	526.79	.10
Std. Deviation	89.63	.51
Skewness	-.293	-.274
Kurtosis	.045	-.164

The relationship between the two scores, illustrated in Figure 1.2, indicates a weak-to-moderate relationship between the two assessment results. The observed correlation is 0.61, indicating that, although there is a general relationship between the two variables, the majority of the variance in each variable cannot be explained by the other variable. It should be noted that the measurement variance of the BC measure is a limiting factor in the observed strength of the relationship. There are several subscales available on the BC/PISA data file, composed of subsets of the BC and PISA scale items; each of these subscales has weaker correlations with each other as well as the overall PISA and BC scales, consistent with attenuation due to measurement error. Although there is a conceptual appeal in linking the subscales with the greatest cognitive similarity, in order to limit the attenuating effects of measurement error, the linking function is estimated using only the overall scales. The overall BC score that is used in this study is a weighted average of the IRT scores for each subscale, while the combined PISA reading score is directly estimated from the item parameters and response, with statistical adjustments (also referred to as *conditioning*) using corollary data from the student and school questionnaires.

Figure 1.2. Scatterplot of assessment results, PISA versus BC.



1.4. Summary

The results of the initial review and comparison of the assessments indicate that the two assessments are measuring the same general construct of reading. However, there is insufficient equity between the assessments to support the notion of interchangeability between the scores of the two assessments, even after the application of a linking function.

Therefore, although the statistical methods used to create the linking function can effectively rescale each assessment's results to be comparable to the scale of the other, the individual scores cannot be interpreted in the same manner as individual scores on the other. Because the error between the two assessment results appears to be random, at an aggregate level, rescaled results from one assessment may provide useful prediction of the results of the other. However, given the lower-than-expected correlation between the two scales, there is the possibility that, for different subpopulations, the 'optimal' linking function may differ. Accordingly, the linking function requires validation for the different subpopulations to which it will be generalized. In practice, this validation is restricted to the subpopulations for which there is a clear definition and sufficient sample size.

2. Methodology

2.1. Equipercentile equating

Given the differences between the two assessments in terms of distributional characteristics (see Table 1.2) the linking function cannot make use of a simple linear transformation. Also, given the goal of the project of representing the existing reporting scales from the two assessments, any form of common item calibration is equally unacceptable, because it would fundamentally alter the properties of the existing reporting scales. Given these constraints, the most useful equating method is that of equipercentile equating (Kolen & Brennan, 1995), which matches the scales by first finding the percentile rank equivalent of each score on one assessments and then finding the percentile corresponding to that score on the other assessment. For example, if score X is equal to percentile rank A on the PISA scale, and score Y is equal to percentile rank A on the BC scale, then score X on the PISA scale and score Y on the BC scale are considered equipercentile equivalent scores.

2.2. Kernel smoothing

The challenge in applying equipercentile methods is that, in the example where score X on PISA corresponds to percentile rank Y, there may be no equivalent score on the BC Exam that corresponds to the same percentile rank. Thus, it is necessary to estimate a smooth percentile rank function (and its inverse) for each scale such that a scaled percentile score may be interpolated for any given percentile rank, even if there is no observed scale score with that value. The requirement for smoothness of the percentile rank function is based on the statistical assumption that, as sampling error decreases, the true distribution of scores should also become smoother. In order to reduce the influence of random error on the estimation of the linking function, the estimated percentile functions should eliminate ‘bumps’ in the observed distribution, while maintaining the core features of the distribution, such as the first four central moments and the observed quartiles.

The most flexible method of smoothing a distribution of observed scores is kernel smoothing, where each observed score is replaced by a unimodal ‘kernel’ function

(Hastie, Tibshirani, & Friedman, 2001). The kernel function values for each score are then added together to represent the complete distribution. The choice of the kernel function depends on the properties of the underlying scale being smoothed. Because both of the scales are unbounded at either end and are assumed to be asymptotic (meaning that there is no theoretical maximum or minimum score, but the probability of a score being observed decreases continuously as the score differs from the average), an appropriate kernel function is the Gaussian function, which shares the same properties.

Because the linking function and estimation are implemented in SPSS, the methodology makes use of the PDF.NORM() function in SPSS to produce the Gaussian kernel. The overall smoothness of the estimated percentile function depends on the bandwidth (or standard deviation, in the case of the PDF.NORM[] function) that is used for the kernel. The optimal bandwidth reduces the irregularities in the distribution while maintaining the salient properties of the distribution. There is no standard method of finding an optimal bandwidth, and visual inspection remains the most useful means of validating a smoothed distribution. However, the iterations in finding the final values were based on a combination of the measurement error and sample optimization[2].

2.3. Linking error

The linking error represents the uncertainty with which one score may be interchanged with another. In practical terms, if a linked score of PISA-to-BC is presented with a linking error equal to *x*, then the interpretation of the link implies that, in 68% of samples using similar students and tests, the true BC score corresponding to the given PISA score would fall within the range of the linked score plus or minus *x*.

There are two components to the linking error. The first is sampling error, representing the degree to which different samples of students would have produced

[2] In a wide variety of smoothing situations, the optimal kernel bandwidth tends to be around the fifth root of the sample size.

different relationships between the two assessments. The second is measurement error, representing the degree to which different selections of test items or score estimation methods would have produced different scores for each assessment. The total error for the linking function is estimated by adding the two error components together.

2.3.1. Sampling error

The influence of sampling error is estimated using the standard PISA BRR methodology, using the 80 replicate weights provided along with the data file (OECD, 2009). Using these replicate weights, the linking function is estimated 80 times; and the variation between these estimates provides an estimate of the sampling error. For ease of calculation, the sampling error is calculated using the first plausible value (see below)

2.3.2. Measurement error

The measurement error for the linking function is estimated using the plausible value methodology used by PISA (OECD, 2009). Because the BC Exam methodology only reports a single point estimate and the standard error, corresponding plausible values were drawn for the BC Exam by constructing an approximate posterior distribution for each student using the score estimate and corresponding standard error for each existing BC score as the mean and standard deviation, respectively, of a normal probability distribution, from which 5 values were randomly drawn. Following the PISA methodology, the linking function is estimated separately using each plausible value. The variability between the estimates (adjusted for the number of plausible values) provides an estimate of the influence of measurement error on the linking function. The measurement error is calculated using the final student weight on the PISA sample data.

2.4. Validation

Because the correlation between the two initial measures is relatively weak, compared to traditional equating scenarios, some validation is required to gauge the extent to which the error between the two measures is sensitive to population variations. Validation is conducted by dividing the population into meaningfully

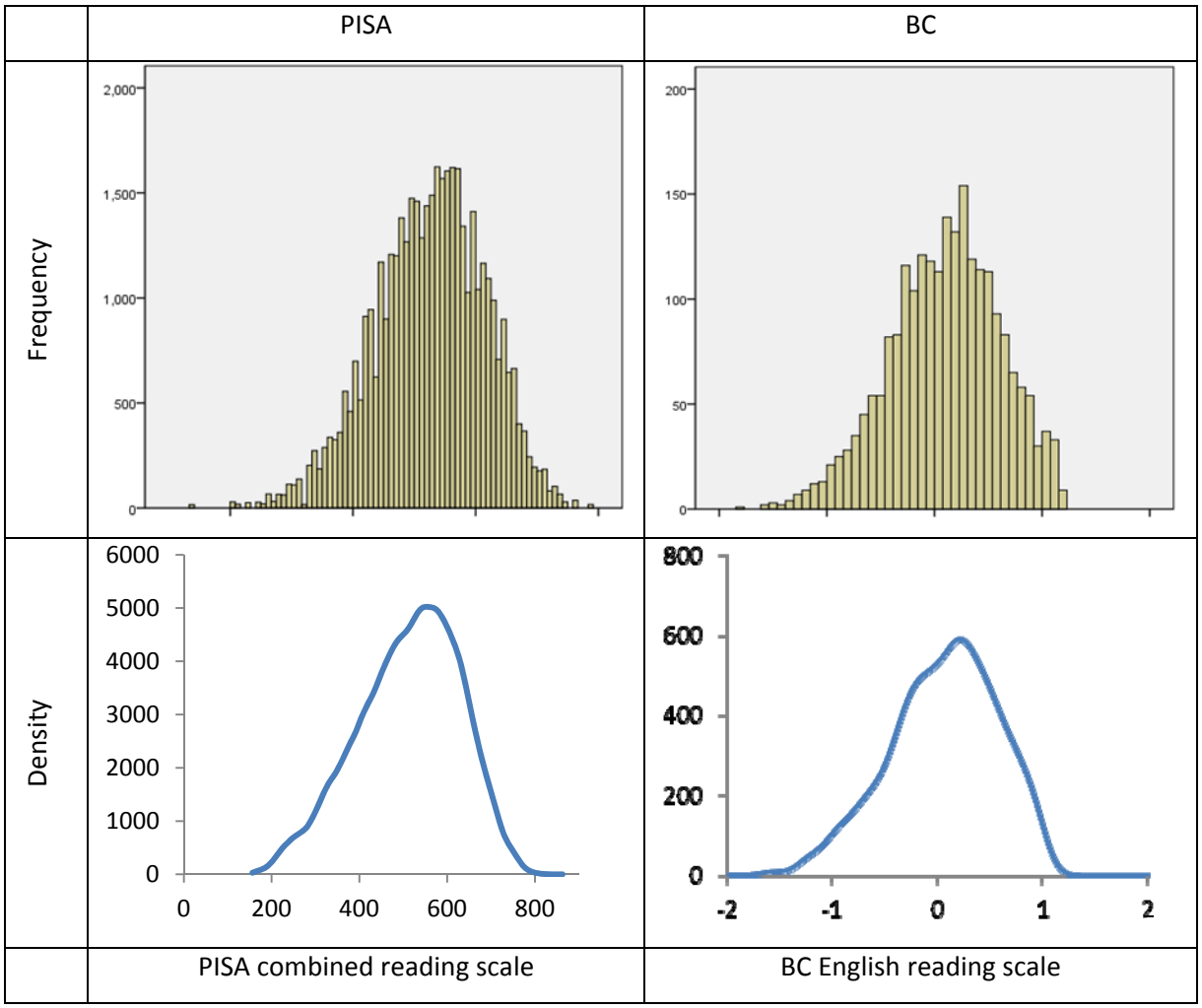
distinct subpopulations and estimating the linking function separately for the two groups. The linking functions are then applied to the entire sample to determine the magnitude of the sensitivity of the linking function to population membership. For this study, the validation was conducted on the male and female portions of the sample separately.

3. Results

3.1. Smoothed distributions

The observed and smoothed distributions of scores are illustrated in Figure 3.1, with the distributional characteristics summarized in Table 3.1. As can be seen in Table 3.1, the smoothed distributions have adequately removed the irregularities in the distribution, while also maintaining fidelity to the salient statistical properties of the original observed score distributions. For both scales, the kurtosis was not well reproduced by the smoothed distributions; the PISA distribution is slightly wider and squatter near the peak and the BC distribution is slightly narrower, compared to their expected observed distributions.

Figure 3.1. Observed and smoothed score distributions for PISA and BC scales.



Note. The horizontal increments on the frequency histograms have the same value as the corresponding increments on the smoothed distributions beneath, regardless of the differences in visual scale.

Examination of the observed frequency distributions suggests that the initial kurtosis values may be the result of sampling and/or measurement error. The PISA observed frequencies indicate several score values close to the mode that spike relative to the scores around them; in the smoothing process, the weight of these spikes becomes distributed, which softens the severity of the peak. In contrast, the difference in kurtosis for the BC distribution is likely the result of the larger standard errors near the high end of the distribution, combined with the negative skewness of the original scores; on the application of the kernel functions, more of the weight of each of the higher scores is redistributed towards the mean compared

to the lower scores. As a result, more of the weight is reallocated towards the peak in the smoothed distribution.

Table 3.1 Distributional properties of score distributions before and after Gaussian kernel smoothing.

	PISA		BC	
	Smoothed	Original	Smoothed	Original
Mean	525.14	526.65	.097	.100
Standard deviation	90.33	89.69	.507	.512
Skewness	-.28	-.29	-.277	-.276
Kurtosis	-.10	.04	-.107	-.163
Percentiles				
10	404.3	407.1	-.57	-.56
25	465.5	468.7	-.24	-.25
50	530.9	532.0	.12	.13
75	589.8	589.2	.45	.47
90	637.4	639.4	.75	.76

Note. PISA values are based on the first plausible value.

3.2. Linking functions

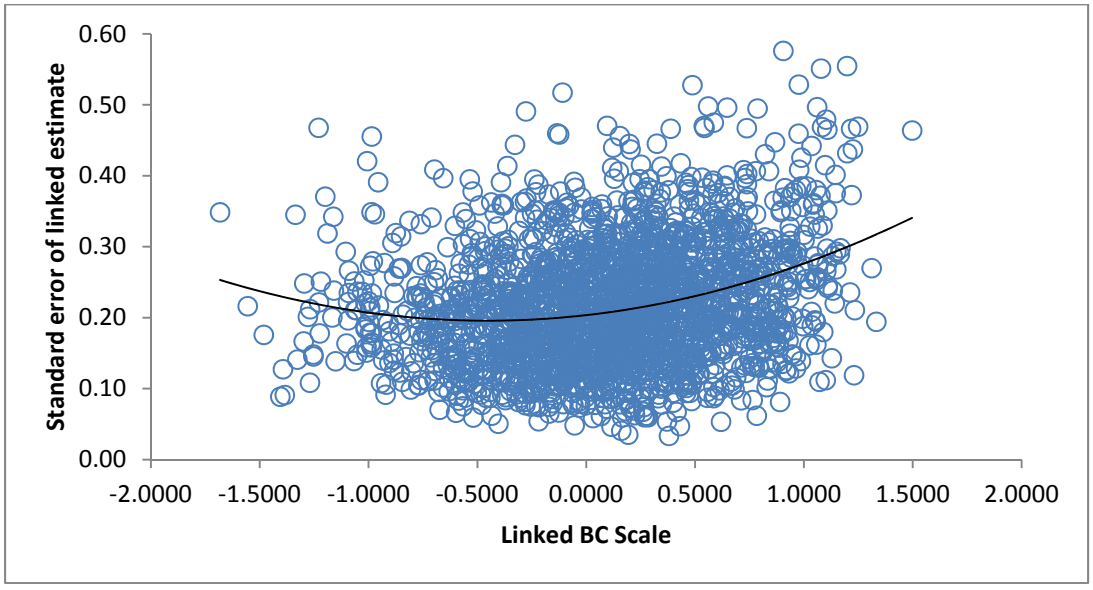
The linking functions are implemented using a net of 300 nodes, distributed across each of the scales using Gaussian quadrature, using a normal distribution with a standard deviation 1.5 times the observed standard deviation in each of the sets of observed scores. The link function itself is implemented through four functions that are discretized on this net: each assessment has both a percentile rank function and an inverse percentile rank function. For example, to find the BC Exam equivalent for a PISA score of x, the percentile rank function for PISA first converts the PISA score to percentile rank. Then, the percentile rank value is used as the independent value in the inverse percentile rank score for the BC Exam, producing the linked score. The inverse operation uses the percentile rank function for the BC Exam and the inverse function for PISA. This function can also be approximated with the polynomial, although the polynomial approximation does not have the same mathematical invertibility as the true linking function.

Precise values are found on each of these functions using triangular interpolation between the quadrature nodes. All functions are implemented in the SPSS syntax files associated with this report.

3.2.1. Standard errors

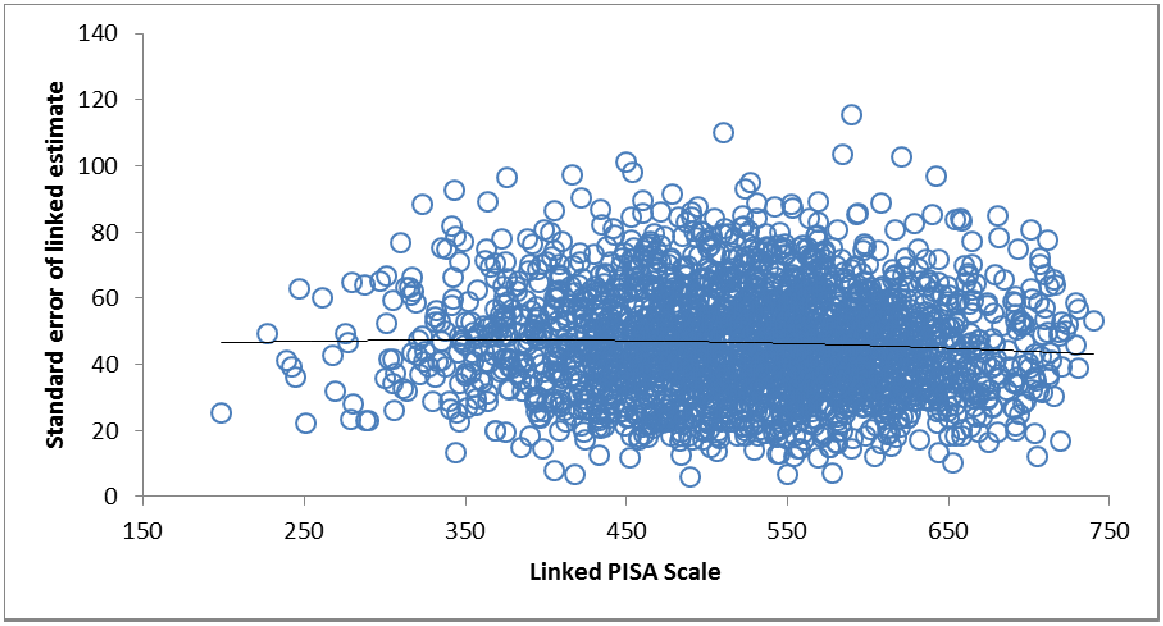
The standard errors for the linking functions are similarly interpolated based on the observed standard linking errors for the observed score values in the linking sample. These errors are illustrated in Figures 3.3 and 3.4. While the standard errors have a weak tendency to be slightly higher for linked estimates on the BC scale at the upper and lower end of the distribution, there is no relationship between the linked estimate and its standard error on the linked PISA scale. In the context of the standard deviations of the original scales (see Table 3.1), the standard errors contribute substantially to the total variation, at the same magnitude as the standard deviation of the scores themselves. In other words, over half of the total variation in linked estimates is a product of linking error, which is consistent with the lower correlation initially observed between the two scales.

Figure 3.3. Relationship between linked score and standard error of linking for BC-to-PISA.



Note.
This figure omits four cases which, due to their abnormally large measurement error on the PISA assessment, also had linking errors that were disproportionately large.

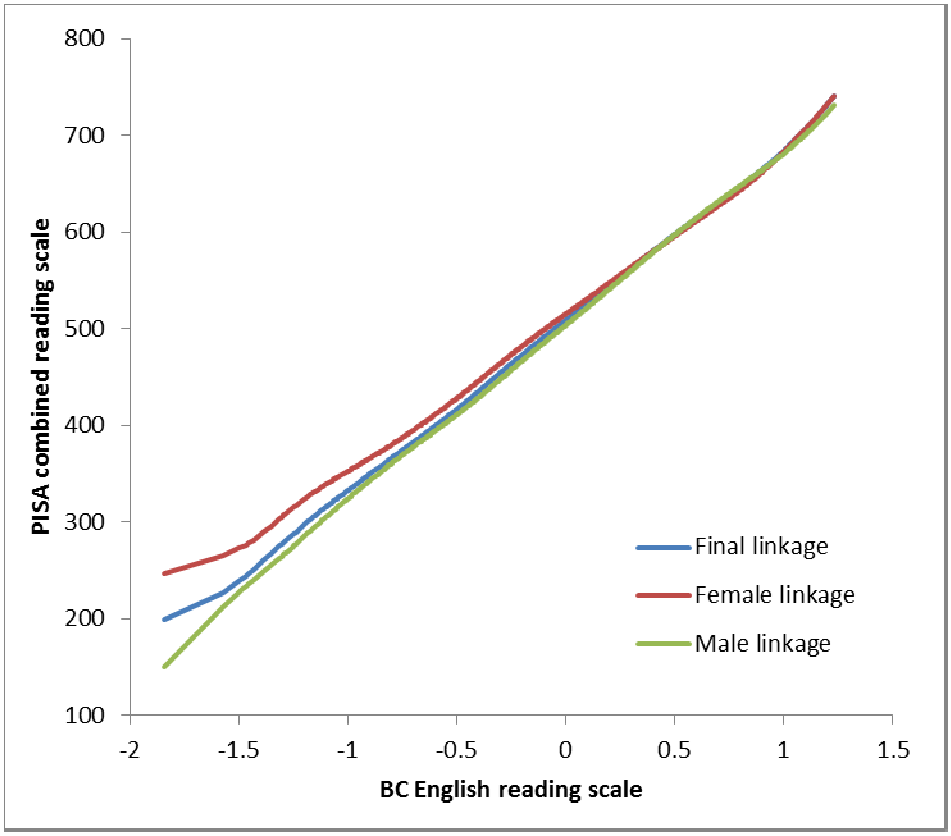
Figure 3.4. Relationship between linked score and standard error of linking for PISA-to-BC.



3.2.2. Validations

The cross-validations were performed separately on males and females, and the results were compared to the final estimated linking function. The score-to-score concordances for all three linking functions are illustrated in Figure 3.5. Each linking function is invertible, so the functions illustrate both the BC-to-PISA link as well as the PISA-to-BC link. For the upper end of either scale, all linking functions provide similar results. However, there is divergence at the lower end of the distribution; the function estimated for males is more similar to the final estimate. This phenomenon is largely due to the overrepresentation of males at the lower end of either scale in the general population.

Figure 3.5. Illustration of the BC-to-PISA linking functions estimated using the full sample, male sample and female sample.



These results suggest that, regardless of the random error calculated for the linking function, the results are relatively stable across subpopulations. However, caution should be used when interpreting the results for subpopulations whose distributions of reading proficiency are distinctly different than the general population.

4. Interpretations

Given that both of the assessments already meet their respective goals, it is unlikely that the results of the linkage will be used as interchangeable scores, in the sense that, in the absence of PISA scores for most students or in each semester, BC Exam scores would be used to predict performance of individual students on PISA (or vice versa). However, the linkage does allow for comparison of other regions, such as other Canadian provinces, against the performance of students in British Columbia on the same scale as the BC Exam. Similarly, interpretations of BC Exam scores

can be compared against the interpretations of PISA scores; for example, what does "A" level performance in British Columbia mean in an international context?

The relative positions of the standards for interpreting reading proficiency on each of the scales are shown in Table 4.1. For example, the A level standard in BC is consistent with a high level 4 performance on PISA[3]. Similar to previous international linking studies (Cartwright et al, 2003) which found that the standard 'meeting expectations' was consistent with Level 3 performance in PISA 2000, the C+ level standard in BC is consistent with Level 3 performance in PISA 2009. According to the BC guidelines, a C+ conveys that "[t]he student demonstrates **good** performance in relation to expected learning outcomes for the course or subject and grade" (Ministry of Education, Province of British Columbia, 2011, p.90). Arguably, this definition is closer in normative interpretation to the FSA level of "meeting expectations" than the more tepid "satisfactory performance" and "minimum acceptable performance" associated with C and C+ letter grades. However, it is worth noting that the concordance between the FSA and PISA was much stronger than that between PISA and the English 10 examination, both in statistical terms and conceptual foundations. These differences are most likely due to the larger emphasis of the English examination on literary content as well as the shorter length of the BC assessment in 2009.

[3] For reference, interpretations of the PISA proficiency levels from the OECD are provided in Annex A.

Table 4.1. Equivalences between interpretive standards (minimum cut-scores) on the BC English reading and PISA combined reading scales.

BC				PISA			
	Threshold	PISA Equivalent	(standard error)[1]		Threshold	BC Exam equivalent	(standard error)[2]
A	0.514	599.70	(51)	Level 6	698	1.0618	(0.30)
B	0.143	534.31	(47)	Level 5	626	0.6734	(0.24)
C+	-0.180	476.57	(50)	Level 4	553	0.2473	(0.24)
C	-0.541	409.42	(49)	Level 3	480	-0.1607	(0.18)
C-	-0.799	365.98	(48)	Level 2	407	-0.553	(0.19)
F	-	-		Level 1	335	-0.9838	(0.23)

Note:
1. Errors for thresholds are calculated by taking running of the surrounding 61 values.
2. Errors for thresholds are calculated by taking averages of the surrounding 31 values.

5. Recommendations

The major weakness of the linkage appears to be dissimilarity in the specifications of the two assessments. While the current methodology makes best use of the existing data given the constraints of the linking study, other alternatives may better leverage the data and optimize future linkages. One alternative, discussed in this section, involves the production of revised scales for the BC Exam, using static calibration of the BC exam test items against the scores produced by PISA. In order to calibrate the items, parameters for a two-parameter logistic model are estimated for each item using the PISA proficiency estimate as the outcome variable. For the two polytomously-scored writing items, the item responses were described using a multinomial logistic model. For greater certainty in estimation, the mean of the plausible values is used as the outcome (consistent with using the singular EAP estimate for the BC data).

Although their simplicity is tempting, omnibus statistical projection methods, such as multivariate regression, would be an unacceptable means of reweighting or recalculating scores using the BC item response data, because they tend to overcapitalize on random relationships in the data. For example, due to the covariance patterns among item responses, a multivariate regression would tend to balance the positive contributions of some items with negative contributions of

others. The resulting model would imply that some correct response would indicate weaker performance, which is non-intuitive.

The revised scale produced using the item responses has several limitations; for example, it would not satisfy the property of invertibility that is required for an equating-type linkage. In addition, it would not make use of the existing BC scale, which makes communication of the results difficult with respect to the existing reported BC results. However, it does represent a theoretically optimal method of creating test scores with the high numerical proximity to the PISA scale, based solely on the student responses to BC items.

5.1. Item characteristics

The statistics in Tables 5.1a and 5.1b describe the relationship of each BC multiple choice item to the PISA scale. The relationship is described by the two parameter logistic model in Table 5.1a, where the *a* parameter represents the maximum strength of the relationship of item performance to reading proficiency, and the *b* parameter describes the level of proficiency at which the item is most accurate. The multinomial logistic parameters in Table 5.1b represent the locations of thresholds describing the change in likelihood of choosing different item response scores.

Table 5.1a. Two parameter logistic model estimates for BC items on the PISA scale.

Item name	Parameter	Estimate	(Standard error)	Item name	Parameter	Estimate	(Standard error)
MCS01	a	0.270	0.030	MCS16	a	0.205	0.027
MCS01	b	-2.775	0.283	MCS16	b	0.778	0.155
MCS02	a	0.323	0.029	MCS17	a	0.287	0.028
MCS02	b	-1.184	0.119	MCS17	b	-0.502	0.100
MCS03	a	0.287	0.029	MCS18	a	0.091	0.026
MCS03	b	-2.015	0.196	MCS18	b	-1.506	0.500
MCS04	a	0.144	0.026	MCS19	a	0.400	0.031
MCS04	b	0.738	0.217	MCS19	b	-1.356	0.103
MCS05	a	0.291	0.029	MCS20	a	0.150	0.026
MCS05	b	-1.732	0.170	MCS20	b	0.831	0.217
MCS06	a	0.329	0.029	MCS21	a	0.104	0.026
MCS06	b	-1.188	0.117	MCS21	b	1.021	0.345
MCS07	a	0.237	0.027	MCS22	a	0.222	0.028
MCS07	b	-0.124	0.109	MCS22	b	-2.377	0.295
MCS08	a	0.228	0.029	MCS23	a	0.293	0.028
MCS08	b	-2.922	0.355	MCS23	b	-0.980	0.120
MCS09	a	0.264	0.028	MCS24	a	0.228	0.027
MCS09	b	-0.885	0.127	MCS24	b	-1.300	0.179
MCS10	a	0.302	0.030	MCS25	a	0.420	0.034
MCS10	b	-1.963	0.181	MCS25	b	-2.493	0.160
MCS11	a	0.196	0.028	MCS26	a	0.225	0.028
MCS11	b	-2.535	0.355	MCS26	b	-1.824	0.231
MCS12	a	0.333	0.030	MCS27	a	0.187	0.027
MCS12	b	-1.897	0.159	MCS27	b	-0.940	0.184
MCS13	a	0.360	0.031	MCS28	a	0.424	0.032
MCS13	b	-2.120	0.161	MCS28	b	-1.447	0.100
MCS14	a	0.320	0.029	MCS29	a	0.294	0.032
MCS14	b	-0.229	0.083	MCS29	b	-2.982	0.283
MCS15	a	0.344	0.034				
MCS15	b	-3.300	0.273				

Table 5.1b. Multinomial logistic model beta parameter estimates for BC writing items on the PISA scale.

Item name	Score category	Multinomial logistic model beta parameter	(Standard error)
Written response	4	0.068	0.014
Written response	5	0.389	0.016
Written response	6	0.955	0.017
Written response	7	1.071	0.017
Writing	4	-0.07	0.015
Writing	5	0.072	0.016
Writing	6	0.875	0.017
Writing	7	0.91	0.017

Using the parameters in Table 5.1, scores are estimated for each student using the same estimation methodology as the original BC scales (expected a posteriori using a single common prior distribution based on the empirical marginal distribution of the sample). The correlations between the original BC scale, linked scale and the PISA scale are in Table 5.2.

5.2 Results

Contrary to expectations, the projected estimates actually have a weaker relationship with the PISA scale scores than the original BC estimate (Table 5.2). Although the relationships are similar in magnitude, the projected results are less strongly related to the PISA results than the original BC results. Again, the weaker relationship is again likely due to the attenuating effects of measurement error, which is most likely due to the absence of the optimizing effect of within-sample calibration that is used to produce the original BC estimates.

The combination of simultaneous IRT calibration and score estimation with a single data set produces the minimum average error in the resulting estimated scores. When items are calibrated against a fixed score estimated from external data, the scores produced using the calibrated items will ultimately have more measurement error than the scores produced by the calibration/estimation process. Thus, although the recalibrated item parameters in principal produce scores that are more aligned

with the PISA scale, this benefit is more than offset by the increased measurement error.

Table 5.2. Correlations between original and linked scales.

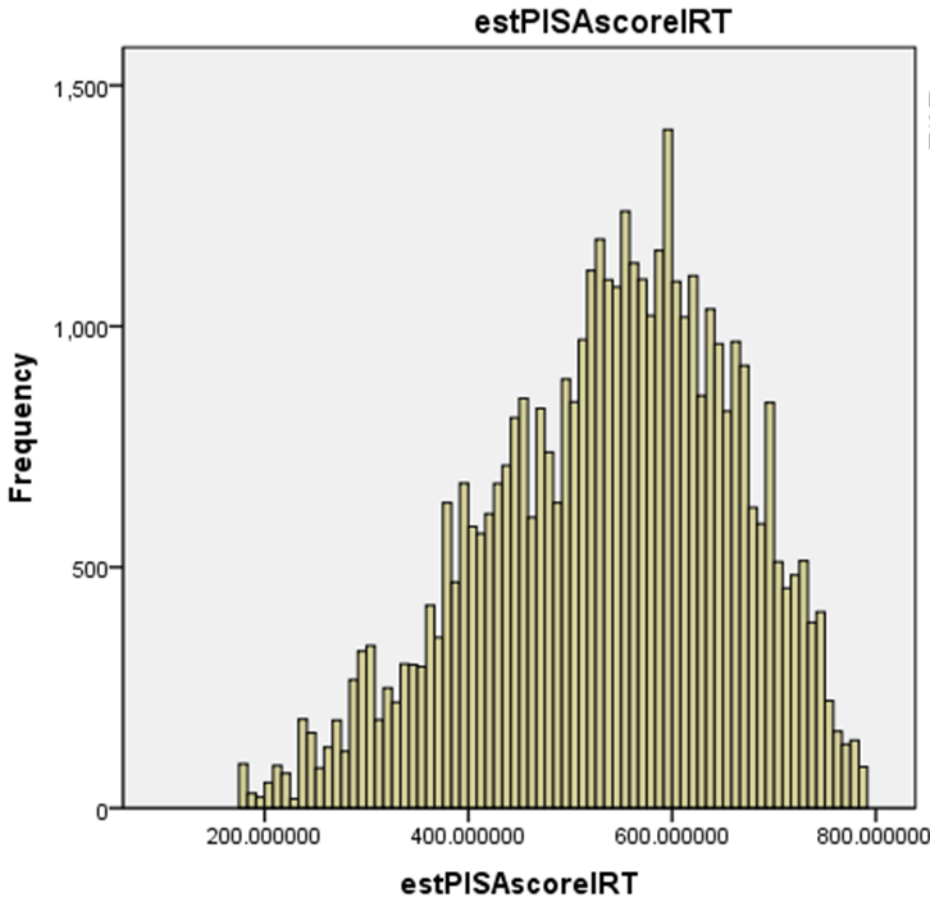
	PISA	BC	Projected PISA estimate
PISA	1	.609	.581
BC	.609	1	.838
Projected PISA estimate	.593	.880	1

The scale properties of each of the scales are in Table 5.3. Clearly, this method of linking does not share the same mathematical properties as the equipercentile method; few of the original distributional properties of the original scale are maintained in the linkage. However, although these results are not true ‘equating’ results, they provide insight into the relationship between the specifications of the tests and the performance of students. The BC test is clearly measuring a degree of proficiency that is oriented to a specific expectation, compared to PISA, which is primarily intended to provide meaningful variations in proficiency estimates across a wider range. The difference in specifications is evident from the ceiling effect in the projected scores (see Figure 5.2) when the items are rescaled according to PISA, which is greater than the ceiling effect observed in the original BC results.

Table 5.3 Distributional characteristics of original PISA scale and IRT-based projections of BC-to-PISA.

	PISA	Projected PISA estimate	BC
Mean	526.8	538.4	.10
Std. Deviation	89.6	123.2	.51
Skewness	-.293	-.416	-.274
Kurtosis	.045	-.328	-.164
Percentiles			
10	407.1	369.4	-.56
25	469.3	453.0	-.25
50	532.0	551.0	.13
75	589.3	630.0	.46
90	639.7	692.3	.76

Figure 5.2. Frequency histogram of IRT-based projections of PISA scores.



5.3 Summary

When the initial correlation between two measures is weak, as with the BC and PISA assessments, it may not be possible to construct a strong linkage. Although equipercentile methods are sufficient to reproduce the scale properties of one assessment onto another, they cannot increase the strength of the relationship between the two. Given the degree of statistical error in the linkage between the BC and PISA assessments, care should be taken to ensure that the linked results are not interpreted with the same degree of credibility as linkages created from assessments with more strongly-related scores or equivalent specifications.

In the current example, although the content review suggests that the two assessments are targeting a common construct, the measurements are not equivalent. However, the lack of content equivalence may be less significant than the large

measurement error in the BC assessment in reducing the equity between the assessments. For example, the correlation between the overall BC scale and PISA scale is 0.61; when the writing component (the largest contributor to construct differences) is removed, the correlation only 'improves' to 0.62. Given the significant error in the linking function, users should not proceed as if the error is negligible and interpret the results of the assessments as interchangeable after applying the linking function. Rather, the results of this linkage should be used primarily to aid stakeholders who have experience making comparisons with one or the other scale to make meaningful distinctions between different results on the scale with which they are less familiar.

By analogy, the recommended use of this linkage is similar to the situation of sporting fans interpreting the results from sports with which they are not familiar. For example, a marathon runner and a NASCAR enthusiast may not have any familiarity with the other sport. Therefore, in order for the two to have a meaningful discussion about one or the other, it might be reasonable for the former to ask how a particular feat of performance, such as placing *n*th in the Sprint Cup at Daytona, compares to running the Boston Marathon in *x* hours. Although the two performances have little in common, establishing the scale of one performance on the scale of the other facilitates a meaningful dialogue. Relative rates of changes or differences in performance may be meaningfully discussed simply based on the commonality of scale. Similarly, the linkage established in the current study should be used primarily to facilitate a dialogue on the comparative demands of performance expected of students by standards in British Columbia and international differences in reading proficiency. Once the space for discourse is established, the conceptual similarity between the two assessments may foster further dialogue about specific aspects of performance that *can* be compared.

References

Cartwright, F., Lalancette, D., Mussio, J. and Xing, D. (2003). *Linking provincial student assessments with national and international assessments.* Report no 81-595-MIE2003005. Ottawa, Canada: Statistics Canada.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*, New York: Springer.

Kolen, M.J., & Brennan, R.L. (1995). Test Equating. New York: Spring.

Linn, R.I., (1993) Linking results of different assessments. Applied Measurement in Education 6 (1), 83-102.

Ministry of Education, Province of British Columbia. (2011). Handbook of procedures. Grade 12 transcripts and examinations. Author. http://www.bced.gov.bc.ca/exams/handbook/1112/handbook_of_procedures.pdf

Mislevy, J. (1992) Linking Educational Assessments. Concepts, Issues, Methods and Prospects. Educational Testing Service, Princeton, NJ.

OECD. (2009), *PISA Data Analysis Manual: SPSS, Second Edition*, PISA, OECD Publishing. doi: 10.1787/9789264056275-en.

Annex A

Description of PISA proficiency levels on combined reading literacy scale: 2009	
Level 6 698	At level 6, tasks typically require the reader to make multiple inferences, comparisons and contrasts that are both detailed and precise. They require demonstration of a full and detailed understanding of one or more texts and may involve integrating information from more than one text. Tasks may require the reader to deal with unfamiliar ideas, in the presence of prominent competing information, and to generate abstract categories for interpretations. Reflect and evaluate tasks may require the reader to hypothesize about or critically evaluate a complex text on an unfamiliar topic, taking into account multiple criteria or perspectives, and applying sophisticated understandings from beyond the text. There is limited data about access and retrieve tasks at this level, but it appears that a salient condition is precision of analysis and fine attention to detail that is inconspicuous in the texts.
Level 5 626	At level 5, tasks involve retrieving information require the reader to locate and organize several pieces of deeply embedded information, inferring which information in the text is relevant. Reflective tasks require critical evaluation or hypothesis, drawing on specialized knowledge. Both interpretative and reflective tasks require a full and detailed understanding of a text whose content or form is unfamiliar. For all aspects of reading, tasks at this level typically involve dealing with concepts that are contrary to expectations.
Level 4 553	At level 4, tasks involve retrieving information require the reader to locate and organize several pieces of embedded information. Some tasks at this level require interpreting the meaning of nuances of language in a section of text by taking into account the text as a whole. Other interpretative tasks require understanding and applying categories in an unfamiliar context. Reflective tasks at this level require readers to use formal or public knowledge to hypothesize about or critically evaluate a text. Readers must demonstrate an accurate understanding of long or complex texts whose content or form may be unfamiliar.
Level 3 480	At level 3, tasks require the reader to locate, and in some cases recognize the relationship between, several pieces of information that must meet multiple conditions. Interpretative tasks at this level require the reader to integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. They need to take into account many features in comparing, contrasting or categorizing. Often the required information is not prominent or there is much competing information; or there are other text obstacles, such as ideas that are contrary to expectation or negatively worded. Reflective tasks at this level may require connections, comparisons, and explanations, or they may require the reader to evaluate a feature of the text. Some reflective tasks require readers to demonstrate a fine understanding of the text in relation to familiar, everyday knowledge. Other tasks do not require detailed text comprehension but require the reader to draw on less common knowledge.
Level 2 407	At level 2, some tasks require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognizing the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent and the reader must make low level inferences. Tasks at this level may involve comparisons or contrasts based on a single feature in the text. Typical reflective tasks at this level require readers to make a comparison or several connections between the text and outside knowledge, by drawing on personal experience and attitudes.
Level 1a 335	At level 1a, tasks require the reader to locate one or more independent pieces of explicitly stated information; to recognize the main theme or author‘s purpose in a text about a familiar topic, or to make a simple connection between information in the text and common, everyday knowledge. Typically the required information in the text is prominent and there is little, if any, competing information. The reader is explicitly directed to consider relevant factors in the task and in the text.
Level 1b 262	At level 1b, tasks require the reader to locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list. The text typically provides support to the reader, such as repetition of information, pictures or familiar symbols. There is minimal competing information. In tasks requiring interpretation the reader may need to make simple connections between adjacent pieces of information.
NOTE: To reach a particular proficiency level, a student must correctly answer a majority of items at that level. Students were classified into reading literacy levels according to their scores. Exact cut point scores are as follows: below level 1b (a score less than or equal to 262.04); level 1b (a score greater than 262.04 and less than or equal to 334.75); level 1a (a score greater than 334.75 and less than or equal to 407.47); level 2 (a score greater than 407.47 and less than or equal to 480.18); level 3 (a score greater than 480.18 and less than or equal to 552.89); level 4 (a score greater than 552.89 and less than or equal to 625.61); level 5 (a score greater than 625.61 and less than or equal to 698.32); and level 6 (a score greater than 698.32). Scores are reported on a scale from 0 to 1,000. SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2009.	

Annex B

Reading Assessment Comparison Study: PISA and English 10

The BC Grade 10 English Provincial Examination

Focus of the Provincial Examination

The Grade 10 English examination (En 10) is a provincial large-scale assessment which is based on the English Language Arts curriculum. It includes multiple-choice and written-response questions. While the curriculum addresses many aspects of English Language Arts, the Grade 10 English examination addresses only reading and writing.

The Grade 10 English examination passages include informational texts and literary texts, both prose and poetry. The informational passages may contain discontinuous text (e.g., timetables, recipes) and material presented in visual or graphical formats (e.g., charts, maps, diagrams, schedules, numerical data, cartoons, web pages).

Test Design

The examination design includes some “process” aspects of both reading and writing, and reading/writing connections. The examination booklet is based on a broad theme. In Part A, students are introduced to the theme. Students read three passages and answer nine multiple-choice questions on each passage. In Part B, students answer two multiple-choice questions based on two of the passages and a “synthesis” written-response question. In Part C, students read a short section “Getting Ready to Write” and a writing prompt based on the broad theme. Students do not need to refer to the reading passages when writing the composition.

English 10 Reading Comprehension

The Grade 10 English examination takes its definition of reading from the National Council of Teachers of English, (NCTE) 1997.

“Reading is the process of constructing meaning from a written text. It is an active process involving the constant interaction between the mind of the reader, the text, and the context.”

The definition reflects numerous current theories, which define reading as a constructive, interpretive, and interactive process. Meaning is constructed in the interaction between reader and text in the context of a particular reading experience, and culturally and socially derived expectations. The reader brings a repertoire of skills, cognitive, and metacognitive strategies, dispositions, and background knowledge to the task of reading. Texts are broadly defined to include print, graphic, and digital forms. This understanding of reading corresponds to that used in the English Language Arts curriculum and the BC Performance Standards for Reading.

Purpose of the English 10 Examination

The main purpose of the English 10 exam is to assess student achievement in reading and writing in relation to a provincial standard. The examination is required for graduation. The examination counts for 20% of the student’s final mark; classroom marks count for the other 80%. Two hours are allocated for the exam, with an additional one hour permitted. The test may be delivered electronically or on paper.

The written-response component of the En 10 exam is locally marked by practising teachers. Measures are in place to ensure the reliability and consistency of marking province-wide. Results may be used by individuals,

schools, school districts and the province as a source of information on achievement, and to guide future improvement.

The Comparison Study

One hundred unique items from the PISA 2009 Reading Assessment distributed across 13 booklets were coded with 30 items from the June 2009 En 10 exam. For the purpose of the comparison study with the PISA instrument, only the En 10 items relating to reading were coded. The En 10 item assessing writing was not included.

Findings

The majority of the 30 En 10 items were classified within the PISA assessment Framework, with the exception of items which assessed specific knowledge of literary terms from the List of Terms and Devices. Also, the single EN 10 written-response item requires a more in-depth response than PISA's written-response items and requires a synthesis of two fairly substantial, multi-paragraph reading passages.

In the reverse process, all of the 100 PISA items were also classified within the En 10 framework. The reviewers noted, however, that some of the PISA items reflected a greater emphasis on the application of reading skills than was evident in the En 10 criteria. As well, a small number of PISA "Retrieve information" items were found to be more low-level than any of the En 10 "Retrieve information" items, possibly because of the different populations who write the assessments.

Overall, both test instruments measure approximately the same reading skills.

Summary Chart

	PISA	English 10	Comments
Text Types	• Continuous • Non-Continuous • Mixed	• Literary Prose • Informational text with graphic • Poetry	*Notes:* -PISA includes a greater number of discrete, shorter texts -Poetry is not included in PISA -PISA includes more graphic and visual text - PISA includes some visual materials that consist of a graphic with little supportive text, whereas BC typically includes graphics that are supported with some text
Question Types	*3 Categories:* • Access and retrieve • Integrate and interpret • Reflect and evaluate	*4 Categories:* • Retrieve information • Recognize meaning • Interpret Texts • Analyze texts	• PISA's "Access and retrieve" items generally correlate with BC's "Retrieve information" • PISA's "Integrate and interpret" items generally correlate with BC's "Recognize meaning" and

			"Interpret Texts" • PISA's "Reflect and evaluate" items generally correlate with BC's "Analyze texts"
Number of Questions	Approx 53 to 62 MC and WR items	• 30 multiple-choice items • 1 written-response item	BC's written-response item requires a more in-depth response than PISA's written-response items and requires synthesis of two reading passages
Additional Questions on Reading Habits	PISA includes two questions at the end on reading for school purposes	En 10 does not include any additional questions on reading habits	