AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

ICSEA 2013: Technical Report

Measurement and Research

March 2014

1. Introduction

There is a substantial body of available research evidence that shows that the educational performance of students is related to, amongst many other things, non-policy, malleable characteristics of their family and school. That is, parental characteristics such as parental education and occupation, and school characteristics such as location and socio-economic background of the students it serves.

Schools serving cohorts of students with educationally advantaging background characteristics are likely to outperform – in terms of average NAPLAN scores –schools that serve cohorts of students that are less educationally advantaged.

In an effort to facilitate more meaningful comparisons of schools' academic performance, ACARA has developed the Index of Community Socio-Educational Advantage (ICSEA) and the student index of socio-educational advantage (SEA).

ICSEA identifies and quantifies many non-policy, malleable characteristics of a school and its student cohort and thus allows comparisons between schools that serve statistically similar students. Quantifying the characteristics that determine a student or school's academic performance is a complex undertaking. As with any real-world estimate, it involves a compromise between the information needed for a model to accurately predict performance and the information that is actually available for the model. No less important is the reliability and completeness of this available information, which greatly impacts the model's accuracy and stability. The ICSEA model developed by ACARA accounts for the following characteristics in a school and the students it serves:

- the level of socio-educational advantage (SEA) of each student and the average SEA of all students in a school. The SEA is derived from the parental education, occupation and qualification variables.
- whether a student is of Aboriginal or Torres Strait Islander descent
- the school's percentage of students of Aboriginal or Torres Strait Islander descent
- the accessibility or remoteness of a school in terms of its accessibility by road to services

In addition to ICSEA, ACARA reports the distribution of students in a school across four SEA Quarters representing a scale of relative disadvantage ('bottom quarter') through to relative advantage ('top quarter'). SEA Quarters provide contextual information about the socio-educational composition of the students in a school.

ICSEA and SEA Quarters have been calculated and released annually by ACARA since 2008. During these years the ICSEA model has been subject to a process of continuous refinement and enhancement. In reporting ICSEA, it is clear that its estimation for some schools can vary substantially from year to year. The reason for that variation appears to be primarily connected to shifts in the information provided by parents and schools. Though missing data issues overwhelm other potential sources of ICSEA variation, the potential for variation also exists within the modelling process itself. In an effort to improve the ICSEA model, ACARA has been exploring possible changes to the ICSEA calculation in consultation with external experts. This work is a continuous effort to adapt to the changes in data availability.

This report provides a technical description to the updated ICSEA and SEA Quarters calculation methodology. Section 2 provides a description of the past ICSEA calculation methodology and the overview of enhancement to the ICSEA and SEA Quarters calculation process introduced in 2013. The 2013 ICSEA calculation process is described in full detail in sections 3 and 4. The SEA Quarters calculation methodology is

presented in Section 5. Section 6 contains a description of the data sources used for the 2013 ICSEA and SEA Quarters calculation. A comparison between the results obtained using the previous and current calculation processes is presented in Section 7.

2. The ICSEA review

The previous ICSEA calculation was a two-stage process, completed in sequential order: (i) socio-educational advantage (SEA) component calculation and (ii) ICSEA construction.

In this process, the SEA index was calculated using the parental responses, from each of the two parents, to the following questions:

- school education
- non-school education (highest certificate awarded)
- occupational information

The proportion of students within a school responding to each category on each variable was then calculated. This process yielded over ten school-level proportions, with different levels of missing data due mostly to missing parental information responses. ICSEA was constructed through a regression analysis using the SEA component, the percentage of Aboriginal and Torres Strait Islander students and the geographical remoteness of the school as predictors of NAPLAN achievement. For each school, ICSEA was based on the six parental variables (direct ICSEA) and on the Australian Bureau of Statistics (ABS) Census data (indirect ICSEA). ICSEA was then calculated separately and equated. Lastly, a data sufficiency rule regarding the choice between direct and indirect ICSEA was applied to determine which ICSEA was published for each school.

There were two main concerns with the previous ICSEA calculation process; the un-expected year-to-year variability of ICSEA values for some schools, and an apparent misalignment of information between ICSEA and the within-school distribution of students across the SEA Quarters for some schools. Investigations showed that the level of missing data, the changing patterns in missing data and the variations in individual school cohorts from year to year were by far the main contributing factors to these unexpected variations, which have a greater impact for small schools. These factors, coupled with the complexity of the previous ICSEA calculation process, interacted with each other in diverse ways. These interactions significantly complicated attempts to isolate and eliminate their negative impact on the stability of ICSEA.

In order to increase the reliability of ICSEA values, it was proposed to improve the calculation process of the SEA component and to streamline and enhance the ICSEA regression analysis. Instead of treating the responses to the six parental background questions as separate indicators of socio-educational advantage (SEA) a modern measurement model was implemented to estimate a single indicator for a student's SEA.

This Item Response Theory (IRT) methodology is the same as that applied in the National Assessment Program – Literacy and Numeracy (NAPLAN) and other large-scale assessments including the Programme for International Student Assessment (PISA). A key advantage of this approach is that it produces more appropriate weights for scaling the parental responses (see Appendix A for the IRT weights for 2012 and 2013 parental background question responses).

A further improvement to the treatment of the parental responses in the estimation of the SEA is the inclusion of two variables which indicate whether or not each student's parent or guardian is in a non-paid occupation. Therefore, the IRT effectively models the SEA using eight parental variables, described in Section 6.

An additional benefit of the IRT approach is that the measurement model is able to

generate estimates of student advantage even when some parental data might be missing, thereby mitigating the impact of missing data on the year-to-year stability of ICSEA.

The enhanced process also explicitly accounts for the effect of any clustering of student educational advantage in a school, in addition to school geographical location and percentage of Indigenous students. This is achieved through the use of a multi-level modelling approach that appropriately combines the respective influences of student and school-level factors to calculate ICSEA.

3. Enhanced SEA estimation process

In the current ICSEA model, the socio-educational advantage is conceptualised as a latent variable and responses to parental questions are treated as indicators of that variable. Item Response Theory (IRT) is a modern measurement method used to measure latent constructs that are not directly observable and must therefore be measured indirectly. IRT is widely used in the field of educational and psycho-social testing.

The Generalised Partial Credit Model (GPCM), a specific form of an IRT model, was developed to handle the scaling of ordered response indicators of the type produced by parental background question responses (Adams and Macaskill, 2012; Adams, Wilson and Wang, 1997). Consequently, the GPCM was implemented to obtain a single indicator of students' SEA and to appropriately scale the eight parental background variables.

The current method for SEA estimation consists of three steps, implemented using the ACER ConQuest software (Adams and Macaskill, 2012):

- step 1: SEA scale calibration using the Generalised Partial Credit Model
- step 2: generation of plausible values for student SEA for the NAPLAN National Report dataset
- step 3: generation of plausible values for student and school SEA for the SBD dataset

It is worth noting that the current model also accounts for the response: "not in paid occupation in the last 12 months" as additional indicators of SEA.

3.1. Step 1: Generalised Partial Credit Model

In step 1, the six responses to parental background questions, plus the two generated variables indicating whether a parent is in a non-paid occupation are treated as items calibrated using the GPCM and anchor item parameters are generated. For this step, all available responses from the NAPLAN dataset are used.

As discussed, GPCM provided an appropriate scaling or weighting of the different parental responses. An example of this weighting, for question about parent 1 occupation, is shown in Table 1. The score column shows the traditional weighting where each response category receives an equidistant score on a scale and the GPCM column shows the appropriate scaling and weighting provided by the IRT model. Appendix A provides the scaled weights for the eight parental variables used in 2013 and modelled for 2012 dataset.

Table 1:

Example comparison of item weightings using the GPCM approach for parent 1 occupation response categories

Response	Score	GPCM
Machine operator	0	0
Tradesperson / clerk / sales	1	0.88
Professional / manager	2	1.99
Senior manager	3	3.53

3.2. Step 2: Student SEA – NAPLAN dataset

In this step SEA estimates for each student in the NAPLAN dataset is drawn using the plausible values methodology. The IRT model for the extraction of plausible uses the GPCM parameters for the parental questions obtained in the step 1 and is conditioned on the following variables for each student:

- NAPLAN reading weighted likelihood estimate score (*wler*)
- a dummy variable indicating whether wler is missing
- school location[1]
- Aboriginal and Torres Strait Islander status (ATSI)
- a dummy variable indicating whether the student's Aboriginal and Torres Strait Islander status information is missing

The plausible values imputation methodology enables that even if a student is missing one or more responses to the parental background questions, their level of advantage can be estimated. Thus, the students SEA estimate can be obtained without the need to impute missing responses to parental questions. The implemented IRT method not only enhances the extraction of SEA but also provides a parsimonious method for the treatment of missing data.

The resulting set of five SEA plausible values ($SEA_{student\ i}$, for i : [1, 5]) for each NAPLAN student are used as student level estimates in the multilevel modeling that provides the final ICSEA regression equation, as described in sections 4.

3.3. Step 3: Student SEA - SBD dataset

In this step, plausible values estimates for SEA are drawn for all students in the SBD dataset. The same GPCM parameters are used however the conditioning variables include only the following:

- the school average NAPLAN reading score based on weighted likelihood estimates (schwler)
- school location[1]
- Aboriginal and Torres Strait Islander status (ATSI)
- a dummy variable indicating whether the student's Aboriginal and Torres Strait Islander status information is missing

The resulting set of five SEA plausible values ($SEA_{student\ i}$, for i : [1, 5]) for each SBD student is used as one of the components of the ICSEA calculation equation in Section 3.2. Also, the SBD SEA plausible values are the key component to the SEA Quarters

[1] School location can have four categories: metropolitan, provincial, remote and very remote.

calculation described in Section 4.

4. Enhanced ICSEA calculation

The 2013 ICSEA calculation methodology is an enhanced statistical model motivated primarily by the need to better address the intrinsic year to year variability of ICSEA and to provide a better treatment of missing data.

The conceptual ICSEA regression model where the student SEA component, the percentage of Aboriginal and Torres Strait Islander students, and the geographical remoteness of the school are regressed on schools' NAPLAN performance, has been retained. However, the regression model has been enhanced to explicitly account for the contribution of a school's cohort SEA component in the prediction of the school's performance in NAPLAN.

Section 4.1 describes the multilevel model used to generate the regression coefficients for the calculation of ICSEA. These regression coefficients are used in Section 4.2 to calculate the schools' ICSEA values.

4.1. Multilevel modelling

The previous ICSEA model was a simple regression model which did not account for the hierarchical structure of the data, i.e. it ignored the effects grouping students within their school. The problem of students nested in schools is a classic problem of hierarchical or nested data. Multilevel models recognise and account for the existence of data hierarchies by allowing for residual components at each level in the hierarchy (Goldstein, 2003). The enhanced ICSEA calculation uses a two-level model which allows for the grouping of students within a school.

$$NAPLAN_{performance} = \beta_0 + \beta_1 * SEA_{student} + \beta_2 * ATSI + \beta_3 * missingATSI + \beta_4 * SEA_{school} + \beta_5 * percentageATSI + \beta_6 * ARIA + \upsilon + \varepsilon \quad (1)$$

The dependent variable in this multilevel model is the averaged performance of students across five NAPLAN tests. Before averaging the NAPLAN plausible values are standardised within each year level for each test.

The $SEA_{student}$ is the IRT estimate of student-level SEA extracted from the NAPLAN dataset. The Aboriginal and Torres Strait Islander status of the student is represented by the acronym ATSI and the lack of the information about the students' Indigenous status is denoted by term *missingATSI*.

The school level component of the multilevel model, SEA_{school}, is calculated as the average IRT SEA estimate for all the students in a school obtained from the SBD dataset. The percentage of Aboriginal and Torres Strait Islander students in the school is represented by *percentageATSI* and Accessibility/Remoteness Index of Australia (*ARIA*) is the remoteness of the school.

The last two terms are the random components of the two-level model accounting for unexplained variation between school-level residuals υ and student-level residuals ε.

A fundamental assumption in the modelling of ICSEA is that the same mechanism governs the relationship between SEA and NAPLAN performance in all Australian schools. Consequently, only the fixed effect coefficients are extracted from this multilevel model and used in the final ICSEA calculation formula.

4.2. Calculation of ICSEA values

Once the fixed effect multilevel model regression coefficients are obtained, β_j (for j : [0, 6], the student-level ICSEA is calculated for every student in the SBD dataset as shown in equation 2:

$$ICSEA_{student} = \widehat{\beta_0} + \widehat{\beta_1} SEA_{student} + \widehat{\beta_2} ATSI + \widehat{\beta_3} missingATSI + \widehat{\beta_4} SEA_{school} + \widehat{\beta_5} percentageATSI + \widehat{\beta_6}\ ARIA \quad (2)$$

Both the student-level SEA component, $SEA_{student}$, and the school-level SEA component, SEA_{school} are derived using the SBD as described in Section 3.3. The school-level SEA is calculated by averaging student-level SEA estimates for all students in a school. Fitting this model produces the $ICSEA_{student}$ value for each student in the SBD dataset, which is the predicted value of each student's NAPLAN score. The student's raw ICSEA values within every school are in turn averaged to obtain each school's $ICSEA_{raw}$:

$$ICSEA_{raw_{i,l}} = \frac{1}{Nsch_l} \sum_{k=1}^{Nsch_l} ICSEA_{student_{i,k}} \quad (3)$$

for k : [1, $Nsch$] where $Nsch_l$ is the number of students recorded for school l in the SBD dataset. Next, the schools' raw ICSEA values ($ICSEA_{raw}$) are standardised so that the distribution of ICSEA values has a mean of 1,000 and a standard deviation of 100:

$$ICSEA_{i,l} = \left(\frac{ICSEA_{raw_{i,l}} - \overline{ICSEA_{raw_{i,l}}}}{\sigma_{ICSEA_{raw_{i,l}}}}\ x\ 100 \right) + 1000 \quad (4)$$

The final ICSEA reported on the *My School* is the mean of the five standardised ICSEA values as shown in the equation (5).

$$ICSEA = \frac{1}{5} \sum_{i=1}^{5} ICSEA_{i,l} \quad (5)$$

5. SEA Quarters calculation methodology

The SEA Quarters are conceptualised as a broad representation of a school's student distribution. The previous SEA quarter calculation methodology was based on a process in which the responses to parental questions were treated in a conceptually different way to the ICSEA calculation process (as the simple rating scale in the former and as the discrete proportions in the latter case). As of 2013, the current SEA estimation calculation allows the allocation of all students in a school to their corresponding Quarter, rather than only those students in the NAPLAN data set.

The benefit of the current process is twofold. First, the range and the distribution of the scale used to calculate SEA Quarters are significantly increased, allowing for a more representative distribution of the SEA level of a school; second, the SEA Quarters and ICSEA are now calculated using the same SEA estimated by the IRT model.

Consequently, the current calculation methodology significantly reduces the misalignment between ICSEA and SEA Quarters.

The current SEA Quarters calculation is based solely on the student-level SEA estimate for all students in the SBD dataset. A key concept of this approach is that any two or more students sharing the same level of SEA will be assigned in the same quarter, independent of the school SEA cohort effect.

The calculation of the schools' SEA Quarters is derived by estimating the quartile cut-off values from the SBD-based student-level SEA estimates. Using these cut-off values, each student is allocated into their corresponding quarter. The resulting distribution and percentages of students in each quarter are then calculated for every school.

This process is performed independently for each of the five sets of SBD-based student-level SEA plausible values. The final distribution of students in SEA quarters is the averaged distribution obtained from the five sets of plausible values.

6. Data preparation and data sources

When enrolling a child in school all parents are asked which of the following options best describes their occupation, and the school education and non-school education levels they achieved. All states and territories, Government Education Departments and Catholic system jurisdictional authorities provided ACARA with the parental background data for all students in their schools.

For some non-government systemic schools and most independent schools, parental background data were only available for students who participated in NAPLAN. Those data were collected and provided to ACARA by the Test Administration Authority in each state and territory.

In 2013, there were 1,134 (~13%) schools that did not provide any student data as part of the *My School* data collection. For these schools, their 2012 and 2013 NAPLAN datasets were merged and used in the SBD dataset. A brief summary of the data prepared by ACARA for the purposes of the 2013 ICSEA and Quarters calculations is presented in Table 2.

Table 2:

Number of students and school in 2013 of NAPLAN and SBD datasets.

Category	NAPLAN dataset	SBD dataset
# of records	1,107,732	3,405,398
# of schools	8,849	8,849
# of schools with no SBD data	–	1,134

In order to calculate ICSEA and the SEA Quarters, the first step required is to recode the available student background information using the nationally agreed definitions of student background characteristics for parental responses, school geographical location information and information as to whether the student is of Aboriginal or Torres Strait Islander descent. See Australian Curriculum, Assessment and Reporting Authority (2012) for further information regarding the recoding conventions.

Next, indicators for cases of missing data values for every student record are generated. This includes the generation of dummy variables indicating the absence of NAPLAN (reading) performance or Aboriginal or Torres Strait Islander information for a particular student.

Additionally, the two indicators of whether each parent or guardian were in a paid occupation is set as 0 if the parent was not in a paid occupation or 1 if the parent was in a paid occupation.

7. Overview of 2013 ICSEA calculations results

The current ICSEA regression model includes both student-level SEA and school-level SEA. The investigations conducted by ACARA and partners have shown that this model increases the predictive power of ICSEA compared to the previous model, as well as reducing the year-to-year variability of ICSEA and SEA Quarters values. This finding is consistent across the results yielded using the current model with the 2012 and 2013 datasets.

Figure 1 shows the comparison of the current model using data for the year 2013 against the year 2012 using the previous model (Figure 1a) and the current model (Figure 1b). As can be seen from the two figures, the 2013 model has greater year to year stability than the previous model.

Figure 1a: Comparison of published 2013 ICSEA values against published 2012 ICSEA values. The black line is the line of best fit and the black cross shows the median in the horizontal and vertical axes. The box-plots at the top and left ends of the graph is a representation of each distribution. The box-plot denote the median, the interquartile range, whiskers at 1.5 interquartile range and the individual points considered as outliers (outside the whiskers).

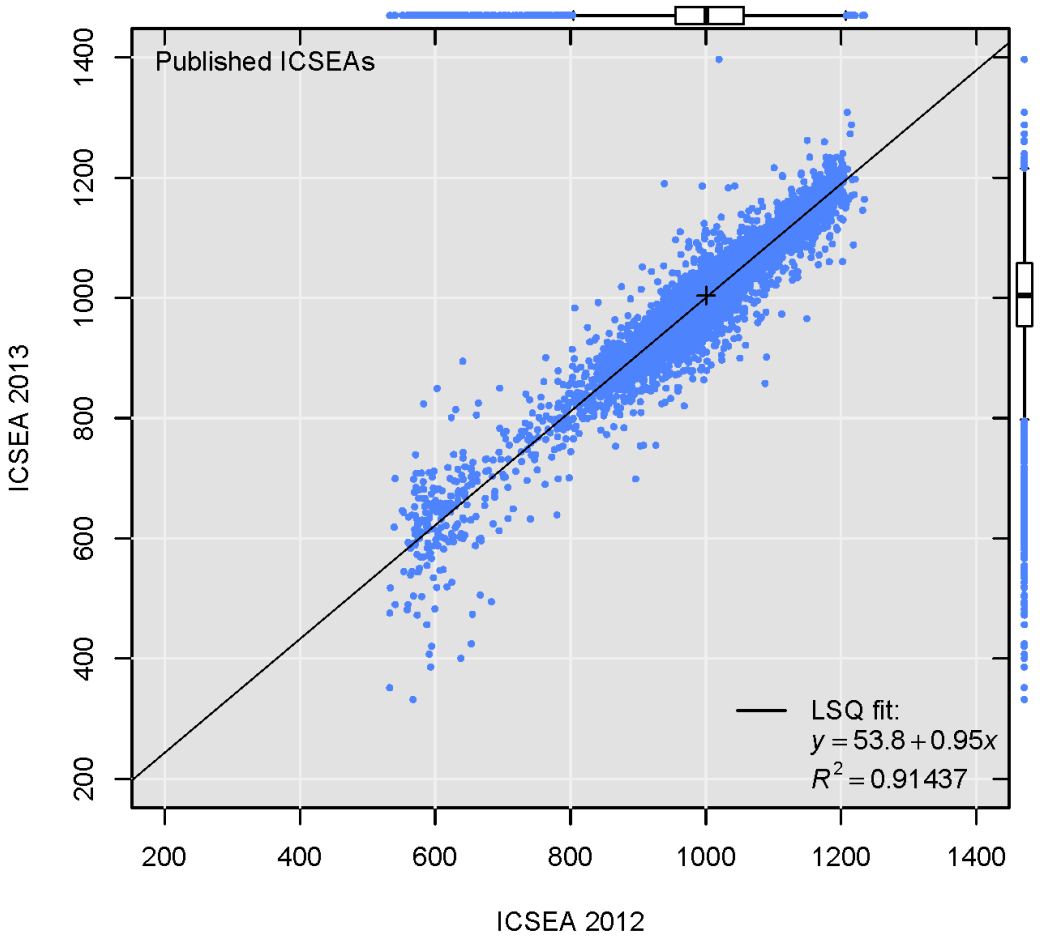


Figure 1b: Comparison of published 2013 ICSEA values (current model) against non-published 2012 ICSEA values (old model)

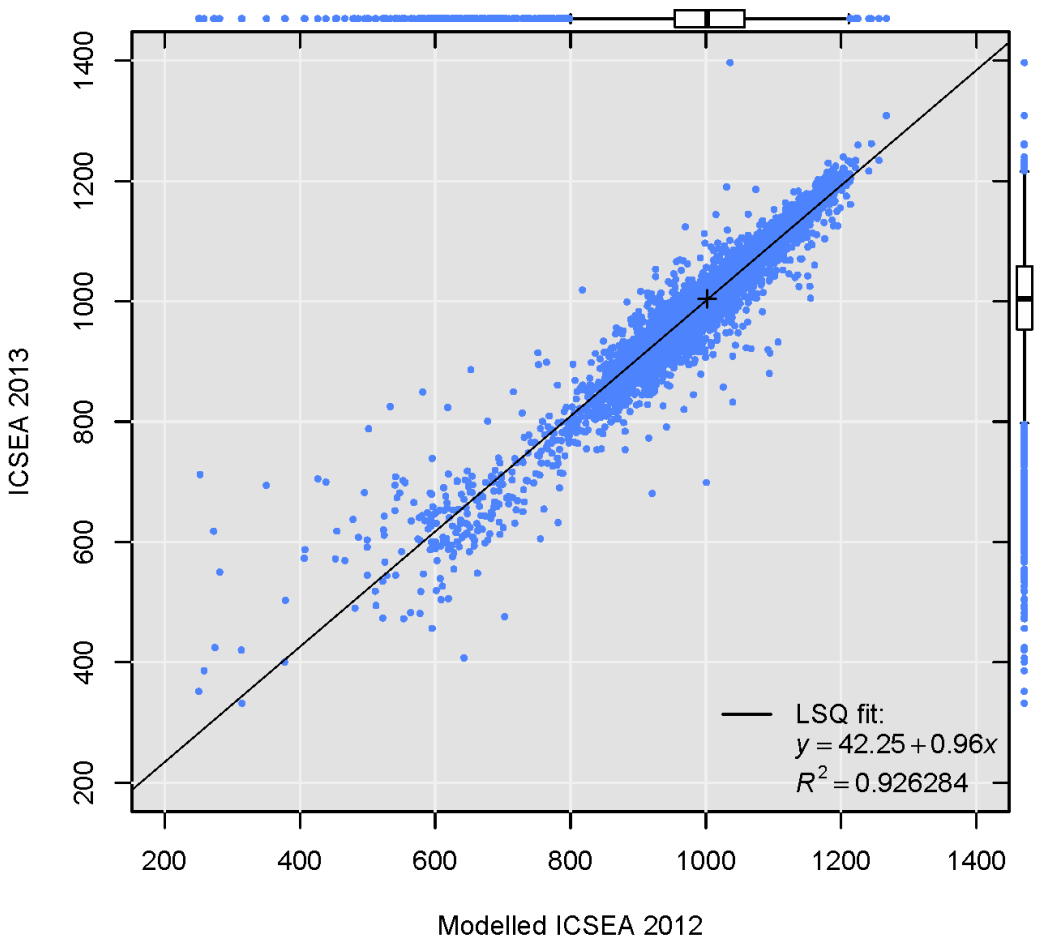


Figure 2 shows the schools' ICSEA values against their school-level NAPLAN performance for the published years 2012 and 2013. The explained variation (R^2) in NAPLAN performance for 2013 is 81% while for 2012 is 72%.

Figure 2a: Published 2012 ICSEA values (old model) against NAPLAN performance.

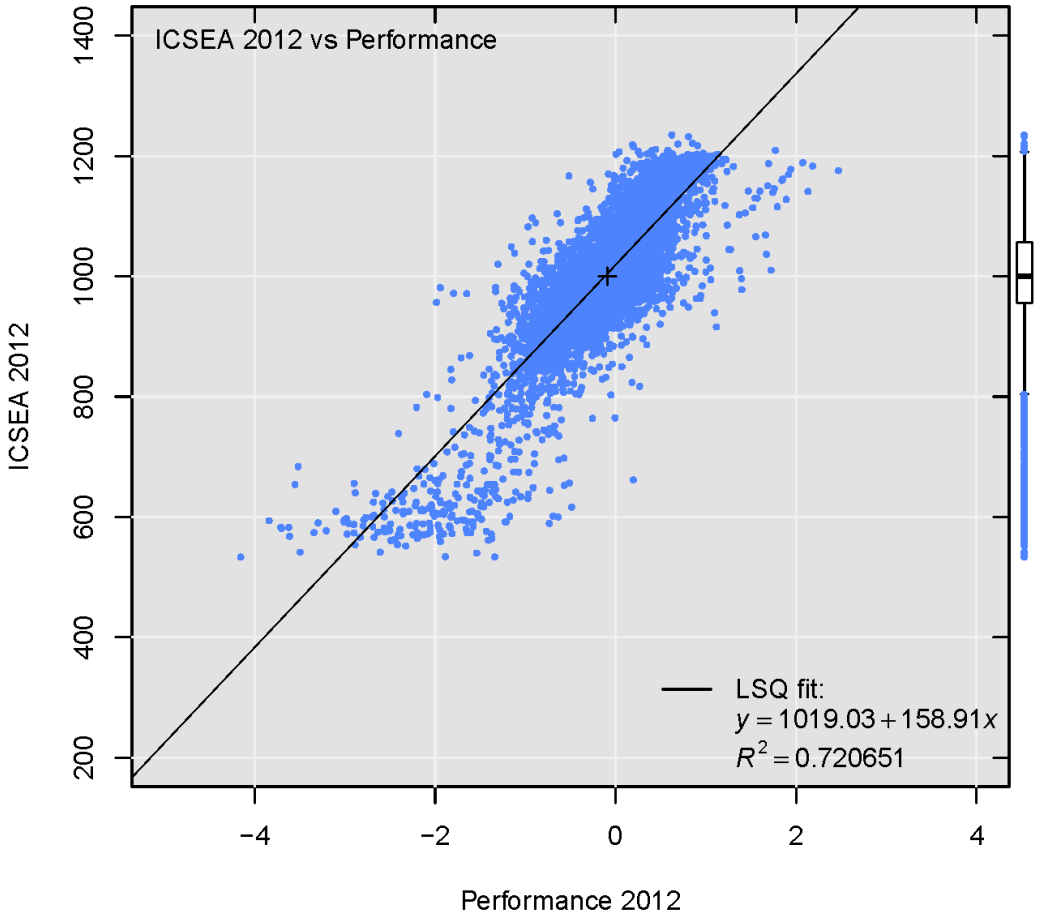
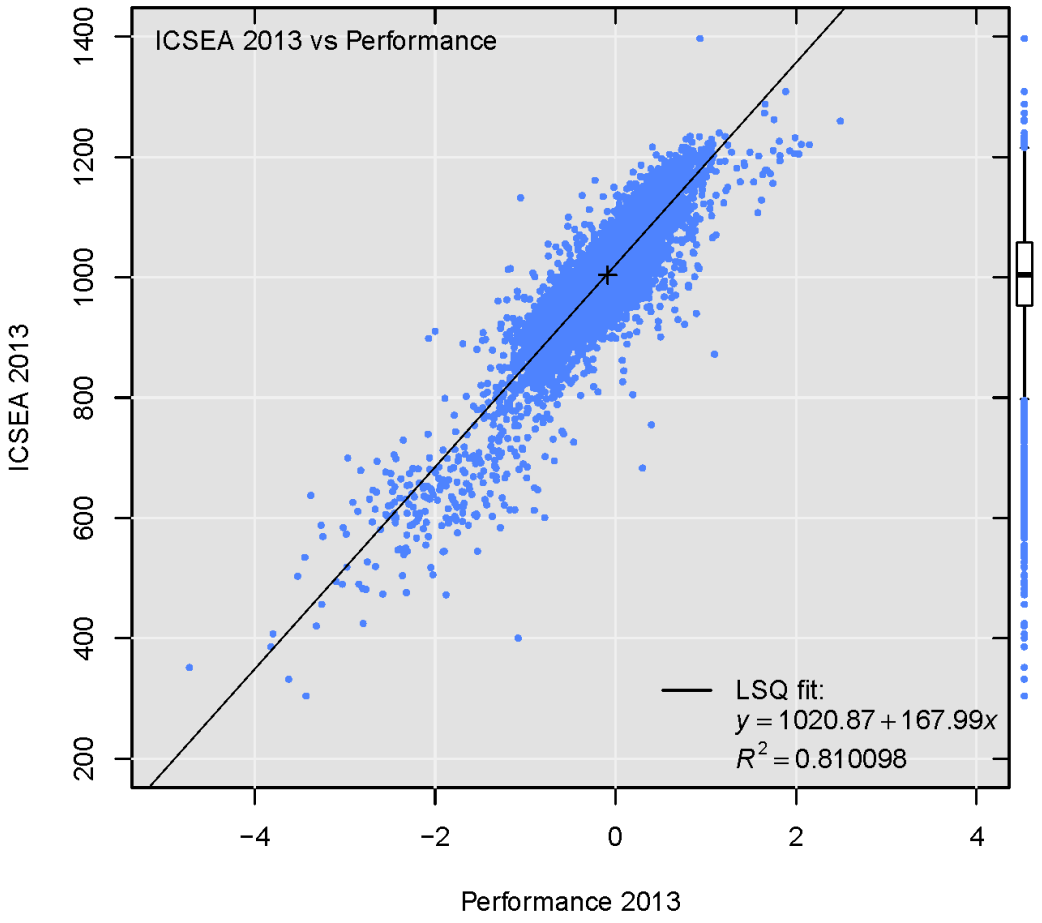


Figure 2b: Published 2013 ICSEA (current model) against NAPLAN performance.



The SEA Quarters are a broad representation of a school's student distribution. As of 2013, this distribution is based solely on each student's level of socio-educational advantage estimation. This means that the school effect is excluded from the Quarters distribution. Thus, the SEA Quarters provide contextual information of a school's socio-educational demographics. Figure 3 shows a comparison between the published Quarters in 2012 (old model) and the 2013 Quarters (current model). The vertical axis values on both graphs were calculated using the following formula:

*value = percentage Q1 * 1 + percentage Q2 * 2 + percentage Q3 * 3 + percentage Q4 * 4*

The distribution obtained using the 2012 data with the current model (not shown) is also consistent with the 2013 data in the right hand-side panel.

Figure 3a: SEA Quarters against ICSEA using the current model with 2013 dataset.

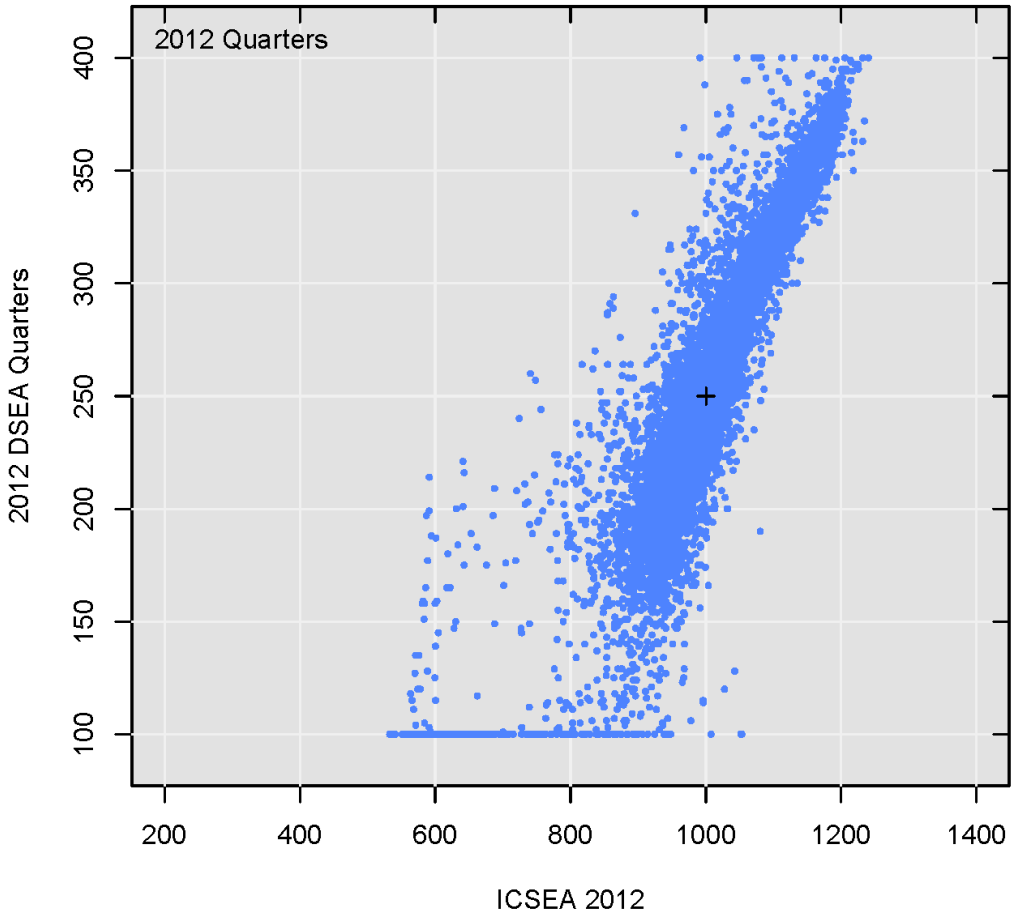
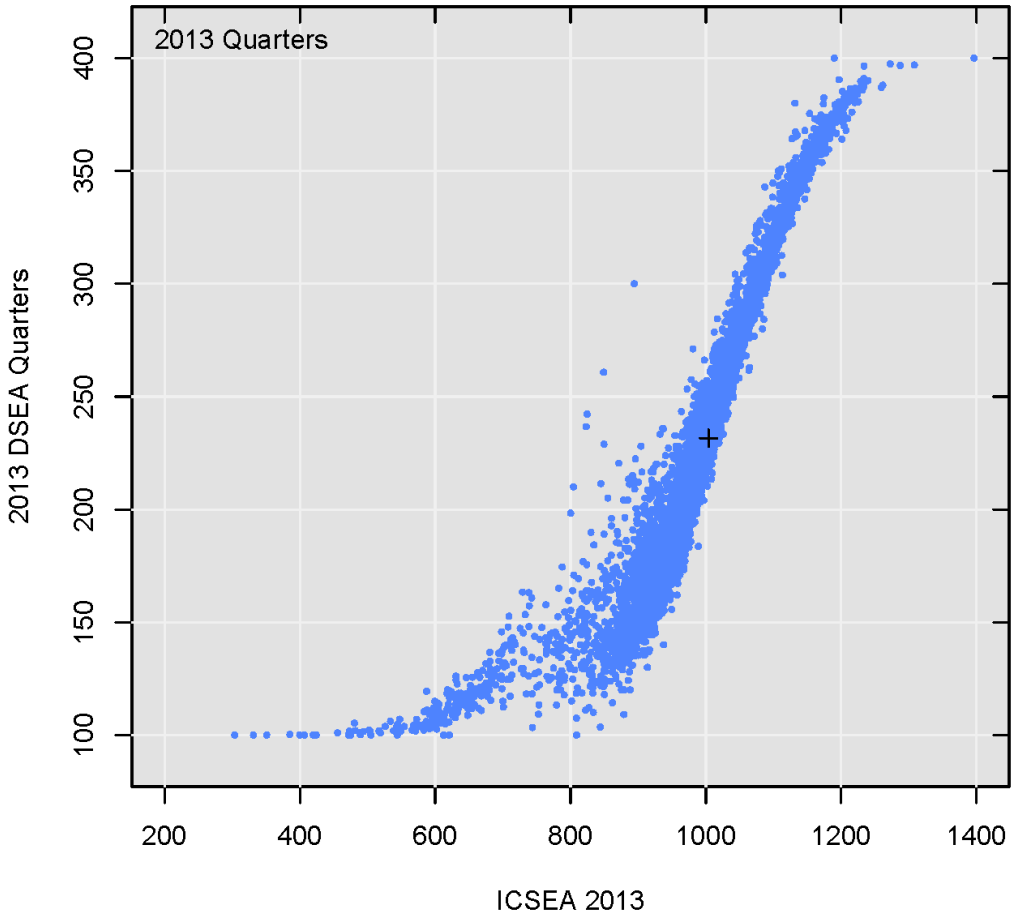


Figure 3b: SEA Quarters against ICSEA using the old model with 2012 dataset.



Appendix A: Generalised Partial Credit Model parameters

Tables 3 -10 contain the parameter scaling factors obtained by the GPCM for the 2013 ICSEA and SEA Quarters calculation (Section 3.2). The Response column shows the responses available to the parental question; the Count column shows the number of instances of a particular response in 2013; the percentage column shows the percentage that the number of instances amounted to in 2013; Score Value column provides the unweighted scores for each response category, and the 2013 and 2012 columns show the item weightings using the GPCM approach for each year.

Table 3: Parent 1: school education

Response	Count	%	Score	2013	2012
Year 9 or equivalent	53,812	6	0	0	0
Year 10 or equivalent	190,294	20	1	1.24	1.29
Year 11 or equivalent	114,981	12	2	1.48	1.52
Year 12 or equivalent	593,643	62	3	3.06	3.11

Table 4: Parent 2: school education

Response	Count	%	Score	2013	2012
Year 9 or equivalent	52,880	6	0	0	0
Year 10 or equivalent	197,362	24	1	1.18	1.19
Year 11 or equivalent	98,025	12	2	1.46	1.44
Year 12 or equivalent	483,305	58	3	3.06	2.99

Table 5: Parent 1: non-school education

Response	Count	%	Score	2013	2012
No non-school education	234,636	27	0	0	0
Certificate I–IV inc. trade certificate	239,419	27	1	0.92	1.01
Advanced diploma / Diploma	142,771	16	2	2.29	2.54
Bachelor degree or above	261,033	30	3	4.04	4.47

Table 6: Parent 2: non-school education

Response	Count	%	Score	2013	2012
No non-school education	165,014	21	0	0	0
Certificate I–IV inc. trade certificate	276,885	36	1	0.94	0.99
Advanced diploma / Diploma	104,499	14	2	2.68	2.75
Bachelor degree or above	225,015	29	3	4.72	4.87

Table 7: Parent 1: occupation

Response	Count	%	Score	2013	2012
Machine operator	140,290	21	0	0	0
Tradesperson / clerk / sales	201,347	30	1	0.88	0.89
Professional / manager	169,143	26	2	1.99	2.00
Senior manager	151,399	23	3	3.53	3.66

Table 8: Parent 2: occupation

Response	Count	%	Score	2013	2012
Machine operator	164,109	22	0	0	0
Tradesperson / clerk / sales	204,989	27	1	0.85	0.83
Professional / manager	204,571	27	2	1.98	1.84
Senior manager	180,054	24	3	3.85	3.55

Table 9: Parent 1: non-paid occupation

Response	Count	%	Score	2013	2012
in non-paid occupation	237,768	26	0	0	0
in paid occupation	662,179	74	1	0.63	0.63

Table 10: Parent 2: non-paid occupation

Response	Count	%	Score	2013	2012
in non-paid occupation	54,595	7	0	0	0
in paid occupation	753,723	93	1	0.80	0.79

Appendix B: Multilevel regression coefficients

For 2013, the regression coefficients obtained in Section 4.1 are shown in Table 11:

Table 11: Multilevel regression coefficients for 2013.

Variable		pv1	pv2	pv3	pv4	pv5
β_0	intercept	-0.017	-0.017	-0.017	-0.017	-0.017
β_1	$SEA_{student}$	0.224	0.223	0.224	0.224	0.224
β_2	ATSI	-0.327	-0.328	-0.327	-0.326	-0.327
β_3	missingATSI	-0.188	-0.19	-0.189	-0.192	-0.187
β_4	SEA_{school}	0.291	0.292	0.290	0.290	0.290
β_5	percentageATSI	-0.007	-0.007	-0.007	-0.007	-0.007
β_6	ARIA	-0.004	-0.004	-0.004	-0.004	-0.003

References

Adams, R. J., Wilson, M. R. and Wang, Wen-chung. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 1, 1-23.*

Adams, R. J., & Macaskill, G. (2012). *Score Estimation and Generalised Partial Credit Model*. ACER Conquest Technical Note. Melbourne, Australian Council for Educational Research. Retrieved from: http://www.acer.edu.au/documents/Conquest-Notes-6-ScoreEstimationAndGeneralisedPartialCreditModels.pdf

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed). Kendall's Library of Statistics. London: Arnold.

Australian Curriculum, Assessment and Reporting Authority. (2012) Data Standards Manual: Student Background Characteristics, Sixth edition. Retrieved from: http://www.acara.edu.au/verve/_resources/DSM_1.pdf