Review of the impact of correlations in the ACT scaling process

Professor Rob J Hyndman

B.Sc. (Hons), Ph.D., A.Stat.

Report for

ACT Board of Senior Secondary Studies

15 February 2011

Contact details:

Telephone:	0458 903 204
Email:	robjhyndman@gmail.com
Web:	robjhyndman.com

ABN: 99 118 173 027

Contents

1 Background

The ACT Board of Senior Secondary Studies (BSSS) uses a scaling process for school-based course scores in T courses for the purpose of calculating Australian Tertiary Admission Ranks (ATARs). The process involves calculating a **scaled score** corresponding to each **school score**. For each student, the best 3.6 scaled scores (called the **aggregate**) are then used, along with the student's score on the ACT Scaling Test (**AST**) to obtain the **scaling score** for each student. While scaling scores are used in the scaling process, it is the aggregate scores that are converted to ATARs.

Recently, a study suggested that the scaling process depended on good correlations between the AST and school-based course scores, and the main purpose of this review is to investigate these correlations and their effect on scaling.

The BSSS has implemented special provisions for students from a culturally and linguistically diverse (CLD) background. In particular, CLD students undertake a modified AST and some tasks are marked against modified criteria. A secondary purpose of this review is to investigate the impact of these special provisions for CLD students on the scaling outcomes.

1.1 Terms of reference

Advice is sought on:

Correlations between school-based course scores, the AST and scaling scores. Specifically:

(a) *How the correlations in the ACT scaling process compare with the correlations in scaling in other jurisdictions?*

(b) *What levels of correlation are deemed acceptable for scaling purposes?*

(c) *The effects on ATARs of courses where there are low correlations in the scaling between school-based course scores and the AST.*

(d) *The effectiveness of current measures used by the BSSS to enhance correlations between school-based assessment and the AST.*

(e) *Additional/replacement measures that could be implemented to increase correlations in the scaling process.*

(f) *What should be done with courses that do not/can not reach an acceptable level of correlation for scaling purposes?*

CLD Procedures
Specifically

(g) *The impact of the presence of CLD students on the outcomes for scaling groups and on overall college results.*

(h) *Abolition of the modified AST papers for CLD students and use of other processes, such as the aberrant score policy, to address correlations for CLD students and their impact on scaling groups.*

Although not covered by the terms of reference, I will also explore the correlations between the AST and aggregate scores as these are directly relevant to the calculation of the ATAR.

1.2 Structure of report

Before discussing these items, I first describe in more detail the scores available for each ACT student in Section 2 and the scaling algorithm used to compute these scores in Section 3. To my knowledge, this is the only complete description of the scaling algorithm currently in use, and so this should be a useful resource separate from the rest of the report.

Section 4 describes the data available to me for this review. In Section 5, I validate the scaling algorithm by attempting to recompute the scaling results for 2007–2009.

To properly consider the issues raised in the terms of reference, I need to be able to generate realistic data with similar characteristics to the real scaling data from 2007–2009. I describe a procedure in Section 6 whereby such artificial data can be produced.

Many of the terms of reference concern the correlations between school scores and AST scores. I consider these correlations in Section 7 and the effect of these correlations on the aggregate (and therefore on the ATAR) in Section 8. The effectiveness of existing measures to improve these correlations is discussed in Section 9. Section 10 explores the impact of CLD students on scaling.

Finally, the conclusions for each of the terms of reference are summarised in Section 11.

A red marker is given in the right-hand margin to highlight when each of the terms of reference is explicitly addressed (from Section 7 onwards).

2 ACT scores and scaling principles

2.1 Scaled scores

For each course undertaken, students in the ACT receive school-based scores with a system mean of 70 and a standard deviation of 12. The scaling process involves a linear transformation of school-based course scores to scaled course scores for each scaling group. The scaled course scores have a system mean of 150 and a standard deviation of 25 (taken across all students in the year).

A scaled course score is calculated as follows. Let $x_{i,j,k}$ be the school course score for student i in course j and college k, and $y_{i,j,k}$ be the corresponding scaled course score. Then

$$y_{i,j,k} = a_{j,k} + b_{j,k}x_{i,j,k}. \tag{1}$$

The values of $a_{j,k}$ and $b_{j,k}$ are to be estimated from the available data as discussed in Section 3.

2.2 Aggregate scores

For most students, the aggregate score is the sum of the best three major scaled course scores plus 0.6 of the next best scaled course score. Where a score corresponds to a major-minor study or a double major study, the course weight is 2.0 or 1.6 respectively.

For students studying an abridged mature age package, the aggregate is sum of the best three minor scaled scores multipled by 1.2.

For students studying an older student package, the aggregate is 3.6/2.6 times the sum of the best two scaled scores of major courses and 0.6 of the next best scaled score.

I shall denote the aggregate score of student i in college k by $z_{i,k}$.

2.3 AST scores

The AST consists of three tests: a Multiple Choice Test, a Short Response Test and a Writing Task. From these three tests, four AST components are calculated: Quantitative Multiple Choice; Verbal Multiple Choice; Short Response; and Writing Task. Each component of the AST is standardised to a system mean of 150 and standard deviation 27.5. For each student, the AST component results are added together in a weighting determined each year to maximise the correlation between the AST and aggregate scores across the system. The total AST results (AST Score) are then re-standardised across the system to a mean of 150 and standard deviation 27.5.

I shall denote the (standardized) AST score of student i in college k by $c_{i,k}$.

2.4 Culturally and linguistically diverse background

With the AST, provisions are made for students who meet the Board's definition of being from a Culturally and Linguistically Diverse (CLD) background. These students sit slightly modified Multiple Choice and Short Response Test papers and have their Writing Task marked against different criteria.

2.5 Aberrant values

The aberrant value for each student is calculated after the AST Scores have been standardised to a system mean of 150 and standard deviation of 27.5. The Culturally and Linguistically Diverse (CLD) background students and the non CLD students are treated as two distinct groups for the calculation of the aberrant value. In this process the aim is to reflect the difference between the average scaled score for CLD students and non CLD students across the system in the average of the two groups' AST Scores, whilst at the same time keeping as many AST Scores as possible in the OCS process.

Let $\delta_{i,k} = z_{i,k}/3.6 - c_{i,k}$ denote the difference between the average scaled score and the AST score. Then the aberrant value is calculated using the following equation:

$$d_{i,k} = \begin{cases} 0 & \text{if } \delta_{i,k} < L; \\ (\delta_{i,k} - L)/10 & \text{if } L < \delta_{i,k} < L + 10; \\ 1 & \text{if } L + 10 < \delta_{i,k} < U - 10; \\ (U - \delta_{i,k})/10 & \text{if } U - 10 < \delta_{i,k} < U; \\ 0 & \text{if } \delta_{i,k} > U; \end{cases} \tag{2}$$

The values of L and U depend on the year and whether a student is classified as CLD or not. Values for 2007–2009 are shown below.

Year	Group	L	U
2009	CLD	−32	40
	non CLD	−57	40
2008	CLD	−32	44
	non CLD	−57	44
2007	CLD	−27	44
	non CLD	−58	44

2.6 Scaling scores

For each student (excluding international fee-paying students) seeking an ATAR, a scaling score is calculated as

$$v_{i,k} = \frac{z_{i,k} + 0.21 d_{i,k} c_{i,k}}{3.6 + 0.21 d_{i,k}}. \tag{3}$$

where $v_{i,k}$ is the scaling score, $z_{i,k}$ is the aggregate score, $d_{i,k}$ is the aberrant value, and $c_{i,k}$ is the AST score for student i at college k.

For each international fee-paying student, the scaling score does not include his/her AST score. So $y_{i,k} = z_{i,k}/3.6$. This is equivalent to setting $d_{i,k} = 0$.

(Prior to 1999, the AST weight was 1.0 rather than 0.21. Between 1999 and 2002, the AST weight was 0.80. It was changed to 0.21 in June 2002.)

2.7 ATARs

The Australian Tertiary Admission Ranks (ATARs) are calculated by ranking the aggregate scores. This rank is then converted to an age rank (including all members of the cohort). In this review, the process of converting aggregate scores to ATARs has not been considered. Because the ATARs are obtained using a monotonically increasing transformation of the aggregate scores, I will only consider the effect of the scaling algorithm on the aggregate scores obtained.

3 OCS scaling method

The ACT uses the Other Course Score (OCS) Scaling method developed by Daley (1989) to scale students' results from school-based assessment. At each college, T courses are placed in scaling groups for the purposes of scaling scores. A scaling group may consist of one or more courses with the approval of the BSSS Technical Adviser.

A scaling group with fewer than 10 T Package students is scaled using small group procedures rather than OCS scaling. A modified version of the OCS scaling algorithm is used for scaling groups with between 11 and 19 students.

Unfortunately, the OCS scaling method used by the ACT BSSS does not seem to be fully documented anywhere as it has changed over the years from the initial description given in Daley (1989, Chapter 11). The following description fills this gap as completely as possible given the information available to me.

3.1 Underlying one-factor model

The underlying model in most scaling algorithms, including that used in the ACT, is a one-factor model where the scaled scores of a student are given by

$$y_{i,j,k} = \mu_{i,k} + e_{i,j,k}, \tag{4}$$

where $\mu_{i,k}$ denotes the "ability" of the student and $e_{i,j,k}$ is an uncorrelated random error with mean zero. Thus the effect of course difficulty is removed because (4) assumes that each student will perform equally well in each course, subject only to different error terms (with zero means). Comparing (4) and (1), it is apparent that

$$a_{j,k} + b_{j,k}x_{i,j,k} = \mu_{i,k} + e_{i,j,k}. \tag{5}$$

The AST provides an "anchor" role in the scaling process. The AST score is treated like any other course and so it is assumed to satisfy the equation

$$A_k + B_k c_{i,k} = \mu_{i,k} + e^*_{i,k}, \tag{6}$$

where $e^*_{i,k}$ is an uncorrelated error term. Because the AST is a general aptitude test, $c_{i,k}$ measures the underlying ability of a student. Consequently, the values of A_k and B_k should be close to 0 and 1 respectively, for all colleges.

It follows from (4) and (6) and the weighted average at (3) that the scaling score $v_{i,k}$ and the aggregate score $z_{i,k}$ will satisfy similar one-factor models:

$$v_{i,k} = \mu_{i,k} + \varepsilon^*_{i,k} \tag{7}$$

$$\text{and} \qquad z_{i,k}/3.6 = \mu_{i,k} + \varepsilon_{i,k}, \tag{8}$$

where $\varepsilon_{i,k}$ and $\varepsilon^*_{i,k}$ are error terms with zero means.

3.2 Method of moments estimates

The method-of-moments approach is a way of estimating the values of $a_{j,k}$ and $b_{j,k}$ (Daley, 1995). First we rearrange the equations in Section 3.1.

Comparing (6) with (7) gives

$$A_k + B_k c_{i,k} = v_{i,k} + e^*_{i,k} - \varepsilon_{i,k}. \tag{9}$$

Taking the means of both sides of (9), we find

$$\text{ave}_i(v_{i,k}) = A_k + B_k \text{ave}_i(c_{i,k}), \tag{10}$$

where ave_i indicates the mean is taken over all values of i (i.e., over all students in college k).

Similarly, we can multiply both sides of (9) by $v_{i,k}$ and take expectations to obtain

$$\text{var}_i(v_{i,k}) = B_k \text{cov}_i(c_{i,k}, v_{i,k}). \tag{11}$$

We can solve these equations for A_k and B_k:

$$B_k = \frac{\text{var}_i(v_{i,k})}{\text{cov}_i(c_{i,k}, v_{i,k})} = \frac{\text{sd}_i(v_{i,k})}{\text{sd}_i(c_{i,k})\text{corr}_i(c_{i,k}, v_{i,k})} \tag{12}$$

$$A_k = \text{ave}_i(v_{i,k}) - B_k \text{ave}_i(c_{j,k}). \tag{13}$$

Similarly, combining (5) and (7) leads to

$$b_{j,k} = \frac{\text{sd}_i(v_{i,k})}{\text{sd}_i(x_{i,j,k})\text{corr}_i(x_{i,j,k}, v_{i,k})}$$

$$a_{j,k} = \text{ave}_i(v_{i,k}) - b_{j,k}\text{ave}_i(x_{i,j,k}).$$

Then, to speed convergence of the estimation procedure and to ensure $A_k \approx 0$ and $B_k \approx 1$, we modify these slightly to give:

$$b_{j,k} = \frac{\text{sd}_i(v_{i,k})}{B_k \text{sd}_i(x_{i,j,k})\text{corr}_i(x_{i,j,k}, v_{i,k})} \tag{14}$$

$$a_{j,k} = \left[\text{ave}_i(v_{i,k}) - A_k\right]/B_k - b_{j,k}\text{ave}_i(x_{i,j,k}). \tag{15}$$

3.3 Taper weights

Rather than take simple averages in the above equations, the scaling process involves taking weighted averages, standard deviations and correlations of students' results. These weights are known as "taper weights" and are defined as follows:

$$w_{i,j,k} = \begin{cases} 1 & \text{if } y_{i,j,k} \ge y^{(4)}_{i,k} \\ (y_{i,j,k} - y^{(4)}_{i,k} + \tau)/\tau & \text{if } 0 < y^{(4)}_{i,k} - y_{i,j,k} < \tau \\ 0 & \text{if } y^{(4)}_{i,k} - y_{i,j,k} \ge \tau \end{cases} \tag{16}$$

where $y_{i,k}^{(4)}$ is the fourth best scaled score of student i in college k. The value of τ is called the "taper" and it is normally set to 40, although it can be overwritten in the scaling process. Where a scaled score is the student's fifth, sixth, etc., best score, the student's results will only be included in the scaling process if the school course score falls within τ marks from the fourth best score.

The purpose of these weights is to ensure continuity in the construction of the aggregate scores as different course scores may be included in successive iterations.

3.4 Iterative process

To carry out the calculation of scaling scores (3), values of $a_{j,k}$ and $b_{j,k}$ are required in (1). We shall first need the following notation:

$$\begin{aligned}
m_{v,k} &= \text{weighted mean of the scaling scores in college } k;\\
s_{v,k} &= \text{weighted standard deviation of the scaling scores in college } k;\\
m_{c,k} &= \text{weighted mean of the AST scores in college } k;\\
s_{c,k} &= \text{weighted standard deviation of the AST scores in college } k;\\
r_{c,v,k} &= \text{weighted correlation between scaling scores and AST scores in college } k;\\
m_{x,j,k} &= \text{weighted mean of school scores for scaling group } j \text{ and college } k;\\
s_{x,j,k} &= \text{weighted standard deviation of school scores for scaling group } j \text{ and college } k;\\
m_{v,j,k} &= \text{weighted mean of scaling scores for scaling group } j \text{ and college } k;\\
s_{v,j,k} &= \text{weighted standard deviation of scaling scores for scaling group } j \text{ and college } k;\\
r_{x,v,j,k} &= \text{weighted correlation between school scores and scaling scores for scaling group } j\\
&\ \text{and college } k.
\end{aligned}$$

The weights used in computing the college-level quantities (the first five quantities) are the Aberrant values ($d_{i,k}$), and the weights for all other calculations (those with a j subscript) are the taper weights ($w_{i,j,k}$) defined in equation (16).

The equations (12)–(15) are solved using an iterative process for each school. In each iteration, the school course scores are scaled, scaling group by scaling group, and then a new scaling score is calculated for each student. The iterative procedure stops when the scaling scores converge or when 30 iterations are completed. For some groups, the scaling scores do not converge and then the Technical Adviser intervenes.

The process begins by setting $a_{j,k} = 25/6$ and $b_{j,k} = 25/12$ for all j and k. Because the course scores (x) have a mean of 70 and standard deviation of 12, these values result in the scaled scores (y) having a mean of 150 and a standard deviation of 25. Once we have initial values of $a_{j,k}$ and $b_{j,k}$, we can calculate scaled scores from (1), and thence scaling scores from (3).

For each college, the estimates of $a_{j,k}$ and $b_{j,k}$ are then improved as follows, based on equations (12)–(15).

Step 1. First compute the values of A_k and B_k for the AST scores in college k:

$$B_k = s_{v,k}/(r_{c,v,k}s_{c,k}) \tag{17}$$

$$A_k = m_{v,k} - B_k m_{c,k}\,. \tag{18}$$

Step 2. For each scaling group j in college k, let $n_{j,k}$ be the number of students in the scaling group with taper weight $w_{i,j,k} > 0.5$, and calculate

$$b^*_{j,k} = s_{v,j,k}/(r^*_{x,v,j,k} s_{x,j,k} B_k)\,, \tag{19}$$

where $r^*_{x,v,j,k} = \max(0.5, r_{x,v,j,k})$.

If $n_{j,k} > 20$, set $b_{j,k} = b^*_{j,k}$.

If $11 \le n_{j,k} \le 20$, set

$$b_{j,k} = 25/12 + (b^*_{j,k} - 25/12)(n_j - 10)/10.$$

(This last equation is known as "modified OCS".)

Step 3. For each scaling group j in college k, calculate

$$a_{j,k} = (m_{v,j,k} - A_k)/B_k - m_{x,j,k} b_{j,k}\,. \tag{20}$$

Step 4. Scaled scores are recomputed from (1) for all students and all scaling groups in the college.

Step 5. Scaling scores are recomputed from (3) for all students in the college.

These six steps are iterated for each college until the values of A_k and B_k converge. Consequently, the values of $a_{j,k}$ and $b_{j,k}$ will also converge for all scaling groups in the college.

4 Data

The ACT BSSS provided individual students' data including AST results, school-based course scores, scaled scores, scaling group identifications, college identifications and final aggregate scores, for each of 2007, 2008 and 2009. Summary statistics for some of the variables are shown in Table 1. Table 2 shows the number of students from each year of the data provided, and Table 3 shows the number of CLD students in each year. The data included all ACT students from 2007–2009. There were some additional overseas students in each year who were not included in the data provided.

Table 1: *Summary statistics for the key score information across all colleges in 2007–2009.*

	Aberrant	School	Scaled	Aggregate	AST	Scaling
Min.	0.0000	28.49	79.54	318.4	46.77	88.42
1st Quartile	1.0000	63.12	132.26	503.2	129.33	139.60
Median	1.0000	71.04	147.57	554.6	149.93	153.82
Mean	0.9198	71.16	147.93	555.3	148.97	154.05
3rd Quartile	1.0000	79.29	164.08	608.4	169.38	168.90
Max.	1.0000	110.32	219.59	778.9	230.16	216.34

	2007	2008	2009	Total
BASS	0	32	59	91
CBRC	197	210	228	635
CGGS	137	138	143	418
CITC	30	43	41	114
COPC	113	84	82	279
DARC	144	124	166	434
DCKC	132	144	168	444
EDMC	53	62	71	186
ERNC	134	153	154	441
HWKC	186	188	161	535
LGNC	105	126	93	324
MARC	131	144	136	411
MERC	60	79	76	215
MKCC	79	108	94	281
NARC	322	348	344	1014
ORAC	10	9	12	31
RDFC	154	157	165	476
SFXC	85	108	108	301
STCC	116	115	147	378
TRCC	42	49	60	151
TUGC	152	136	187	475
Total	2382	2557	2695	7634

Table 2: *Number of students from each college in each year.*

	2007	2008	2009	Total
BASS	0	3	1	4
CBRC	22	20	28	70
CGGS	11	10	6	27
CITC	1	13	3	17
COPC	21	18	11	50
DARC	1	1	0	2
DCKC	22	17	7	46
EDMC	2	1	1	4
ERNC	13	17	2	32
HWKC	30	35	5	70
LGNC	8	22	11	41
MARC	11	4	0	15
MERC	4	1	1	6
MKCC	7	7	0	14
NARC	46	26	50	122
ORAC	0	0	0	0
RDFC	1	2	0	3
SFXC	8	3	0	11
STCC	9	6	1	16
TRCC	3	0	0	3
TUGC	32	26	19	77
Total	252	232	146	630

Table 3: *Number of CLD students at each college in each year.*

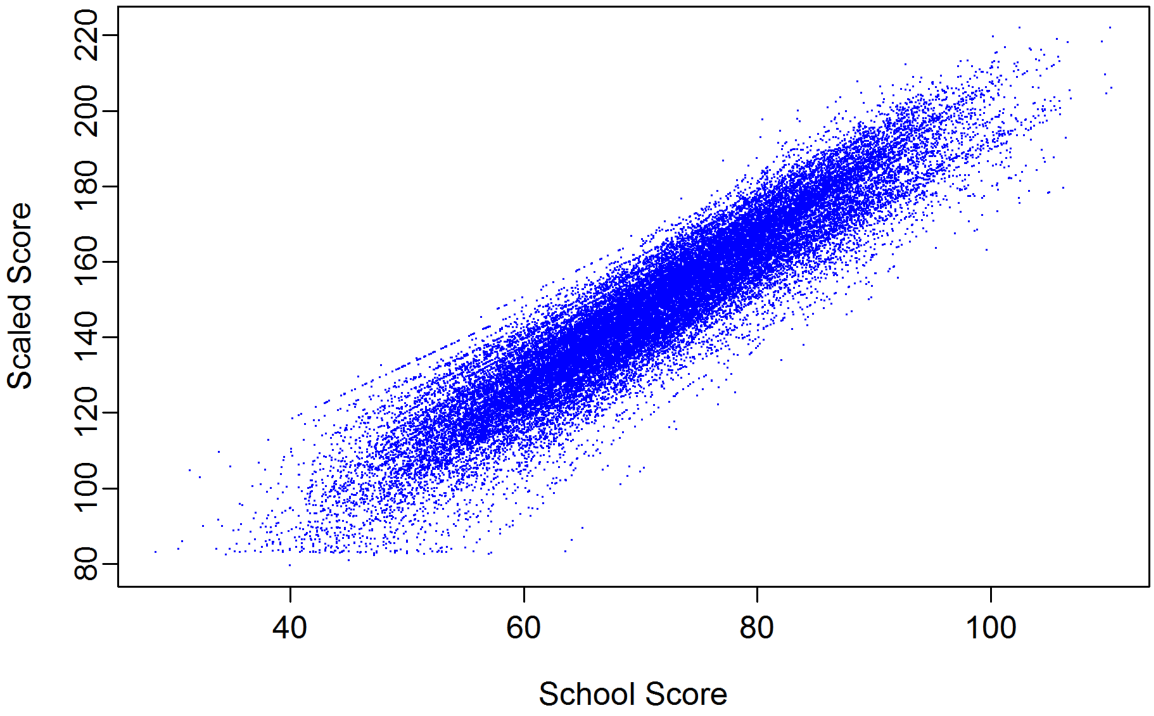

Figure 1: *School scores and scaled scores for all students in all years.*

Figure 1 shows the individual scaled scores and school scores for each student and Figure 2 shows plots of the scaling scores, aggregate scores and AST scores for all students across all years. The results for each year are very similar, and so the years have not been separated in these figures. While the correlation between school scores and AST scores is not particularly high, once the data are scaled, the scaled results are averaged, and the scaling score computed, the correlation has increased to nearly 0.70 which is comparable to the results in other states (see Section 7).

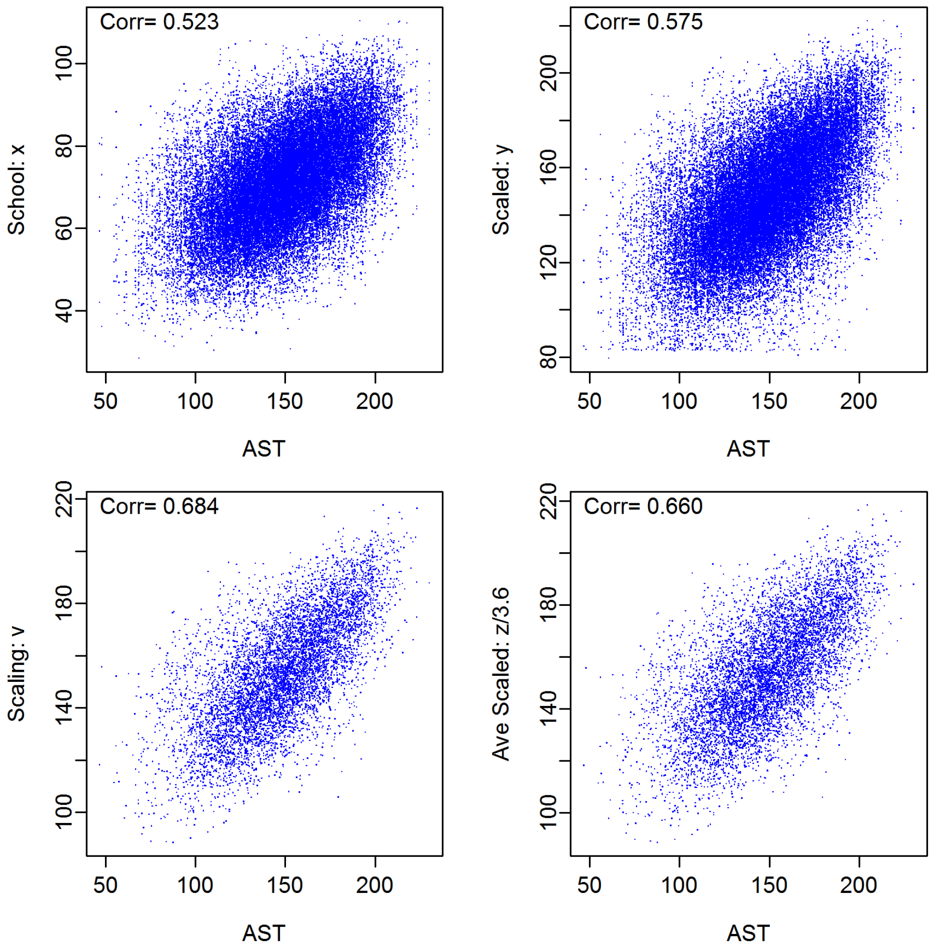

Figure 2: *School scores, Scaling scores and Scaled scores plotted against AST scores for all students in all years.*

5 Validation of scaling results

Extensive testing was carried out to ensure that the scaling algorithm implemented for this review was equivalent to the algorithm implemented by the ACT BSSS. Any differences were explained after discussion with ACT BSSS staff.

I calculated the Scaling scores using (3) based on the Scaled scores, AST scores and Aberrant values provided by the ACT BSSS, and compared the results with the Scaling scores provided by the ACT BSSS. The results are shown in Figure 3. The scores match (within 0.1) for all but 34 students (30 from 2007 and two each from 2008 and 2009). These students were either not in the original scaling or had course scores adjusted as the result of H courses.

I checked the calculation of Aberrant scores based on the AST values and Scaled scores provided by the ACT BSSS. Figure 4 shows the scores obtained from (2) plotted against the Aberrant scores provided by the ACT BSSS for 2007–2009. Some randomness has been added to enable the points to be seen distinctly. Nearly 97% of the Aberrant scores match with 1.2% of scores differing by more than 0.1 and 0.14% of scores differing by more than 0.25. It is assumed that most of the differences are due to the Aberrant scores provided by the ACT BSSS being calculated in an earlier iteration of the algorithm than those obtained using the final Scaled scores.

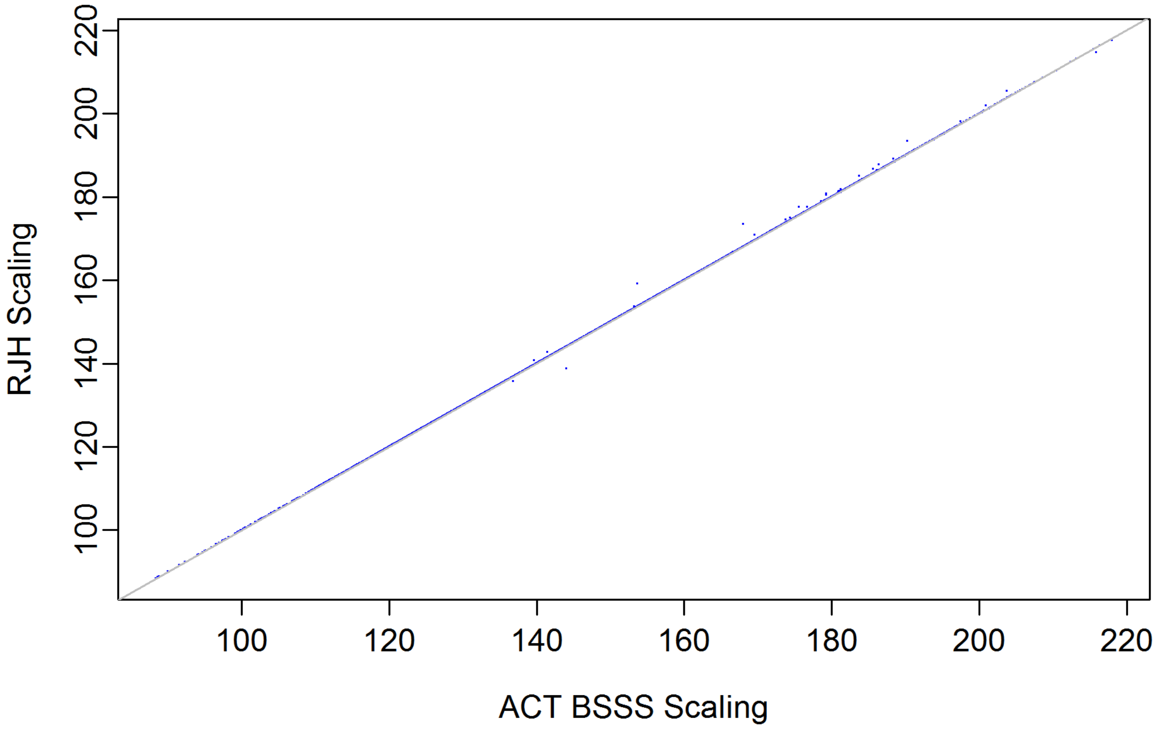

Figure 3: *Scaling scores computed from the formula (labelled "RJH Scaling") plotted against Scaling scores provided by the ACT BSSS.*

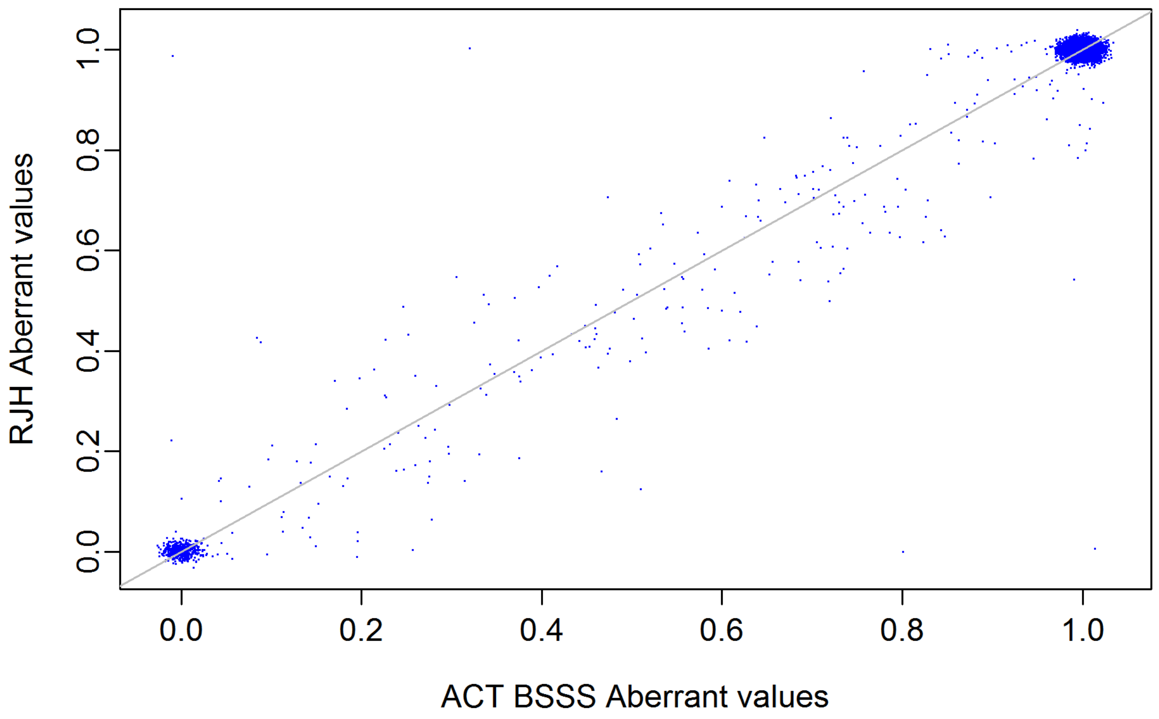

Figure 4: *Aberrant scores computed from* (2) *(labelled "RJH Aberrant values") plotted against Aberrant scores provided by the ACT BSSS. A small amount of random noise has been added to each point to prevent over-plotting, especially at (0,0) and (1,1).*

It is also necessary to check the computation of scaled scores via the iterative procedure. Figures 5–7 show the $b_{j,k}$ obtained via my implementation of the iterative algorithm against the $b_{j,k}$ values obtained by the ACT BSSS for each school in each year. So each point on these graphs represents one scaling group. Scaling groups with fewer than 11 students have not been included in these plots.

The values of $b_{j,k}$ obtained via this algorithm ranged between 0.87 and 3.86, and the values of $a_{j,k}$ ranged from -155.30 to 81.25. For almost all scaling groups, the $b_{j,k}$ values I obtained are very close to those obtained by the ACT BSSS. Where they differ, it is likely that this is due to slight variations in the populations of the scaling groups at the time the scaling was undertaken.

Figure 8 show the scaling scores obtained via my implementation of the iterative algorithm (using the $b_{j,k}$ values shown in Figures 5–7) against the scaling scores provided by the ACT BSSS for all colleges in 2007–2009. Small groups (fewer than 11 students) and other groups that were not scaled by the ACT BSSS using the (modified) OCS algorithm were not re-scaled here. All scaled scores from groups that were not scaled in this exercise were set to the values specified by the ACT BSSS. This figure differs from Figure 3 because I've recalculated the Scaled scores in Figure 8 but I've used the Scaled scores provided by the ACT BSSS in Figure 3.

In most cases, the results are very close. The discrepancies are probably explained by the slightly different populations of students used in this review and by the ACT BSSS in their annual scaling exercise.

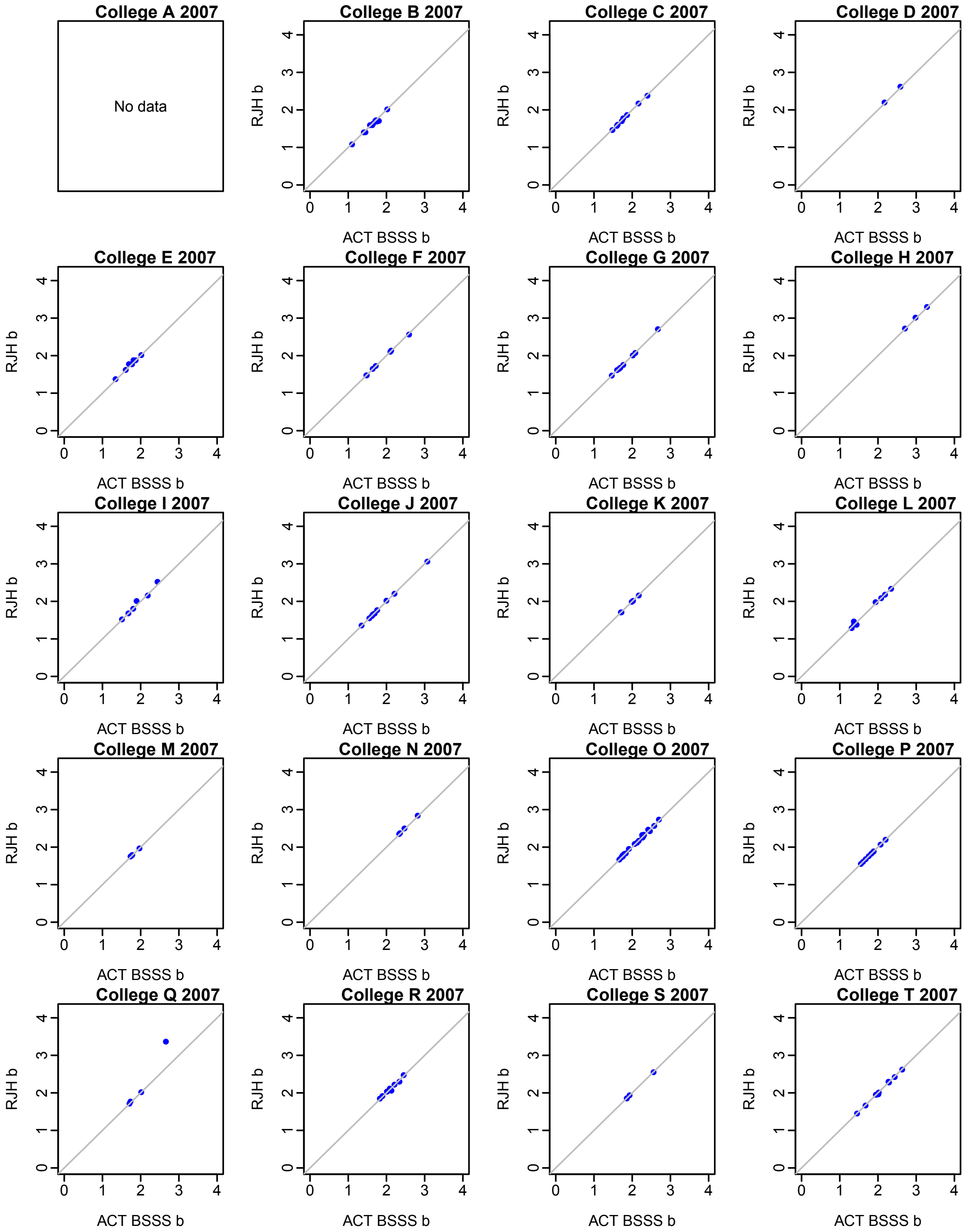

Figure 5: *Values of $b_{j,k}$ computed via the iterative algorithm against the $b_{j,k}$ values obtained by the ACT BSSS for each school in 2007.*

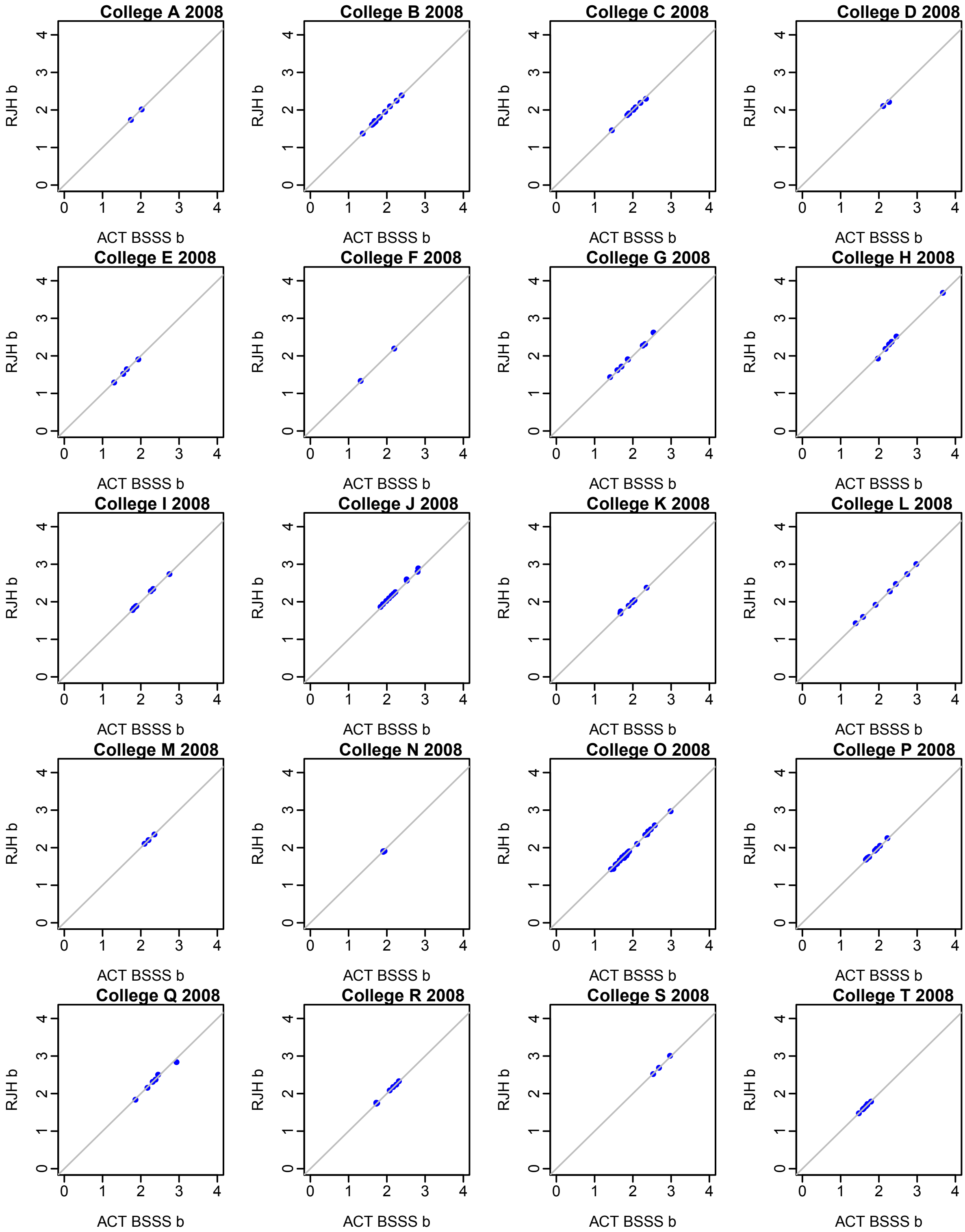

Figure 6: *Values of $b_{j,k}$ computed via the iterative algorithm against the $b_{j,k}$ values obtained by the ACT BSSS for each school in 2008.*

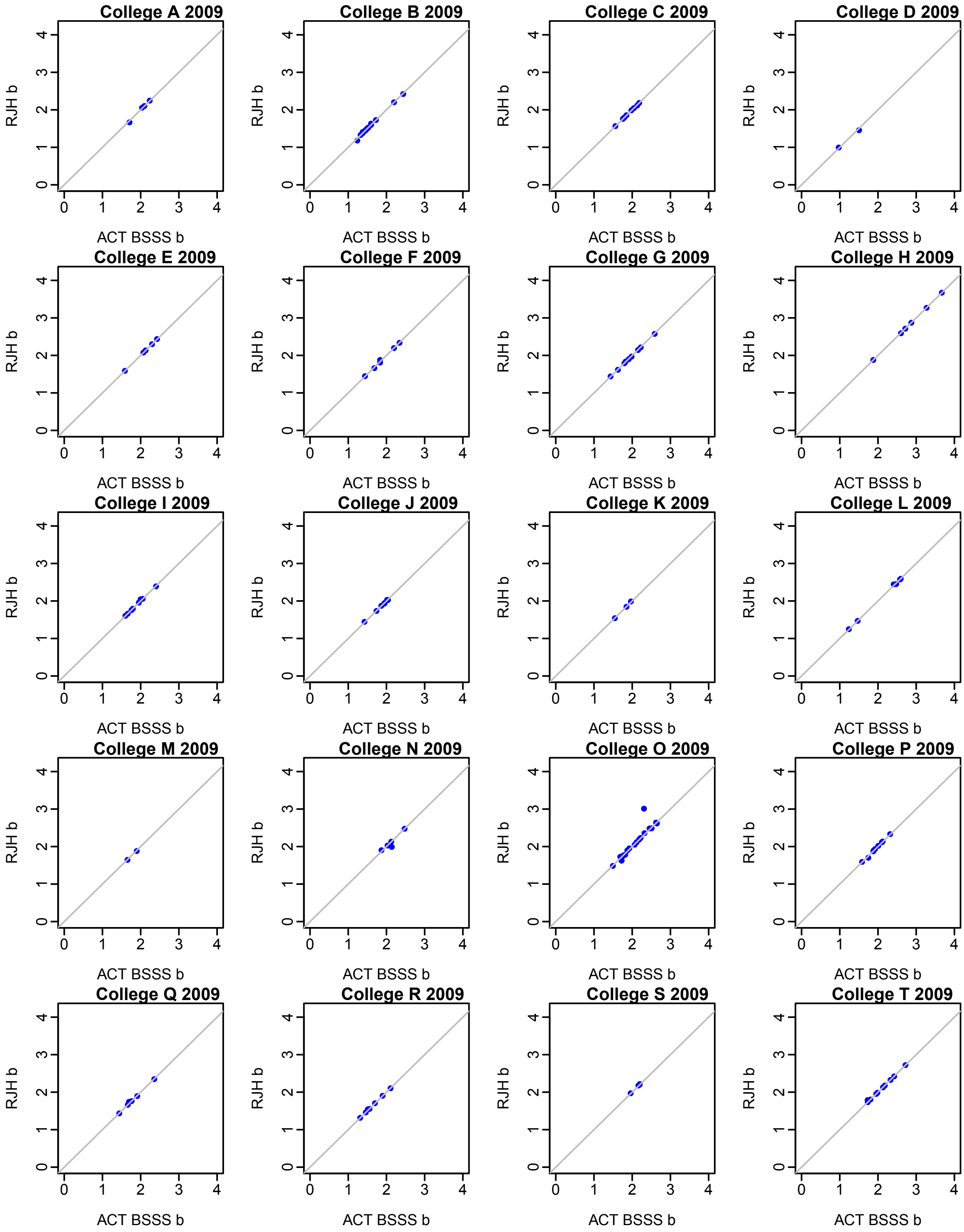

Figure 7: *Values of $b_{j,k}$ computed via the iterative algorithm against the $b_{j,k}$ values obtained by the ACT BSSS for each school in 2009.*

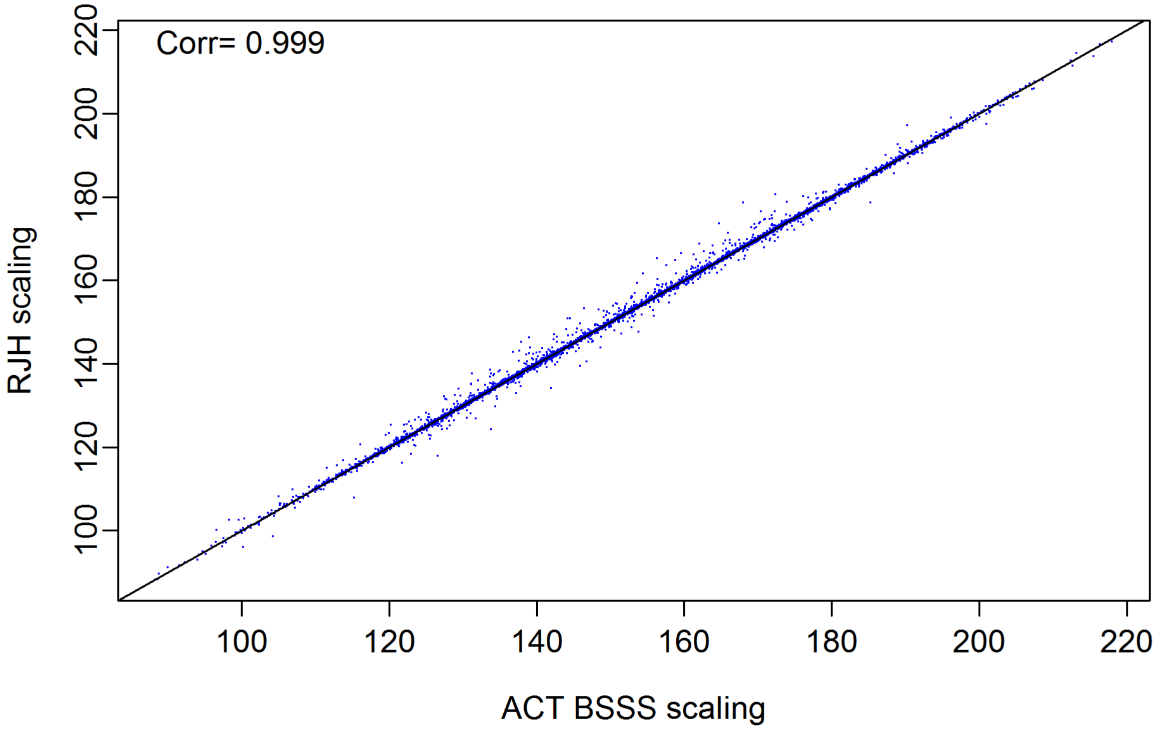

Figure 8: *Scaling scores computed via the iterative algorithm plotted against Scaling scores provided by the ACT BSSS. All colleges and all years combined.*

We will need the values of $a_{j,k}$ and $b_{j,k}$ in Section 6. Histograms of the estimates of these quantities are shown in Figure 9 for those groups with at least 20 students.

We will also need the standard deviation of the scaled scores within each scaling group. These values, denoted by $s_{j,k}$, are shown in Figure 10 where the group size was at least 20 students.

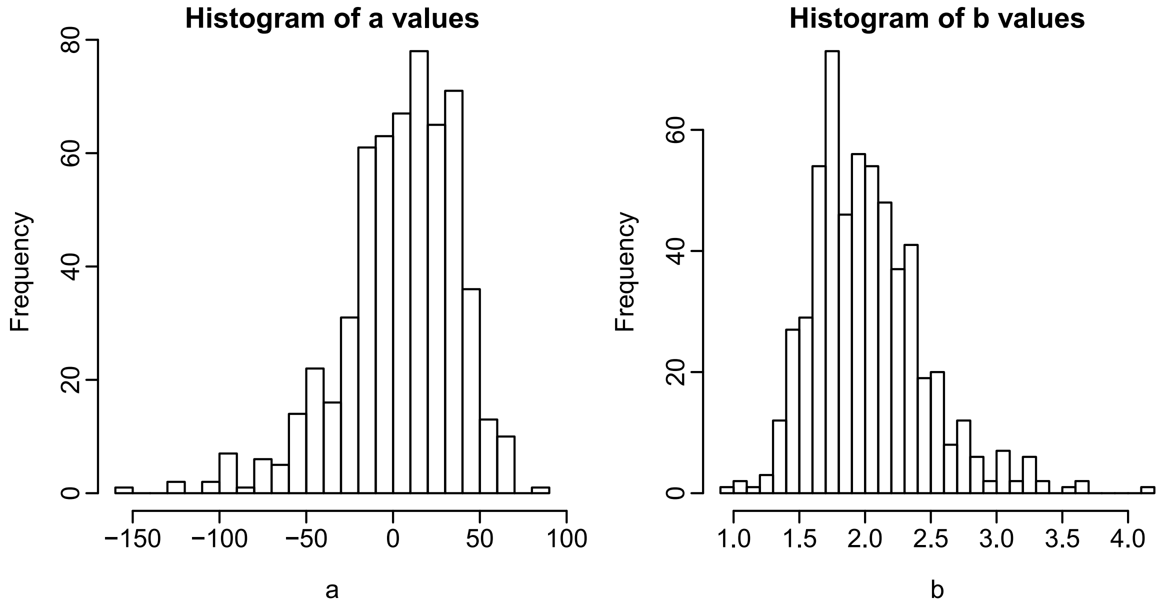

Figure 9: *Estimated values of $a_{j,k}$ and $b_{j,k}$ for all scaling groups with at least 20 students in all colleges and all years.*

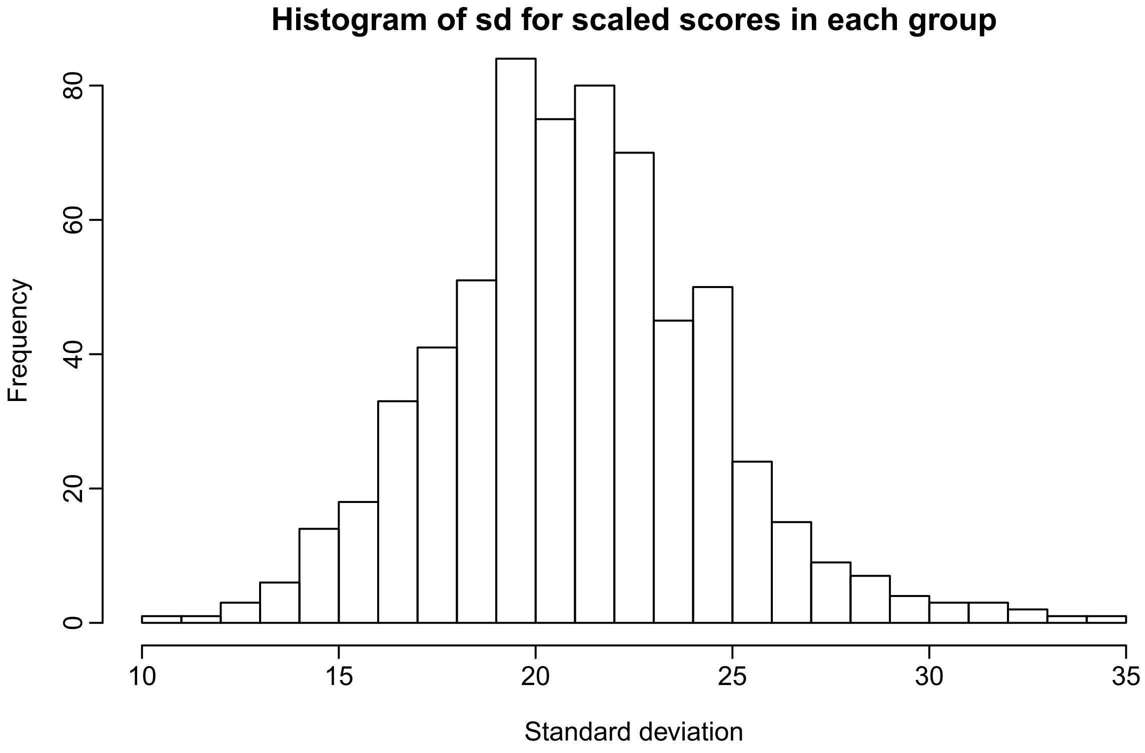

Figure 10: *Standard deviations of the scaled scores within each scaling group for groups with at least 20 students.*

6 A simulation approach

In order to measure the effectiveness of a scaling algorithm, we will use the Mean Absolute Error based on equation (7):

$$\text{MAE} = \text{ave}_i |z_{i,k}/3.6 - \mu_{i,k}|.$$

The ATAR is a non-linear monotonically increasing function of $z_{i,k}$; therefore the MAE is a measure of how well the ATAR reflects the underlying student ability, $\mu_{i,k}$. The smaller the value of the MAE, the better the algorithm. However, since we do not know μ_i this is difficult to estimate from a set of real data. A convenient approach is to simulate data similar to the real data, but where μ_i values are known.

As each college is scaled separately, the simulations only need to involve one "college". This imaginary college contains 250 students each doing 4 subjects (3 majors and 1 minor). There are 16 scaling groups with approximate group sizes $\{20, 22, 24, 26, 28, 30, 35, 40, 50, 60, 70, 80, 90, 100, 125, 200\}$, matching the typical distribution of scaling group sizes in the real data and ignoring all small or medium-size groups.

Because there is only one college, we drop the k subscript that was used in earlier equations. We simulate data based on (1), (4) and (6) as follows:

$$\begin{aligned}
(a_j^*, b_j^*) &\text{ randomly sampled from } (a_{j,k}, b_{j,k}) \text{ estimates obtained from 2007–2009 data;}\\
\sigma_j &\text{ randomly sampled from } s_{j,k} \text{ obtained from 2007–2009 data;}\\
\mu_i &\sim \text{N}(150,\ \sigma_\mu^2)\\
y_{i,j}^* &= \mu_i + e_{i,j}\\
x_{i,j} &= A_x + B_x(y_{i,j}^* - a_j^*)/b_j^*\\
c_i &= A_c + B_c(\mu_i - \Delta_i + e_i^*)
\end{aligned}$$

where 8% of students are randomly classified as CLD, $\Delta_i = 30$ if student i is classified as CLD and 0 otherwise, $e_{i,j} \sim \text{N}(0, \sigma_j^2)$ and $e_i^* \sim \text{N}(0, \sigma_c^2)$. Half of the CLD students are randomly selected to be fee paying. The values of A_c and B_c are chosen to ensure the c_i values have mean 150 and standard deviation 27.5, and the values of A_x and B_x are chosen to ensure the values of $\{x_{i,j}\}$ have a mean of 70 and standard deviation of 12.

Once c_i and $x_{i,j}$ have been simulated using these equations, the scaling algorithm described in Section 3.4 is used to obtain the scaled scores $y_{i,j}$ and scaling scores, v_i. In computing the aberrant values from (2), 2009 values of L and U are used. Note that the scaled scores, $y_{i,j}$, will not be the same as the $y_{i,j}^*$ scores used in generating the school scores; similarly, the values of a_j and b_j will not be the same as the values of a_j^* and b_j^*. The only information directly used from the simulation are the AST scores (c_i) and the school scores ($x_{i,j}$) which are used as inputs to the scaling algorithm. The values of μ_i will also be used in assessing the effectiveness of the scaling algorithm.

To find appropriate values for σ_μ and σ_c, I calculated

$$r_{z,c} = \text{corr}_{i,k}(c_{i,k}, z_{i,k}) = 0.660, \quad r_{x,c} = \text{corr}_{i,j,k}(c_{i,k}, x_{i,j,k}) = 0.523, \quad r_{y,c} = \text{corr}_{i,j,k}(c_{i,k}, y_{i,j,k}) = 0.575,$$

$$s_{v,c} = \text{ave}_{i,j,k}(v_{i,k} - c_{i,k})^2 = 470, \quad \text{and} \quad s_{v,y} = \text{ave}_{i,j,k}(v_{i,k} - y_{i,j,k})^2 = 205,$$

from the real data where each quantity is computed using all students in all years and all colleges. See Figure 2 for a graphical display of the correlations. Then σ_μ and σ_c were selected to minimize

$$(r_{z,c} - 0.660)^2 + (r_{x,c} - 0.523)^2 + (r_{y,c} - 0.575)^2 + (s_{v,c}/470 - 1)^2/5 + (s_{v,y}/205 - 1)^2/5,$$

where each of $r_{z,c}$, $r_{x,c}$, $r_{y,c}$, $s_{v,c}$ and $s_{v,y}$ were computed from the simulated data. This yielded values of $\sigma_\mu = \sigma_c = 25$. Then, using these values of σ_μ and σ_c, a realistic simulated set of data can be generated for the 250 students in the imaginary college.

For the specific values of $\sigma_\mu = \sigma_c = 25$, we obtain MAE=8.86. That is, the average scaled score, $z_{i,k}/3.6$, is on average about 8.86 points different from the true underlying ability, μ_i. Figure 11 shows some simulated data obtained using the above equations. This figure can be compared with similar real data shown in Figure 2 on page 14.

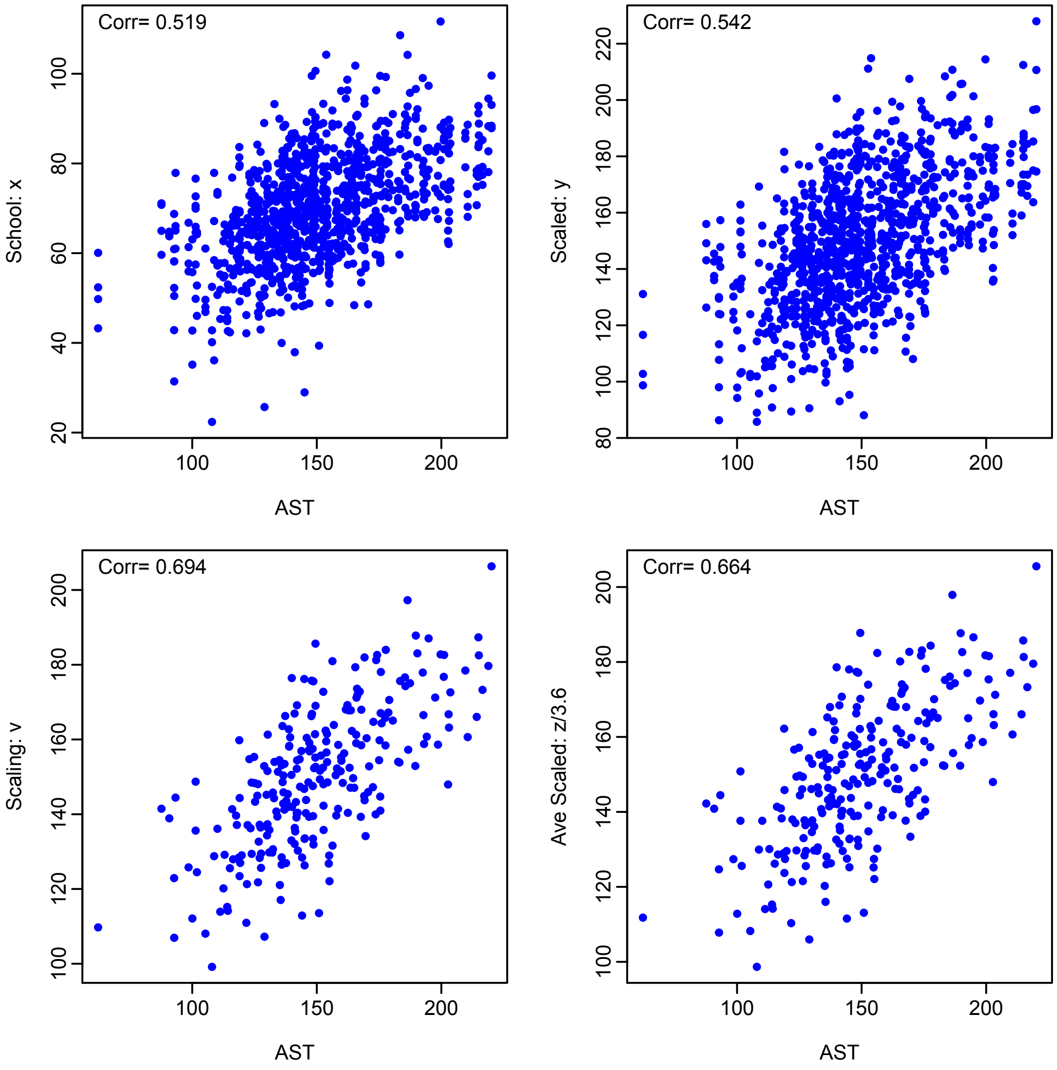

Figure 11: *Simulated data for AST (c), School scores (x), Scaled scores (y), Scaling scores (v) and Aggregate scores (z).*

7 Correlations between school scores and AST scores

One of the issues raised in the terms of reference for this review was the problem of low correlations between school-based course scores and AST scores. Figure 12 shows the correlations between these two scores for each scaling group containing at least 50 students. The average correlation is 0.55 with a minimum of −0.03 and a maximum of 0.83. It was thought that the very low correlations and negative correlations may be associated with scaling groups containing large numbers of CLD students. To test this hypothesis, Figure 13 shows the correlations plotted against the percentage of CLD students in each scaling group, demonstrating that there is no such relationship.

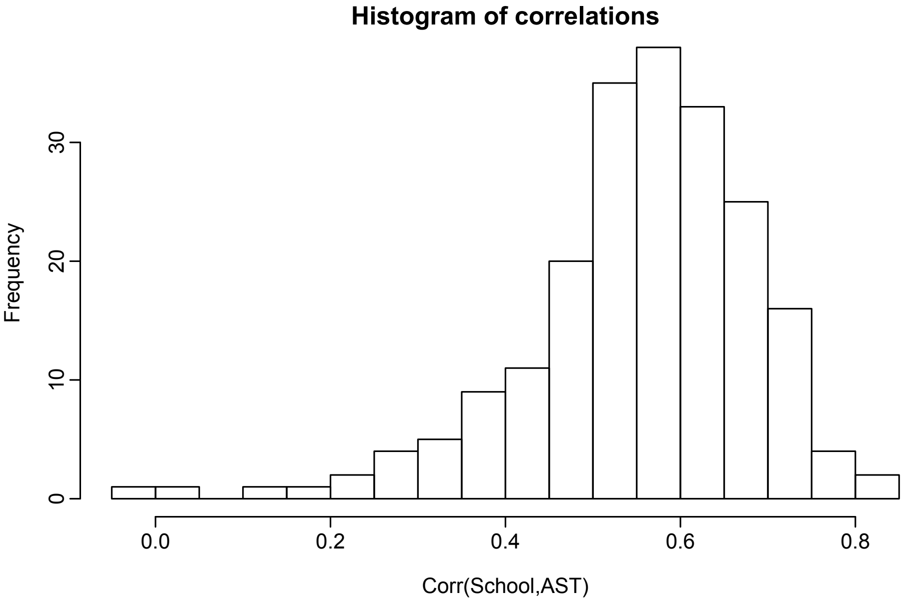

Figure 12: *Correlations of School scores and AST scores for each scaling group containing at least 50 students.*

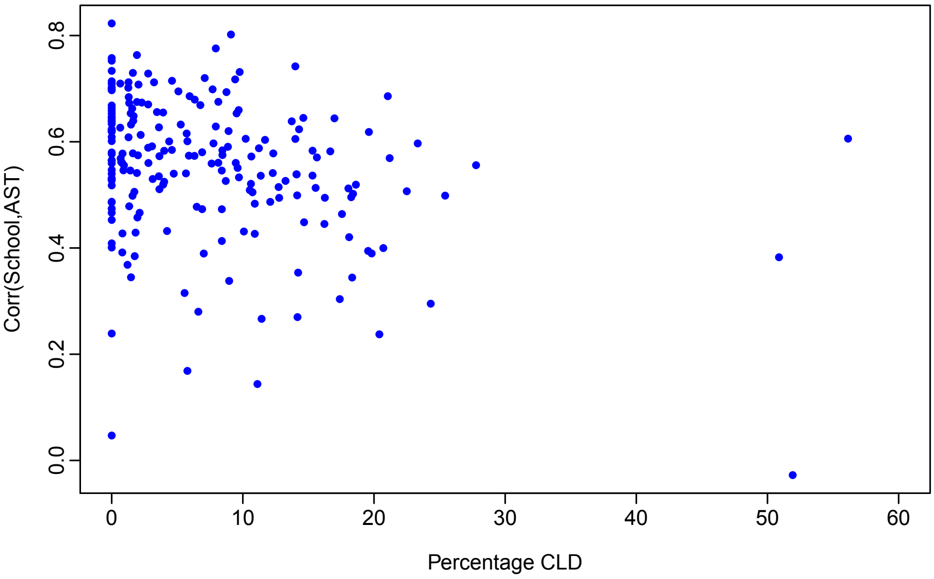

Figure 13: *Correlations of School scores and AST scores plotted against the percentage of CLD students for each scaling group containing at least 50 students.*

It is clear from the algorithm described in Section 3.4 that the correlations between school-based course scores and AST scores have no direct role in the scaling algorithm implemented in the ACT. The only correlations that have a direct impact on scaled scores and scaling scores are (1) the correlation between scaling scores and AST scores and (2) the correlation between school scores and scaling scores. In this section I explore the impact of AST scores on these two correlations and on the resulting scaling scores.

Table 4 shows the correlations of AST scores with Scaling scores for each college in each year. The mean correlation is 0.75 with only CITC in 2009 having a correlation lower than the minimum of 0.50. Note that these are not correlations of independent observations because the scaling scores include AST scores.

	BASS	CBRC	CGGS	CITC	COPC	DARC	DCKC	EDMC	ERNC	HWKC	LGNC
2007		0.73	0.73	0.87	0.75	0.70	0.74	0.66	0.70	0.66	0.74
2008	0.72	0.80	0.85	0.74	0.67	0.75	0.77	0.63	0.73	0.76	0.77
2009	0.79	0.72	0.81	0.47	0.77	0.79	0.72	0.69	0.69	0.75	0.74

	MARC	MERC	MKCC	NARC	ORAC	RDFC	SFXC	STCC	TRCC	TUGC
2007	0.69	0.79	0.80	0.75	0.58	0.81	0.81	0.77	0.71	0.76
2008	0.76	0.74	0.71	0.75	0.84	0.81	0.81	0.76	0.87	0.75
2009	0.73	0.79	0.79	0.75	0.69	0.82	0.76	0.71	0.78	0.77

Table 4: *Correlations of AST scores with scaling scores for each college in each year.*

Table 5 shows the average correlations (across scaling groups) of school scores and scaling scores for each college in each year. The mean correlation is 0.87 and all correlations were greater than 0.46. Only two scaling groups had correlations lower than the minimum of 0.50 (one from COPC in 2007 and the other from MERC in 2009). Again, these are not correlations of independent observations because the scaling scores include school course scores.

	BASS	CBRC	CGGS	CITC	COPC	DARC	DCKC	EDMC	ERNC	HWKC	LGNC
2007		0.88	0.83	0.88	0.85	0.82	0.88	0.88	0.85	0.87	0.85
2008	0.89	0.89	0.87	0.89	0.87	0.84	0.87	0.87	0.88	0.88	0.85
2009	0.86	0.88	0.87	0.77	0.88	0.83	0.87	0.88	0.86	0.86	0.86

	MARC	MERC	MKCC	NARC	ORAC	RDFC	SFXC	STCC	TRCC	TUGC
2007	0.85	0.90	0.87	0.87	0.90	0.91	0.90	0.91	0.89	0.87
2008	0.86	0.87	0.90	0.85	0.82	0.90	0.90	0.88	0.91	0.85
2009	0.88	0.86	0.89	0.84	0.88	0.88	0.91	0.87	0.91	0.89

Table 5: *Average correlations of school course scores with scaling scores for each college in each year.*

(a)

There is almost no information available from other states about the relationship between scaling scores and general aptitude tests such as the AST. Further, individual scores on general aptitude tests are not used directly in scaling in most other jurisdictions, and so students may perceive them differently. In Victoria, the GAT (General Achievement Test) is used to monitor the assessment process, and as an additional selection score for a small number of courses, but is not used in the calculation of a student's ATAR. The correlation between the ATAR predicted by the GAT (using a nonlinear regression) and the ATAR itself is about 0.72 (Hyndman, 2010). Because the ATAR is a nonlinear function of the aggregate score, this is also an upperbound on the correlation between the aggregate score and the GAT. The actual correlation is probably lower due to the assumed linearity in a correlation calculation. Consequently, the correlation between the AST and Scaling scores in the ACT is almost certainly larger than the correlation between the GAT and aggregate scores in Victoria.

Note that the correlations between internal and external assessments for particular subjects in other states is not a valid comparison as the external assessments are subject specific and not intended to be general aptitude tests such as the AST. Further, it is expected that correlations between assessments for a specific subject will be higher than correlations between general aptitude tests and subject scores because a more restricted and homogeneous body of material is being assessed.

(b)

The scaling process in every state and territory uses the available information as effectively as possible in constructing an appropriate estimate of μ_i. Scaling is possible provided the data have *positive* correlations, and a sufficiently large population to ensure significant correlations. This is clearly true in the ACT and everywhere else in Australia.

If the AST scores had very low (or even negative) correlations with scaling scores, that would suggest that the AST was not a good measure of general aptitude. Then it would be necessary to reform the AST to ensure it was a better measure of general aptitude. There are no statistical procedures that can cover up a poor test. However, the evidence suggests that the AST is a relatively good measure of general aptitude (based on the comparison with Victoria's GAT).

8 The effect of low correlations on ATARs

(c)

One of the issues to address is the effect on ATARs of courses where there are low correlations between school-based course scores and the AST. In order to study this effect, I simulated some data while controlling the correlation between school scores and the AST for one particular scaling group of size 50. This can be done by specifying the standard deviation, σ_j, for the group. Large values of σ_j correspond to small correlations between the School scores and AST scores. The relationship between σ_j and correlation between the School scores and AST scores is shown in Figure 14. This plot shows results for 2000 simulations of the specified scaling group of size 50 with different values of σ_j; the mean correlation for each value of σ_j shown as a red line. Note that the induced correlations are similar to those seen in real data (see Figure 12 on page 24).

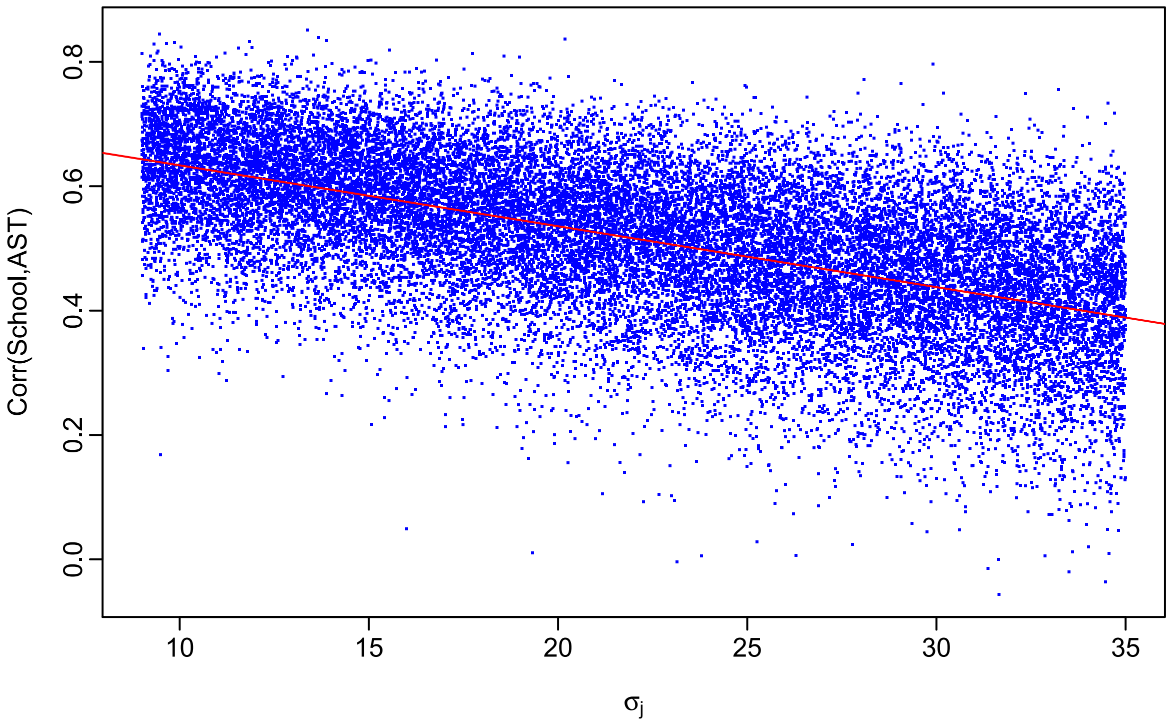

Figure 14: *Effect of σ_j on the correlation between School scores and AST scores.*

I allowed σ_j to increase from 10 (just lower than the smallest value of σ_j in the real data) to 35 (just higher than the highest value of σ_j seen in real data). For each value of σ_j, 2000 simulations were carried out and the mean and standard deviation of the Aggregate scores for the students in the particular group were computed. The average of the means and standard deviations are shown in Figure 15. I also computed the value of the MAE for the particular scaling group.

Figure 15 shows that the average aggregate score is unaffected by the changing σ_j, but that the standard deviation decreases as the correlation between school scores and AST scores decreases. However, the effect is very small within the range of values that are seen with real data. Figure 16 shows that the MAE increases as σ_j increases (and the correlation between school scores and AST scores decreases). Again, the effect is relatively small with a range of 1.4 for realistic values of σ_j.

(f)

The main thing to be learned from this exercise is that the effect of one scaling group on the Scaling scores is small, even when the correlations are low. In other words, the OCS method-of-moments procedure appears to be a relatively robust estimator of the value of $\mu_{i,k}$ over a wide range of error variances (corresponding to a range of correlations).

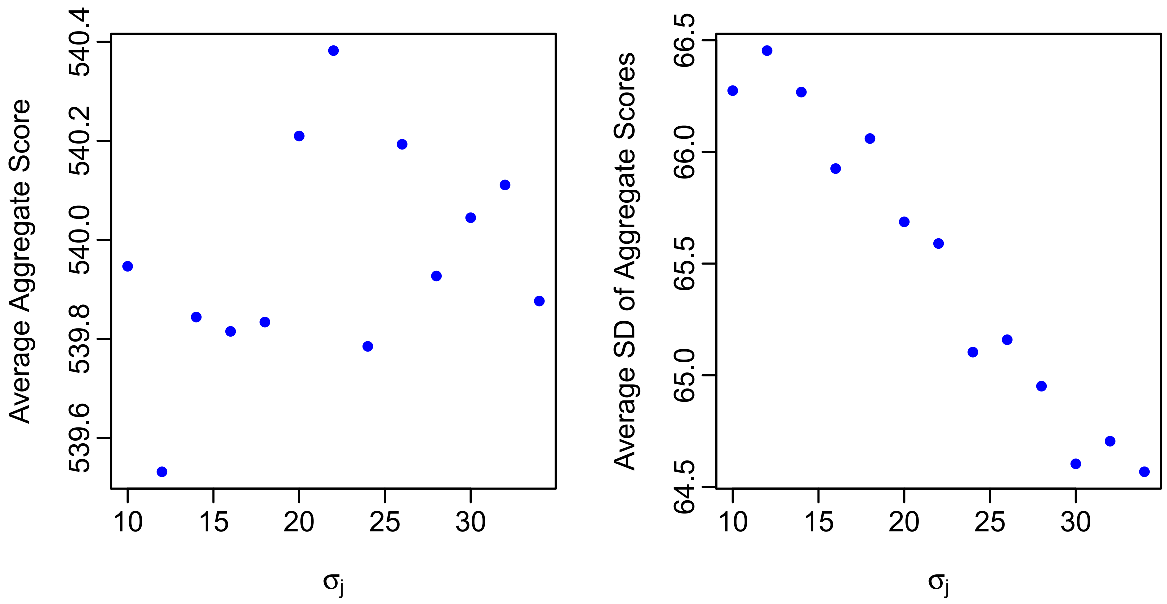

Figure 15: *Effect on scaling scores of a scaling group having low correlation with AST scores.*

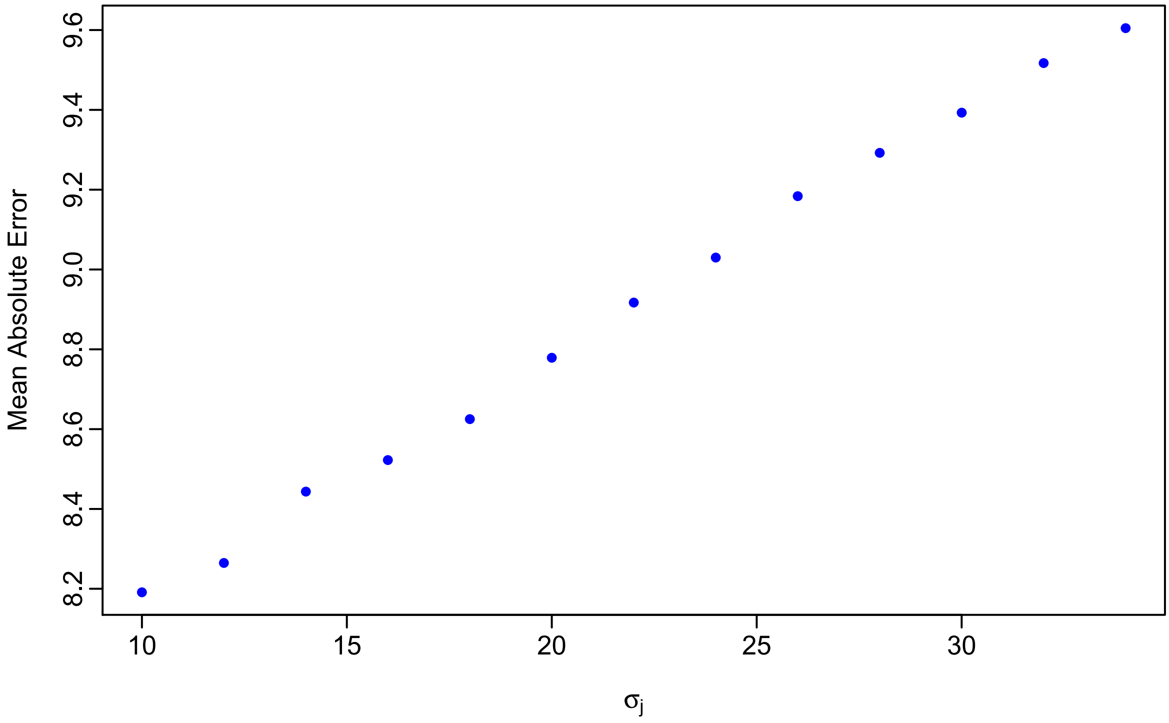

Figure 16: *Effect on MAE of a scaling group having low correlation with AST scores.*

9 Effectiveness of existing measures to improve correlations

(d)

Analyses in Daley (1989) showed that the single most important action to be taken by the ACT BSSS to improve correlations was to abandon the pure AST scaling procedure and use an Other Course Scaling procedure. Since the OCS procedure was adopted, the ACT BSSS has introduced several measures intended to improve the correlations or mitigate against low correlations:

1. introduction of a short response test to the AST;
2. the inclusion of AST results directly in the scaling score;
3. implementation of an aberrant score process.

The short-response test was introduced to also attempt to alleviate the gender-linked discrepancy between the reference scale scores and the school-based assessments (Daley, 1989, Chapter 7).

I will consider the effect of items 2 and 3.

To measure the effectiveness of including AST results, I recomputed the scaling scores for 2009 with all AST results omitted. (Of course, this could not actually be done in practice, because then there would be no course to anchor the scaling process.) That is, the aberrant value, $d_{i,k}$, was set to zero in (3) for all students. These revised scaling scores are plotted against the original scaling scores in Figure 17 (left).

As a further test, 2000 sets of simulated scores were generated with d_i set to zero for all students. The average value of MAE for these revised simulated scores is 9.3 compared to 8.9 for the original simulated scores. So the inclusion of AST results directly in the scaling score has reduced the mean absolute error by 4%.

To measure the effectiveness of the aberrant score process, I recomputed the scaling scores for 2009 with the aberrant value, $d_{j,k}$, set to one in (3) for all students. These revised scaling scores are plotted against the original scaling scores in Figure 17 (right).

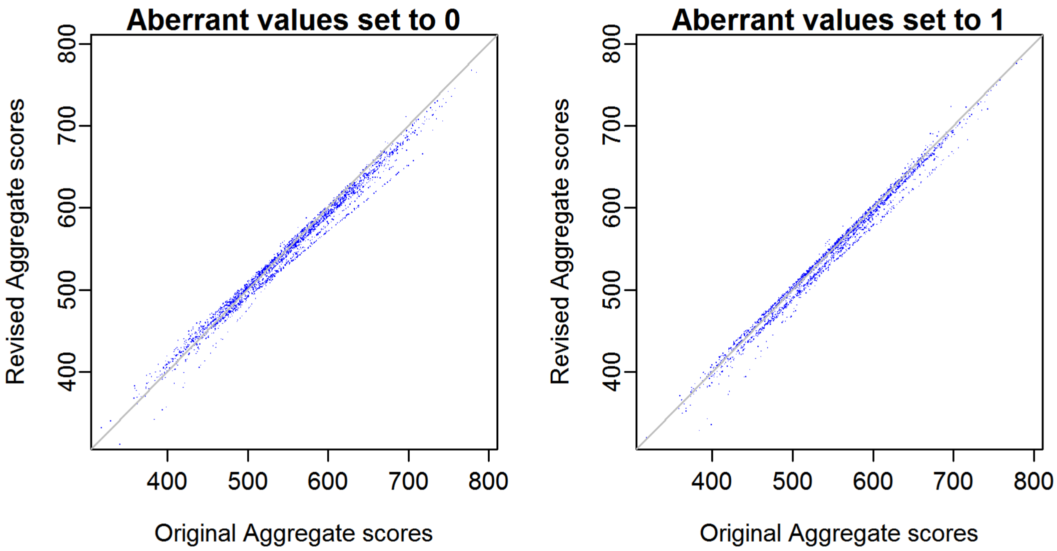

Figure 17: *Effect on Scaling scores when all aberrant values are set to 0, and when all aberrant values are set to 1.*

As a further test, 2000 sets of simulated scores were generated with d_i set to 1 for all students. The average value of MAE for these revised simulated scores is 8.8 compared to 8.9 for the original simulated scores. So the aberrant score process has a very small (but statistically significant) *negative* effect on the effectiveness of the algorithm.

This counter-intuitive result can be possibly be explained by the increase in sample size. The effect of the aberrant score process is to remove potential outliers, thus reducing the effective sample size, but also removing observations that may not adequately reflect student ability. However, with fewer observations it is harder to estimate the underlying student ability levels, thus causing a small increase in the average MAE. In any case, the effect is very small and I do not recommend that the aberrant score process be modified on the basis of this result.

(e)

Given the improvement in the scaling scores with the inclusion of the AST, it is possible that weighting the AST more highly may result in further improvement. To test this idea, I simulated scores with the AST weight in (3) ranging from 0 to 1 (instead of the current 0.21). For each possible value of the AST weight, I simulated 500 data sets and computed the MAE for each one. The average MAE values are shown in Figure 18 with the smooth line being computed using loess (Cleveland and Devlin, 1988). The red point indicates the current value of 0.21 for the AST weight. From this graph it is clear that an increase in the AST weight will result in a decrease in the MAE value, thus give a better scaling algorithm with more accurate estimates of each student's ability. An MAE value of about 8.2 is possible with an AST weight of 1.0.

The reduction in MAE that occurs with an increase in the AST weight is relatively large and I recommend that a change be made to increase the AST weight to at least 0.8.

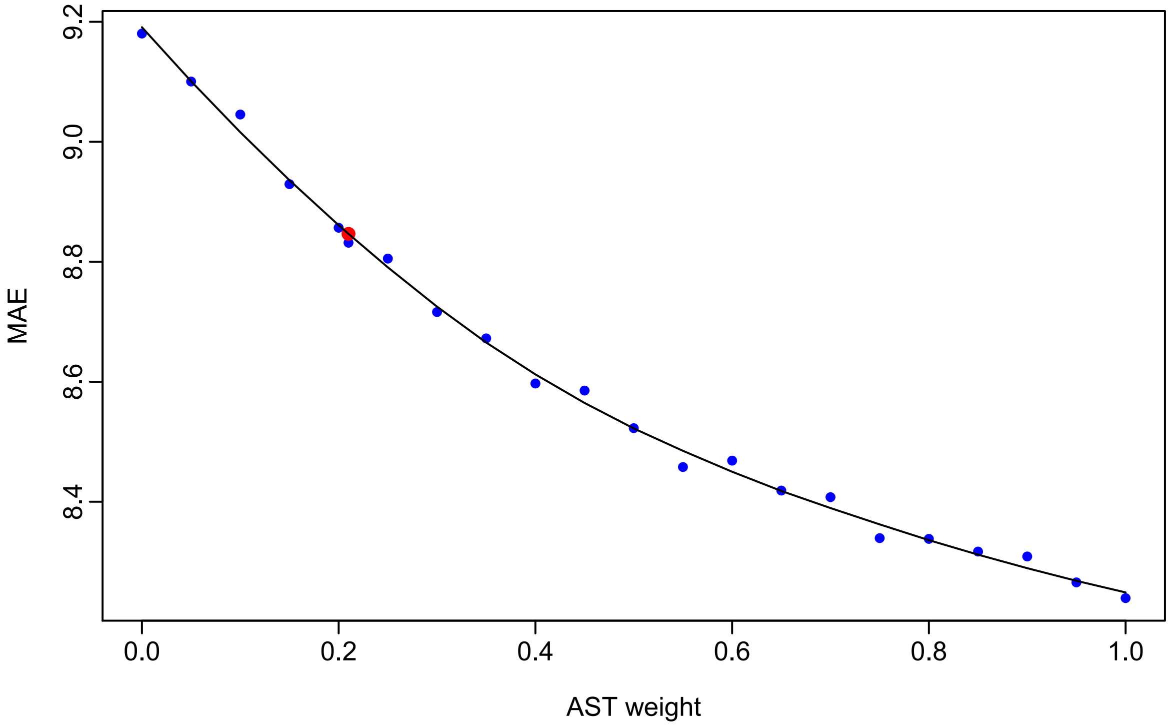

Figure 18: *Effect of AST weight on MAE. The existing weight of 0.21 is marked in red.*

10 Impact of CLD students

(g)

To assess the impact of students from a culturally and linguistically diverse background, I considered what would happen to the scaling results if these students were treated the same as non-CLD students. This has already partly been considered in the analysis of the effect of making all aberrant values equal to 1 (see previous section). As a further analysis, I simulated data according to the scheme described in Section 6, but then in scaling the scores, the CLD status of students was ignored; that is, all students were treated as if they were non-CLD in the calculation of aberrant values. Two thousand simulations were carried out and the average MAE change was an increase of 0.02. While this is a very small difference, the result was statistically significant. The change in MAE for CLD students and for non-CLD students was of a similar size and significant for both groups. The results are shown in the table below along with 95% confidence intervals for the mean difference.

Group	Normal scaling	Scaling ignoring CLD status	95% CI for difference
All students	8.86	8.88	[0.008, 0.033]
CLD students	8.85	8.88	[0.011, 0.041]
non-CLD students	8.86	8.88	[0.008, 0.033]

Table 6: *MAE values for normal scaling and modified scaling (treating all students as non-CLD).*

This indicates that the different weights for CLD and non-CLD students have slightly improved the scaling results. It also suggests that further improvement may be possible by modifying the values of L and U. I tried varying L from -80 to 10 and varying U from 10 to 80. For each combination of L and U, I simulated 200 data sets and calculated the average MAE. However, there was no significant improvement over existing results.

11 Conclusions

Each of the items in the terms of reference have been addressed in the report. In this concluding section, I summarise the results.

(a) *How the correlations in the ACT scaling process compare with the correlations in scaling in other jurisdictions?*

Page 25: The correlation between the AST scores and scaling scores is slightly larger than the equivalent correlation for Victoria. No other information is available about correlations in other states.

(b) *What levels of correlation are deemed acceptable for scaling purposes?*

Page 26: Positive correlations are required for scaling purposes.

(c) *The effects on ATARs of courses where there are low correlations in the scaling between school-based course scores and the AST.*

Page 27: Simulations showed that a scaling group with low positive correlations has almost no effect on average aggregate scores and results in a small decrease in the standard deviation of aggregate scores. Consequently, the effect on ATARs will also be small. The Mean Absolute Error of a scaling group with low correlation is about 9.6 compared to 8.2 for the scaling groups with highest correlation and an average of 8.9 across all students.

(d) *The effectiveness of current measures used by the BSSS to enhance correlations between school-based assessment and the AST.*

Page 29: The inclusion of AST scores in the scaling score has improved the Mean Absolute Error from 9.3 to 8.9. The introduction of aberrant values in the scaling score equation has resulted in a small increase in Mean Absolute Error from 8.8 to 8.9.

(e) *Additional/replacement measures that could be implemented to increase correlations in the scaling process.*

In the preceding sections, two variants on the scaling algorithm have been seen to give improved Mean Absolute Errors:

1. Increasing the AST weight from 0.21 to at least 0.8 (page 30);
2. Dropping the aberrant score weighting (page 29).

Increasing the AST weight to at least 0.8 will reduce the MAE from 8.86 to less than 8.35, and it is recommended that this change be made. Dropping the aberrant score weighting has only a very small effect, and is not recommended at this stage.

(f) *What should be done with courses that do not/can not reach an acceptable level of correlation for scaling purposes?*

Page 28. The effect of one scaling group on the aggregate scores is small, even when the correlations are low. It is worth noting that small- and intermediate-group procedures exist to avoid possibly highly variable (including low) correlations.

(g) *The impact of the presence of CLD students on the outcomes for scaling groups and on overall college results.*

Page 31: If all students were treated as non-CLD in the aberrant score process, the aggregate scores would be slightly less accurate for both CLD and non-CLD students.

(h) *Abolition of the modified AST papers for CLD students and use of other processes, such as the aberrant score policy, to address correlations for CLD students and their impact on scaling groups.*

In general, the different weights used for CLD students seem to have very little effect on scaling. I cannot comment on the abolition of modified AST papers as there is no information available about the effect that would have. The aberrant score policy is addressed under item (d).

References

Cleveland, W. S. and S. J. Devlin (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* **83**, 596–610.

Daley, D. J. (1989). “Determining relative academic achievement for fair admission to higher education”. Report to ANU, Canberra CAE and ACT Schools Authority.

Daley, D. J. (1995). Two-moment scaling formulae for aggregating examination marks. *Australian Journal of Statistics* **37** (3), 253–272.

Hyndman, R. J. (2010). *Nominal ATAR scores predicted by GAT components: 2010 analysis*. Report for VTAC. Monash University Statistics and Econometrics Consulting Service.