Caribbean Curriculum
Vol. 26, 2018/2019, 178-197

TO SCALE OR NOT TO SCALE: Insights from a Study of Grade Comparability in CXC Examinations

Stafford Alexander Griffith

This study sought to ascertain the extent to which the use of statistical scaling procedures to establish comparable Grade III/IV cut scores for different examinations of the same subject resulted in cut scores that were comparable to those obtained when judgmental procedures are used. The study used data from three subjects of the Caribbean Examinations Council (CXC) where the same Grade III/IV cut scores were retained over a five-year period. Through linear transformation, the Grade III/IV cut scores for each subject were converted to scale scores on a base form. The extent to which scaling procedures validated CXC's maintenance of the same Grade III/IV cut scores across years was considered. For all but one of the 11 cut scores considered in this study, the calculated scale scores were dissimilar from those retained by CXC. The calculated scale scores, therefore, could not be regarded as comparable to the cut scores established by CXC through the use of its judgmental procedures. However, it was found that the direction of change in the proportion candidates obtaining Grades I to III with the CXC judgmental procedures was consistent with the outcome of scaling.

The Importance of Comparability of Grades in a Public Examination

Comparability of standards between examinations and the grades awarded from them has been the subject of discussion and debate for many years. Comparability is an important consideration in a public examination. It is the extent to which results from separate examinations embody the same standard (Ofqual, 2015). Given that students taking the same examination at different sittings may compete in the same market for the same jobs, the same scholarships or the same places in institutions of further education and training, public examinations must find a way of assuring comparability of the scores and/or grades of the same examination regardless of the sittings at which it was taken.

Although it may well be true that comparability theory is not at all

well-developed (Newton, 2007), several significant contributions have been made to the development of a sound conceptual and theoretical grasp of issues of comparability which have practical application in treating with test scores from different tests (Angoff, 1971; Baird, Cresswell, & Newton, 2000; Christie & Forrest, 1981; Cresswell, 1996; Crisp, 2017; Flanagan, 1951; Goldstein, 2001; Holland & Dorans, 2006; Linn, 1993; Newton, 2005; Peterson, Kolen, & Hoover, 1993; Wiliam, 1996; Young, Holtzman, & Steinberg, 2011). Essentially, comparability refers to the application of the same standards across different examinations, or different versions of the same examinations (Crisp, 2017; Educational Testing Service [ETS], 2014; Newton, 2007). When more than one form or version of an examination is being used, a way must be found to equate scores and/or grades so that the outcomes reported are comparable, regardless of the particular examination taken. As Newton (2007) declares: "It either involves adjusting grade boundary marks on version 2 to reflect the standard of version 1, or it involves adjusting all marks on version 2 to reflect the standard of version 1" (p. 32).

Scaling techniques have been widely used as a way of obtaining comparability of standards across different examinations, or different versions of the same examinations. Following an examination of several scaling methods for comparability of standards in examinations, Lamprianou (2009) concluded that it was very difficult to take a clear position either for or against the use of statistical or scaling methods to achieve comparability between subjects. While many public examination boards use various scaling procedures, there are a number that continue to use the alternative of a judgemental approach to achieve comparability across different examinations, or different versions of the same examinations. The Caribbean Examinations Council (CXC) falls in the latter category. The jury may still be out on whether the two procedures used in public examinations (scaling and judgements) can achieve similar outcomes or whether one is better than the other. To scale or not to scale is still the question over which independent researchers and academics in the international community continue to ponder in seeking comparability of grades and scores from different examinations. This paper explores the issue of scaling compared to judgement, with special reference to the comparability of scores across different forms of an examination offered by the CXC. The Council is a public examinations board, established in 1972, that currently serves 16 member territories, most of which are independent, past colonies of Britain.

Scaling Versus Judgement

Scaling

Flanagan (1951), in the first edition of *Educational Measurement*, had wrestled with the need for comparability of test scores across different forms and ways of obtaining that comparability. Flanagan referred to comparability of test scores obtained from different forms of a test as a "fundamental necessity" (p. 747). He further suggested a number of ways of achieving that comparability, including a method for equating of raw scores (p. 750).

Equating is a form of scaling that involves the conversion or transformation of raw scores on a form of a test so that there is comparability with scores obtained from a previous form of that test. Various statistical methods have been proposed and used to achieve this transformation of the scores of one test to the scores of another. Angoff (1971), in the second edition of *Educational Measurement*, recommended an equating measure which would "convert the system of units of one form to the system of units of the other" (p. 562). As Holland and Dorans (2006) explained, the goal of such equating is to allow scores from both tests to be used interchangeably. Proponents of this approach, therefore, would see equating procedure as a means of establishing equivalence of the cut score of one form of a test or examination to that of another form of the test or examination taken subsequently.

Cut scores partition the distribution of scores into categories which identifies the range of scores associated with various grades. The identification of cut scores or grade boundary marks is an important step in many public examinations, including those of the CXC. As Robinson (2007) explains:

> This process is important because even ostensibly parallel versions of the same subject examination can differ, from one year to the next, in terms of how easy or hard it is to achieve marks… The process of adjusting grade boundaries is what allows examining boards to claim comparability. (p. 25)

Essentially, equating is undertaken to adjust for differences in difficulty when two forms of the same test have been constructed to be as similar as possible, by utilising the same test construction blueprint. Note that the tests are already assumed to be equivalent but, because of the use of different items, this may not be fully achieved. Writing in the fourth edition of *Educational Measurement*, Holland and Dorans (2006) point out that "the purpose of equating tests is to allow the scores from each test to

be used interchangeably, as if they had come from the same test" (p. 193). The *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME), 2014) adds further clarity to the understanding of equating, by pointing out that:

> Equating involves small statistical adjustments to account for minor differences in the difficulty of the alternate forms. After equating, alternate forms of the same test yield scale scores that can be used interchangeably even though they are based on different sets of items. (p. 97)

Felan (2002) proposes that there are three traditional methods of equating – mean equating, linear equating and equipercentile equating. While mean equating may be acceptable under certain conditions (Felan, 2002; Kolen & Brennan, 2014; Livingston & Kim, 2010; Skaggs, 2005), it is the second and third methods (linear and equipercentile equating) that are more frequently used and referenced in the literature (Angoff, 1971; von Davier, Holland, & Thayer, 2004; Flanagan, 1951; Holland & Dorans, 2006; Kolen, 1984; Kolen, 2006; Peterson et al., 1993; Wang, Lee, Brennan, & Kolen, 2008; Yin, Brennan, & Kolen, 2004). Both provide good approaches for establishing comparability of scores from two tests constructed to be similar.

Equipercentile equating, where practicable, is likely to provide greater similarity between distributions of equated scores than does linear equating (Felan, 2002). However, linear equating has some advantages in calculating comparable scores since this procedure relies largely on means and standard deviations. Therefore, in cases where the raw scores on the two tests in question have not been converted into percentile rankings, or may not be available to allow the researcher to do the conversion to percentiles due to security and other related reasons, linear transformation procedures provide an alternative for calculating scores on the second tests that are comparable to scores on the first test. This was an important consideration in selecting equating procedures for the investigation reported in this paper.

In the equating process, the linear transformation of the raw scores distribution of one form of the test to the distribution on the original of first form, is usually a first step in the process. A second step often involves the conversion of the equated score from the new form to a reporting scale previously developed for the first or base form (Holland & Dorans, 2006).

The second step was not necessary for the investigation reported in this paper since the interest was limited to the extent to which scaling procedures validated the maintenance of the same cut scores by CXC across years.

Judgements

A number of public examination boards rely on expert judgements in determining grade awards. However, this reliance of examinations has been questioned for several decades. For example, Skurnik and Hall (1969) reported on a 1966 monitoring experiment where a wide range of individual judgements were found when 100 Certificate of Secondary Education (CSE) English scripts were re-graded by 14 chief examiners or chief moderators. Skurnik and Hall found wide differences between the individual judgements of the markers.

Also, a study by Good and Cresswell (1988) examined whether teams of awarders made grade-awarding judgements that were consistent with that of other teams, and whether standards that were judgementally determined and statistically determined were comparable. The latter is, of course, the central issue in the current article. The authors used scripts from a sample of 414 History candidates, 574 Physics candidates and 393 French candidates in a March 1986 "experimental differentiated examination" for candidates who had entered for the General Certificate of Education (GCE) O-level and/or the Certificate of Secondary Education (CSE) examination in the relevant subjects. Based on their findings, Good and Cresswell (1988) raised questions about the capacity of experts to establish grade boundaries on two versions of an examination paper, one of which was clearly more demanding than the other. Good and Cresswell (1988) found that the examiners failed to compensate adequately for the difference in difficulty between the two papers. As a result, it was easier for the same candidates to gain higher grades on lower level versions of the examination than they did on the harder paper. The authors suggested that comparable standards of performance should not be defined solely in terms of the awarders' judgements but that statistical comparability based on the candidates' responses should also be taken into account.

Subsequent research has confirmed the doubts cast on expert judgement of examiners. These suggest that with the advances made in statistical techniques, examination boards may employ statistical procedures in grade awarding which produce much smaller margins of error than the judgemental approach (Cresswell, 1997; Baird & Dhillon, 2005; Scharaschkin & Baird, 2000; Stringer, 2012).

The CXC grade awarding procedure mirrors that of the UK boards

which has been well described by Robinson (2007). It requires the division of mark scales into mark bands, so that the marks within each band represent a particular grade. Such partitioning of the distribution of scores into categories is often seen as a critical step for the proper use of test scores (AERA, APA & NCME, 2014).

The judgemental approach to grade awarding usually require several days of review of information and discussions. The CXC Subject Awards Committee (SAC) is critical to this process. A SAC is established for each subject examination. It comprises the Examining Committee for the subject and a CXC measurement specialist.

The critical part of the grade awarding exercise requires the SACs to assure comparability of the standards for a particular grade across years. This requires the reading of scripts at each grade boundary from previous years to determine whether the application of the cut scores that were established at the paper setting exercise, or in previous year or years, should be maintained in light of the actual demands of the examination. The SAC must review all available information about the paper-setting, marking and other factors which may impact the marks obtained by students to determine whether a previously recommended boundary satisfactorily represents a particular grade boundary performance. The relevant statistics on the performance of students, available at the grade-awarding exercise, is carefully examined and any implications for the cut score considered. If it is confirmed that a previously recommended grade boundary is satisfactory, this mark is confirmed as the grade boundary or cut score. If this is not the case, then the SAC must determine what adjustment must be made to the cut score for a grade by making either an upward or downward adjustment based on the combination of factors considered.

Essentially, at the grade awarding exercise, the SACs are required to make a judgement about whether the paper-setting and marking exercise resulted in examinations which are comparable in demand to examinations for previous years and, if not, what adjustments should be made to cut scores to assure equivalence of the demand for the award of a grade in the current year. In recent years, changes to the cut scores have been made very infrequently. Any deviation recommended during grade awarding must be fully justified.

Despite the rigor of these processes, they lack the level of precision associated with statistical approaches. The conclusion by Zeiky and Perie (2006), following their discussion of various procedures for establishing cut scores, summarises how elusive cut scores based on

judgements may be. According to the authors:

> It is impossible to prove that a cut score is correct. Therefore, it is crucial to follow a process that is appropriate and defensible. Ultimately, cut scores are based on the opinions of a group of people. The best we can do is choose the people wisely, train them well in an appropriate method, give them relevant data, evaluate the results, and be willing to start over if the expected benefits of using the cut scores are outweighed by the negative consequences. (p. 23)

The point made by Zeiky and Perie (2006) is worth noting. It strengthens the case for seeking alternative approaches that are less elusive than judgements for establishing cut-scores.

Aim of the Study

The main aim of the study was to explore the extent to which the use of statistical scaling procedures to establish comparable Grade III/IV cut scores for different CXC examinations of the same subject across years, provide results that are comparable to those obtained when a judgemental approach was used. The Grade III/IV cut scores define the performance level at and above which students are awarded grades generally regarded as satisfactory.

Procedures

Given the nature of this study, the specific years for which data were used was not a material consideration. It was important, however, that data for comparability across years be derived from examinations that were constructed as parallel forms. This required the identification of subject examinations for which the syllabuses that guided the development of the test specifications and the development of the tests for the particular examinations had not changed over the years spanned by this research.

Subjects with small entries, in this case less than 10,000, were excluded from consideration. The small entries were not considered sufficiently stable across years. Generally, they fluctuated considerably from year to year. It was felt that the size of the entries and the possible fluctuating nature of the population taking the examination may introduce variances which could not be effectively addressed in this study.

Of particular interest were those subject examinations where CXC did not change the Grade III/IV cut scores over years but where, nevertheless, there was considerable variance across years in the

proportion of candidates who obtained the passing Grades of I to III (at or above the Grade III/IV cut score). The retention of the same cut scores across years, without adjustment, suggests that the judgemental procedure used by CXC endorsed them as requiring the same grading standard from year to year in the particular subject, even where there was considerable variance across years in the proportion of candidates who obtained satisfactory or passing grades.

Guided by these considerations, the researcher reviewed the syllabuses of 34 CXC Caribbean Secondary Education Certificate (CSEC) examinations to identify those that were appropriate for the study over a five-year period. The five-year period was considered adequate for the data requirements of this study. The period 2007 to 2011 was selected since a number of syllabuses could be identified in that period which satisfied the aforementioned considerations. These, therefore, were subjects with larger entries for which (a) the tests and examinations were constructed as parallel forms to assess achievement over that period; (b) the Grade III/IV cut score remained unchanged over the five-year period; and (c) there was considerable variance in the proportion of candidates obtaining Grades I to III across years. Given the nature of this study, data from any period would be acceptable as long as the three aforementioned considerations were satisfied.

Three CSEC subjects which best satisfied these criteria were selected from the 34 CSEC subjects offered by CXC during the period 2007 to 2011. These were Chemistry, English A (English Language), and English B (Literature).

Analysis and Findings

Comparability of Judgemental Cut Scores and Scale Scores

The procedures outlined by Kolen (2006) for the linear transformation of scores were used to convert the Grade III/IV cut scores used by CXC for each subject to scale scores on the base form. In doing so, the 2007 year of the examination for a subject was the targeted base year.

The procedures involve a raw-to-raw equating which would link the raw score used by CXC as the Grade III/IV cut score for the base year of a subject examination and the Grade III/IV cut scores used for subsequent administrations of the same examinations. For this study, it was not necessary to pursue the second step in an equating process which

would require the placing of the scores from both the base year examination and the equated scores from the subsequent examinations on a reporting scale, constructed from the base examination. The interest was limited to the extent to which scaling procedures validated the maintenance of the same Grade III/IV cut scores across years by the CXC judgemental approach. For this purpose, the first step, the raw-to-raw equating of cut scores, was sufficient.

The formula for raw-to-raw transformation of Grade III/IV cut scores for an examination to Grade III/IV cut scores of the base year examination which was adopted from Kolen (2006) is as follows:

$$S(x) = \frac{\sigma_s}{\sigma_x}x + [\mu_s - \frac{\sigma_s}{\sigma_x}\mu_x]$$

where

S(x) is the scale score or transformed score

μ_x and σ_x are the mean and standard deviation of the raw score

μ_s and σ_s are the desired mean and standard deviation of the scale scores

In this study μ_s *and* σ_s were derived from the distribution of scores of the base year of the examination while μ_x *and* σ_x were derived from the distribution of scores for a subsequent year of the examination. The interest in this analysis was the extent to which the calculated scale score validated the Grade III/IV cut scores which the judgemental procedure used by CXC maintained unchanged over the five-year period 2007 to 2011 for each of the three subjects investigated. In the analysis which follows, an absolute difference scores of 2.00 points and above (that is, ±2 beyond the CXC cut score) is regarded as a deviation of significance. The use of an absolute deviation of 2.00 took into account that in CXC examinations, a difference of 1.00 or 2.00 may be viewed as a marginal difference from a particular cut score and this may be taken into account, under particular circumstances, for example, in hardship cases, in adjusting the achieved grades of particular candidates.

Table 1 shows the comparison of the Grade III/IV base examination cut score for Chemistry and the calculated scale scores for subsequent examinations. It became evident during the analysis that the 2007 means and standard deviations obtained through CXC could not reasonably be accurate, given how different they were from the statistics for the other years and the very high and unusual scale scores they generated. No explanations which were considered helpful for this study were obtained from the organisation. The researcher, therefore, decided to use the 2008 as the base year for Chemistry and calculated scale scores for 2009, 2010 and 2011.

Table 1. ***Comparison of the Grade III/IV base examination cut score for Chemistry and calculated scale scores for subsequent examinations***

Year	Mean	Standard Deviation	Cut Score	Scale Score	Difference
2008	101.52	32.45	84	-	-
2009	110.21	31.84	84	74.79	-9.21
2010	103.29	32.23	84	82.04	-1.96
2011	99.40	32.48	84	86.12	2.12

Table 1 shows that CXC applied a cut score of 84 for Grade III/IV for all years for CSEC Chemistry. It shows, however, that the calculated scale scores were 74.79, 82.04 and 86.12 for 2009, 2010 and 2011, respectively. The deviations of the scale scores from the CXC cut score of 84 were -9.21, -1.96 and 2.12 for the three consecutive years 2009, 2010 and 2011, respectively, when 2008 was used as the base year.

For English A, 2007 was retained as the base year and scale scores were calculated for 2008, 2009, 2010 and 2011. Table 2 shows the comparison of the Grade III/IV base examination cut score for English A and the calculated scale scores for subsequent examinations.

Table 2. ***Comparison of the Grade III/IV base examination cut score for English A and calculated scale scores for subsequent examinations***

Year	Mean	Standard Deviation	Cut Score	Scale Score	Difference
2007	77.99	23.34	79	-	-
2008	75.46	24.40	79	81.39	2.39
2009	81.36	23.90	79	75.68	-3.32
2010	88.24	23.59	79	68.84	-10.16
2011	88.34	23.32	79	68.65	-10.35

Table 2 shows that for English A, the CXC Grade III/IV cut score of 79 was maintained across the five years. However, using 2007 as the base year, the scale score for the years 2008, 2009, 2010 and 2011 were calculated to be 81.39, 75.68, 68.84 and 68.65, respectively. The table shows that the deviations of the scale scores from the CXC cut score of 79 were 2.39, -3.32. -10.16 and -10.35 for the four consecutive years 2008 to 2011, when 2007 was used as the base year.

Table 3 shows the comparison of the Grade III/IV base examination cut score for English B and the calculated scale scores for subsequent examinations.

Table 3. ***Comparison of the Grade III/IV base examination cut score for English B and calculated scale scores for subsequent examinations***

Year	Mean	Standard Deviation	Cut Score	Scale Score	Difference
2007	71.96	21.34	65	-	-
2008	62.94	21.68	65	73.98	8.98
2009	66.66	22.85	65	70.42	5.42
2010	80.49	22.26	65	57.09	-7.91
2011	76.16	22.55	65	61.36	-3.64

Table 3 shows that for English B, the CXC Grade III/IV cut score of 65 was retained across the five years and that the calculated scale scores were 73.98, 70.42, 57.09 and 61.36 for the years 2008, 2009, 2010 and 2011, respectively, and that the deviations of the scale scores from the CXC cut score of 65 were 8.98, 5.42, -7.91 and -3.64 for the four consecutive years 2008 to 2011, when 2007 was used as the base year.

All deviation scores for the three subjects, except the one for 2010 Chemistry, were above the absolute deviation of 2.00 points (that is, ±2 beyond the CXC cut score) established for considering a deviation score to be dissimilar (or to deviate significantly) from the CXC cut score. It was concluded, therefore, that in all cases, except one, the calculated scale scores were dissimilar from the Grade III/IV scores used by CXC.

Comparability of Candidate Results Obtained from Scaling Procedures and the Judgemental Approach

This aspect of the study sought to ascertain the extent to which the results obtained from the application of the scaling procedures, were similar to the results obtained through the CXC judgemental approach, in assuring comparability of the Grade III/IV cut scores across years.

In the case of Chemistry, the scale scores of 74.79 for 2009 resulted in a difference score of -9.21 points from the CXC Grade III/IV cut score. This suggests that the CSEC examination for that year was more difficult than the examination for the base year. The judgemental approach used by CXC did not adjust the Grade III/IV cut score of 84. The application of the scale score of 74.79 would allow more candidates to obtain Grades I to III in 2009 than was the case with the CXC cut score of 84.

It would appear, however, that the CXC procedure, while maintaining the same cut score, would have instituted measures prior to the grade awarding exercise to take into account an examination that appeared to be more difficult than the examination in the base year. The CXC marking exercise for the written papers makes provisions for such adjustments. As part of the standardisation process, which precedes actual marking of scripts, a representative sample of scripts of candidates for the subject examination is reviewed, using the mark schemes prepared earlier.

During that exercise, the mark schemes may be modified to take into account "insights gained from studying candidates' responses… encountered in the selected sample" (CXC, 2004, p. 4).

Psychometrically imprecise though these measures may be, they seemed to have responded to the recognition that the 2009 examination was more difficult than that of 2008; used in this study as the base year for Chemistry. Consequently, it was observed that despite CXC's maintenance of the same cut scores, the proportion of students who obtained Grades I to III was higher in 2009 than in 2008. The direction of the change in the proportion of candidates obtaining Grades I to III accords with what may be reasonably expected, had the scale scores been used to adjust the proportion of students obtaining Grades I to III.

A similar consistency of the direction of change in the proportion of students obtaining Grades I to III, which the shift of cut score would suggest, was observed in the percentages of students obtaining those grades in 2011. Here, the positive deviation of 2.12 points from the CXC cut score of 84 would suggest that the 2011 examination was less difficult, and that the cut score should have been higher than the 84 used by CXC. Consequently, one would expect a downward shift in the proportion of students obtaining Grades I to III in 2011 compared with the base year. Table 4 shows that this was in fact the case.

The scale score of 82.04, with a deviation of -1.96 for 2010, does not satisfy the condition of ±2 for it to be regarded as dissimilar from the CXC cut score of 84, and not much difference could be expected between the proportion of candidates obtaining Grades I to III in the 2008 base year and 2010. This was the case as 62.35 percent of candidates obtained Grades I to III in 2010 compared to 62.37 in the 2008 base year.

Table 4 shows that the pattern observed for Chemistry, was also observed for all years in English A and English B. Table 5 summarises the findings about the extent to which the results obtained from statistical scaling procedures, generally supported the direction of change resulting from the use of CXC's judgemental approach to grade awarding. It shows that the general direction of the adjustments was in keeping with the calculated scale scores.

Table 4. ***Percentage obtaining Grades I to III for the period 2007 – 2011 in the three selected CSEC subject examinations***

Year	Subjects								
	Chemistry			English A			English B		
	Entries	% Grades I – III with CXC cut score of 84	Deviation of scale score from CXC cut score	Entries	% Grades I - III with CXC cut score of 79	Deviation of scale score from CXC cut score	Entries	% Grades I - III with CXC cut score of 65	Deviation of scale score from CXC cut score
2007	12158	NA	-	103380	45.80	-	20086	64.13	-
2008	12475	62.37	NA	106161	42.03	2.39	19891	46.91	8.98
2009	12902	70.92	-9.21	110219	53.09	-3.32	18913	51.96	5.42
2010	14344	62.35	-1.96	123546	57.76	-10.16	20605	74.03	-7.91
2011	16234	55.34	2.12	114381	59.81	-10.35	21502	66.85	-3.64

Table 5. ***Consistency of the general direction of change in percentage of candidate obtaining Grades I to III using CXC's judgemental procedures with direction of adjustment suggested by the calculated scale scores***

	Subjects								
	Chemistry			**English A**			**English B**		
Year	Deviation from CXC Cut Score	Observed Shift in Percentage Obtaining Grades I-III	General Direction of Change in Percentage Grades I-III	Deviation from CXC Cut Score	Observed Shift in Percentage Obtaining Grades I-III	General Direction of Change in Percentage Grades I-III	Deviation from CXC Cut Score	Observed Shift in Percentage Obtaining Grades I-III	General Direction of Change in Percentage Grades I-III
2008	NA	NA	NA	2.39	Smaller	Consistent	8.98	Smaller	Consistent
2009	-9.21	Larger	Consistent	-3.32	Larger	Consistent	5.42	Smaller	Consistent
2010	-1.96	Similar	Consistent	-10.16	Larger	Consistent	-7.91	Larger	Consistent
2011	2.12	Smaller	Consistent	-10.35	Larger	Consistent	-3.64	Larger	Consistent

Discussion

The findings in this study are similar to those of Cresswell (2000). He compared a total of 108 boundary marks set by the examiners with those that would have been set to produce statistically equivalent outcomes. Although it might be expected that random fluctuations in the sample of students taking examinations in any one year would result in some changes in outcome, Cresswell found that most changes represented large swings in outcome compared with the previous year.

As in the current study, Cresswell found that there was clear evidence that examiners had, in fact, responded to changes in difficulty of the examinations. He found that 77 per cent of the boundary marks moved in the direction predicted by the statistical evidence. The current study found that, for all examinations where the size of the deviations of the scale score from the CXC cut scores warranted further attention (based on an absolute ±2-point difference), the proportion of candidates obtaining Grades I to III moved in a direction consistent with the increased or decreased difficulty of the examination. Also, in the current study, it was found that though the direction of change was correct, these changes seem to represent overestimates or larger swings than the statistical approach to defining boundary marks would suggest.

Wilmott (1977) points out that, almost by definition, any approach to a study of the comparability of grading standards needs to make a number of assumptions regarding the issues under consideration. He further notes that there can be considerable disagreement in the interpretation of information based on the extent to which it is believed that the assumptions made are justified (1977, p. 97).

The CXC examinations used in this study were constructed to be equivalent across sittings. The judgemental procedure used by CXC was intended to assure the equivalence of grades across years for the same examination.

One might reasonably infer that CXC's maintenance of the same cut scores across the five-year period for those examinations in the investigation (CSEC Chemistry, English A and English B), rested on the assumptions that the judgemental procedure assured the equivalence of the standards for those cut scores, from year to year. However, the results of the statistical scaling procedure used in this study, raise questions about that assumption.

It is accepted that some small changes in the proportion of students

obtaining Grades I to III in a CXC examination may be expected from year to year, given random fluctuations of attributes of the student population taking a particular subject examination each year. However, the larger changes in the proportion of students obtaining Grades I to III would suggest that there were other factors at play. These may include intrinsic differences in the nature of the examinations which are not readily evident, or inconsistencies in grading practices across years.

Conclusion

This study sought to ascertain the extent to which the use of statistical scaling procedures to establish comparable Grade III/IV cut scores for different examinations of the same subject across years, resulted in cut scores that were comparable to those obtained when the judgemental approach was used.

It was found that for all but one of the 11 cut scores considered in this study for the three subjects (three for Chemistry, four for English A and four for English B), the calculated scale scores were dissimilar from the Grade III/IV cut scores used by CXC. The calculated scale scores, therefore, could not be regarded as comparable to the cut scores established by CXC through the use of its judgemental procedure.

The study also examined the extent to which the proportion of candidate obtaining Grade I to III (the acceptable or passing grades), based on the judgemental approach used by CXC to establish comparability of the Grade III/IV cut scores across years for a subject examination, were comparable to the proportion obtained when the calculated scales scores were used. It was found that, despite the maintenance of the same Grade III/IV cut scores across years in a subject examination, the direction of change in the proportion of candidates obtaining Grades I to III accorded with what may be reasonably expected, had the calculated scale scores been used to adjust the proportion of students obtaining Grades I to III. It appears that although CXC maintained the same Grade III/IV cut scores across years for the subjects investigated, the organisation instituted measures prior to the grade awarding exercise to take into account an examination that appeared to be more difficult or less difficult than the examination for the base year.

The judgemental gauge seems to have been good enough to determine when an examination was more difficult or less difficult than the examination for the base year. However, the calibration appeared to be psychometrically imprecise, based on the magnitude of the change in the

proportion of candidates obtaining Grades I to III, compared with the proportion in the base year. It seems that the CXC procedure may be overcompensating for examinations which were less difficult or more difficult.

Based on the findings of this study, it is concluded that there was a lack of comparability between the cut scores used by CXC to maintain standards across years and the scores that were derived through linear scaling. It was further concluded that although the proportion of students obtaining passing grades (Grades I to III) when the CXC cut scores were applied were different from the proportion that would obtain those grades when the scaling procedure was applied, the CXC judgemental procedure has been picking up the instances where the subject examination for a particular year was more difficult or less difficult than that of a base year, and measures, though imprecise, were implemented to deal with this change in difficulty.

Recommendations

A linear transformation of scores was used in this study to convert the Grade III/IV cut scores used by CXC to scale scores on a base form. Taking into account some possible lack of equivalence in populations across years, it may be more appropriate, in future studies, to consider the use of other scaling procedures which take this factor into account. These include the use of embedded equating items on two forms of the test, on the basis of which adjustments may be made on scores derived from the second form of the test.

This paper raises a number of issues related to the use of scaling and judgements. It is very challenging to determine whether one of these procedures should be preferred over the other. What is clear is that they both have limitations. The results of this research suggest that no radical shift should be made from the use of one procedure in favour of the other, without further and more compelling evidence than was found in this study. It has implications for decision making about continuing or discarding the use of judgemental or scaling procedures, for the award of scores or grades by examinations units and boards in the Caribbean, as well as in the wider global community.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014).

Standards for Educational and Psychological Testing. Washington, DC: Authors.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Educational Research Association.

Baird, J., Cresswell, M. J., & Newton, P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education*, *15*(2), 213–229.

Baird, J., & Dhillon, D. (2005). Qualitative expert judgements on examination standards: Valid, but inexact. *Internal report RPA05 JB RP 077*. Guildford, UK: Assessment and Qualifications Alliance.

Caribbean Examinations Council. (2004). *Guidelines for Marking CXC Examinations.* St. Michael, Barbados: Author.

Christie, T., & Forrest, G. M. (1981). *Defining Public Examination Standards.* London: Schools Council Research Studies, Macmillan Education.

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis, (Eds.), *Assessment: Problems, Developments and Statistical Issues* (pp. 57-58). Chichester, UK: John Wiley.

Cresswell, M. J. (1997). *Examining judgements: Theory and practice of awarding public examination grades*. London: Institute of Education, University of London.

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 69-104). Oxford, UK: Oxford University Press for The British Academy.

Crisp, V. (2017). Exploring the relationship between validity and comparability in assessment. *London Review of Education*, *15*(3), 523-535. Retrieved from https://doi.org/10.18546/LRE.15.3.13

Educational Testing Service. (2014). *Standards for quality and fairness*. Princeton, NJ: Author.

Felan, G. D. (2002). *Test equating: mean, linear, equipercentile, and Item Response Theory*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, Texas, February 16.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 695-763). Washington, DC: American Council on Education.

Good, F. J., & Cresswell, M. J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, *14*(3), 263–281.

Goldstein, H. (2001). Using pupil performance data for judging schools

and teachers: Scope and limitations. *British Educational Research Journal*, *27*(4), 433–442.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education and Praeger Publishers.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, *9*(1), 25-44.

Kolen, M. J. (2006). Scaling and norming. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger Publishers.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking* (3rd ed.). New York, NY: Springer.

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education*, *35*(2), 205-226.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, *47*(2) 175-185.

Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education*, *12*(2), 105–123.

Newton, P. E. (2007). Contextualising the comparability of examination standards. In P.Newman, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 9-42). London: Qualifications and Curriculum Authority.

Ofqual. (2015). *Inter-subject comparability: A review of the technical literature: ISC Working Paper 2*. Coventry, UK: The Office of Qualifications and Examinations Regulation.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1993). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). Phoenix, AZ: American Council on Education and Oryx Press.

Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In P. Newman, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 97-123). London: Qualifications and Curriculum Authority.

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, *26*(3), 343–357.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, *42*(4), 309-330.

Skurnik, L. S., & Hall, J. (1969). *The 1966 CSE monitoring experiment:*

A report to the Schools Council. London: HMSO.

Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: The case for contextualised cohort-referencing. *Research Papers in Education*, 27(5), 535-554.

von Davier, A. A, Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating.* New York, NY: Springer.

Wang, T., Lee, W., Brennan, R. L., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651.

Wiliam, D. (1996). Standards in examinations: A matter of trust? *The Curriculum Journal, 7*(3), 293–306.

Wilmott, A. S. (1977). *CSE and GCE grading standards: The 1973 comparability study*. London: Macmillan.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*(4), 274-289.

Young, J. W., Holtzman, S., & Steinberg, J. (2011). *Score comparability for language minority students on the content assessments used by two states*, (Research Report. ETS RR-11-27). Princeton, NJ: Educational Testing Service.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service.