An intra-board comparison at syllabus level based on outcomes of rank-ordering exercises at component level

Louis Yim Cambridge International Examinations (The author is currently at the Singapore Examinations and Assessment Board, Singapore.)

Introduction

Ensuring the equivalence of standards of similar qualifications across different awarding bodies or across time within the same awarding body has been a salient area of research in educational assessment in England for some time. For the former, the rationale behind this research is that a number of examination boards in England offer public examinations which lead to the same qualifications, i.e. GCE A level and GCSE. Although each examination syllabus must conform to general qualifications criteria approved by the examinations regulator[1], and also to a common core of subject content, the syllabuses may differ between boards in other respects. A crucial question of whether it is equally difficult to obtain a given grade in a particular examination with one board than with another arises. In fact, this issue is not limited to England alone, but extends to other countries where candidates sit examinations which are claimed to be equivalent qualifications to the GCE A level and GCSE. For examinations taken within the same awarding body but at different times, the issue of equivalence of standards is more commonly known and is about maintaining the same standard for a particular examination between different administrations (e.g. in different years) within an awarding body. For example, the standard of a given grade in a particular examination from three years ago should be comparable to the same grade in an examination a year later within an awarding body; or the standard of a given grade between two administrations (e.g. in different time zones) from the same examination session within an awarding body.

Rank-ordering is one of many comparability methodologies, and has been used relatively effectively to compare standards quantitatively between two exam sessions at component level[2] (Bramley, 2005; Bramley, 2007). Such a method has been modified to measure the equivalence of standards at syllabus level, based on examiners' holistic evaluation of scripts from prescribed components for each syllabus. Several studies (Yim, Shaw & Lewis, 2008; Yim & Shaw, 2009) have been conducted to demonstrate its feasibility and capability. The method has been used in both inter-board (Yim & Forster, 2010) and intra-board (Yim, 2012) studies pertaining to Cambridge International Examinations' (CIE) time zone question papers administered within the same exam session. Results so far have shown that the rank-ordering method could, to a large extent, produce comparable results when conducted repeatedly[3]. The qualitative feedback from questionnaire responses, on the other hand, revealed that some expert judges lacked confidence in their final rank-order judgements because the method's large cognitive demand requires them to retain the script information from several candidates holistically before making rank-order decisions. Interestingly, the quantitative results supported their judgements (Yim, 2012).

This paper describes a variant comparability methodology which uses the rank-ordering method at component level to derive results at syllabus level for intra-board comparison. In other words, instead of judges holding several components' information about each candidate in their minds and making a holistic evaluation of individual candidates during comparison, judges only rank-order candidates' performance within each prescribed component. The final rank order at syllabus level of each judge is derived based on his/her component level's rank orders. This variant methodology could enhance judges' experience during the exercise, as well as generate quantitative evidence of comparison for each prescribed component in order to inform threshold adjustment at component level during grading, in addition to the syllabus level only evidence from the holistic approach. This piece of information should provide an improvement in terms of clarity for grading advice, compared to that for the syllabus level only methodology.

The rationale behind conducting research at syllabus level is that quantitative results can generally help inform CIE's grading decisions in terms of threshold adjustment of an entire option/syllabus. The materials used in this study were question papers, mark schemes and syllabus specifications. Real candidates' component scripts with the same scheme of assessment and subject content from the same examination session within the same examination board were used in this study. These were then evaluated by external consultants (or judges) to generate rankings of candidates' scripts for each component. The rank-order data for each component were analysed using the multifacet Rasch modelling technique (Linacre, 1987). The outputs (or 'measures') from each component were combined by a weighted average method to generate the overall measure at syllabus level. The difference in standards between candidates' scripts at component as well as syllabus levels was deduced from the graphs. The methodology, the research outcome, and judges' feedback are described in detail below.

Background to comparability exercises

In this context, comparability is concerned with the application of the same standard across different examinations (Newton, 2007). The purpose of inter-board comparability studies is to compare standards across different examination boards. In making this comparison, it is important to distinguish between *content standards* and *performance standards*: "Content standards refer to the curriculum (or syllabus/ specification) and what examinees are expected to know and to be able to do … performance standards communicate how well examinees are expected to perform in relation to the content standards" (Hambleton, 2001). In fact, a more precise definition of comparability is paramount since many different aspects of qualifications can be compared, such as the demand of the curriculum, similarity of content materials,

[1] The Office of Qualifications and Examinations Regulation, England.

[2] In CIE, an exam *syllabus* usually comprises several ***components*** which assess different areas of skills/competencies in order to cover the subject knowledge to be assessed. For example: *Component 1:* Algebra, *Component 2:* Calculus, and so on, in a Maths *syllabus*. A component level comparability means, say, only comparing *Component 1s* between 2010 and 2013 exam sessions.

[3] The same set of scripts from prescribed components within each concerned syllabus was used in two separate research studies, i.e. inter- and intra-board comparisons.

difficulty experienced by candidates, demand of assessment materials, perceived quality of candidate outcome based on scripts and standards of attainment, etc.

One way to compare performance standards across assessments from different boards (or across parallel assessments from the same board) is to ask experts to compare pairs of scripts from each assessment and make judgements about which one demonstrates better quality. Such exercises address the question: "Which syllabuses' grade boundary scripts[4] are perceived by expert judges to be of better quality (after allowing for slight differences in syllabus content, question paper and mark scheme difficulty)?"

One way of analysing the data from these paired comparison judgements is by Thurstone's model (case 5) for comparative judgements (Thurstone, 1927). For a discussion of how Thurstone's method has been applied in the context of examination comparability, see Bramley (2007). For recent applications of the method see Yim, Shaw and Lewis (2008), and Yim and Shaw (2009).

The main advantage of this approach is that the use of candidates' scripts provides explicit evidence of the knowledge, understanding and skills of examinees. As such, direct comparison of performance standards can be achieved. For inter-board comparisons it should be noted that it is only possible to compare performance standards if the content standards across the examination boards are similar enough for the different assessments to be considered to be measuring the same construct (underlying trait). If the question papers, mark schemes and syllabus specifications are very different, examiners will be expected to make judgements about the relative performance standards in a context of possible differences in content standards. The outcome of such an exercise would be rendered less reliable due to disparate schemes of assessment and syllabus contents.

In practice, the nature of the scripts (objects) being compared is such that the scripts take a long time to read, and paired comparisons are unlikely to be independent, because of the repeated use of shared scripts. Hence examiners might already have the knowledge of either or both of the scripts before the paired comparisons, which violates the assumption of local independence between paired judgements. Therefore instead of asking judges to make paired comparisons, it is less time-consuming to ask them to put sets of scripts into rank-order of perceived quality. It is then possible to extract paired comparison data from the rank-order in the form of '1 beats 2', '2 beats 3', '1 beats 3' and so on (Bramley, 2007). These extracted paired comparisons are not statistically independent, because they are constrained by the ranking, but as explained above even genuine paired judgements would arguably not be independent either. In other words, a rank-ordering method is a time-saving variant of the paired comparison method for comparing performance standards. Such comparison exercises draw heavily on the expertise of senior examiners, and their ability to judge the quality of examinees' work, taking into account the demand placed upon examinees by the individual syllabuses/specifications, question papers and mark schemes.

Method

This study rank-ordered each prescribed intra-board component individually at component level using the same procedures as Yim and Forster (2010) with respect to the algorithm for selecting real candidates, the pack design, the instructions given to expert judges, and the data analysis method. Each judge's rank-orders for each prescribed component were then fed into the FACETS software (Linacre, 1987) to generate the outcome for the multifacet Rasch analysis, which would then be presented in graphical form for standards' comparison at component level. The outcome (or 'measure') of each component was then standardised by linear scaling such that they could be combined with each other in association with the weighting factor of each component specified in the syllabus specification, i.e. standardised weighted average, to yield the measure at syllabus level. The advantage of this approach is that the amount of script information that judges hold cognitively before making a rank-order decision is reduced, which is likely to improve on the accuracy of the rank-order results, enhance judges' ranking experience, and help boost their confidence in the exercise. Furthermore, the weighting factor of each component is applied during the generation of the weighted average at syllabus level. This is in contrast to the method used in a holistic evaluation approach, in which the application of a weighting factor could be less rigorous. Quantitative results in terms of differences in standards at component level can be generated in addition to those at syllabus level to inform grade boundary adjustment during awarding meetings[5] if there is a need to align standards with another assessment option (or exam board in the case of inter-board comparison).

The materials used in this project were question papers, mark schemes, syllabus specification and real candidates' scripts from the examination board. The first assessment is referred to as 'Option AA' and the second as 'Option BB' in this article. Each option has the same three components, namely, multiple choice (Component 1), structured questions (Component 2) and analysis and critical evaluation (Component 3). Thirty-four (or 17 from each assessment option) exact 'flat' profiles of real candidates' scripts at grade boundaries, A, B, C, D and E, and their intermediate grade boundaries at 2/3 and 1/3 of a grade above each grade, and 1/3 and 2/3 of a grade below each grade for both assessments were selected. A candidate with an exact 'flat' profile on a three-component assessment could be a candidate who achieves a mark exactly at the grade boundary of, say, B[6] at syllabus level with all three components also being at a mark exactly at the grade boundary of B; a candidate with an uneven profile could achieve a mark at the grade boundary of B at syllabus level, but with uneven grades at component level, for example, a mark at well above grade A in Component 1, a mark at the boundary of grade B in Component 2 and a mark at the middle of grade C in Component 3. The latter is more common/authentic in examination practice. The use of the exact 'flat' profile is to indicate to judges that a clear-cut standard across component level, for example, all components at the boundary of grade B, will lead to the same syllabus grade level, that is, grade B.

As a result of using the exact 'flat' candidate profile, real candidates whose script components' marks fit within ±1% of each targeted component mark at particular syllabus marks/grade levels were selected. It should be noted that the selection of real candidates' scripts meeting this criterion can only work well in an examination with a large entry, because there are enough scripts to choose from.

[4] Grade boundary scripts are scripts whose marks are exactly at the grade boundaries which were set during a grading (or an awarding) meeting.

[5] At awarding meetings the grade boundary locations on the raw mark scale of each component are decided.

[6] There is a subtle difference between a candidate with an exact even (or 'flat') profile and one with an even profile in this discussion. The criteria of the former are a candidate with the targeted component marks at exactly the same point relative to the grade boundary; whereas the latter only requires the same grades across prescribed components (e.g. BBB) within a syllabus and no stipulation of any targeted component marks.

After selecting the real candidates' scripts, examiner markings/annotations were removed electronically via a scanner so that they did not have an influence on the rank-ordering judgements during the experts' judging process. Each candidate was then allocated into different packs of scripts in accordance with the pack design at component level. An example of the pack design layout for component 1 is illustrated in *Appendix A* for readers' reference; other components follow the same pack design layout. Each pack comprised six candidates (three from Option AA and three from Option BB). Altogether there were eight packs (A to H) for component 1; eight packs (J to Q) for component 2; and eight packs (R to Y) for component 3. The candidates and hence their scripts in each pack were randomised, coded and labelled such that the original scripts' rank-order based on marks was concealed.

The same pack design was used for each component, i.e. the same set of candidates appeared in packs A, J and R, etc. Each candidate's scripts were photocopied for each expert judge.

In each pack of six scripts for each component, two were common to the pack above and two were common to the pack below (where 'above' and 'below' refer to the rank order by total mark). The top pack had two scripts in common with the pack below and the bottom pack had two scripts in common with the pack above. This linked design allowed a common scale of 'perceived quality' to be created from the ranking judgements.

Five senior examiners (expert judges), all with marking/moderating experience of the syllabus concerned, were recruited to make judgements about the real candidates' scripts. Their task was to rank-order scripts within each pack from best (highest quality = 1) to worst (lowest quality = 6) on each component and record their outcomes in the tables provided on a record sheet. Each expert judge was asked to complete a questionnaire towards the end of the exercise for the qualitative analysis of the study.

Analysis and results

Once the rank-order data at component level were received from judges, data for each component were deconstructed into paired comparison data and then analysed using the Rasch analysis (FACETS) software to estimate the difficulty/ability of each script/candidate for each component based on the inter-relationship of examiners' rankings. It should be noted that the percentage mark at component level, instead of a raw mark, was used in the analysis in order to achieve a common scale for both Options. The FACETS outputs are given in Appendices B, C and D for component 11 vs. 12, component 21 vs. 22 and component 31 vs. 32 respectively. The separation reliability index (analogous to Cronbach's Alpha) was high in all three cases, i.e. 0.99, 0.98 and 0.97 from Appendices B, C and D respectively, showing that the variability in perceived quality among the scripts could not be attributed to chance. There are different views on what fit index is actually acceptable; McNamara (1996) suggests that the usual limits of acceptability are the mean ±0.3 (so anything between 0.7 and 1.3 will be acceptable). According to Lunz and Wright (1997:83) "Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistics acceptance or rejection." They go on to use a level of ±0.5 in their studies. Operational experience, however, would suggest lower and upper bound limits of 0.7 and 1.6 respectively for mean squares to be useful and acceptable for practical purposes; and these were used in this analysis. Fit statistics of 1.7 or greater indicate too much unpredictability in examiners' scores, while fit statistics of 0.6 or less indicate over-fit or not enough variation in examiners' scores. The fit statistics from the infit and outfit columns of the FACET outputs for scripts and judges showed a slight tendency towards over-fit in all three cases suggesting that the judges were perceiving the trait in the same way and that there was less variability in their judgements than modelled. All these scale statistics need to be treated with caution because the paired comparison analysis violates the assumption of local independence between paired judgements when derived from the rank-ordering outcome (Bramley, 2012).

The Measure column in the bottom table of each component in Appendices B, C and D indicates the ability of each candidate's script. After taking the mean and standard deviation of the Measure column of each component and standardising them to the same mean and standard deviation, the total standardised weighted average Measure at syllabus level could be obtained. This was done by combining the respective Measure of candidates' scripts of each component with the weighting factor of each component designated in the syllabus specification. The results/graph of the total standardised weighted average Measure obtained using the component-derived-syllabus approach can then be re-scaled to compare with those evaluated by the holistic approach in 2012 (Yim, 2012), i.e. Figure 5.

Figures 1, 2, 3 and 4 show the results of the comparability plots for the three prescribed components and that for the component-derived syllabus approach respectively. The vertical axis along the left of the figures represents the Measure (or script quality) scale in logodds units (logits). In these graphs each data point (diamond – Option AA and square – Option BB) represents a script. Each script (a data point) is positioned according to its measure. Thus performances are rank ordered with the most able candidates at the top of the axis and the least able at the bottom, that is, the scripts in the top half of the graph (above 0 logits) are judged to be of better quality than those in the bottom half (below 0 logits). The horizontal axis shows the component/overall syllabus aggregate percentage mark obtained from conventional marking of the scripts.

The two straight lines in each comparability plot shown in Figures 1, 2, 3 and 4 are linear regression lines whose equations are given in the boxes. It should be noted that the legend 'linear' in the graphs refers to the regression line, and not to a linear exam (as opposed to a modular exam). The parameter *R* is the correlation coefficient. The magnitude of *R* indicates the extent to which the two sets of measurements (Measure and Syllabus %) are linearly related. Each pair of regression lines, that is, Options AA and BB, in the four cases shares similar features such as strong correlation and similar gradient. Figures 1, 2 and 3 provide the comparison between Options AA and BB at different grade boundaries in each component; whereas Figure 4 gives an overall syllabus aggregation with the weighting factor of each component taken into consideration.

Yim (2012) used the same set of scripts and judges as the current study, but used a holistic evaluation approach at the syllabus level. By comparing the results from Yim (2012), a comparison of results from the holistic evaluation approach at syllabus level and those from the current study can be made. Figure 5 shows an intra-board comparability plot using a holistic evaluation approach (Yim, 2012). The pair of regression lines in Figures 4 and 5 shows some similar features: strong correlation, similar gradient, no reversal of position; that is, option AA regression line is consistently above Option BB. Tables 1 and 2 show a comparison of some numerical findings between the holistic approach and the component-derived-syllabus approach.

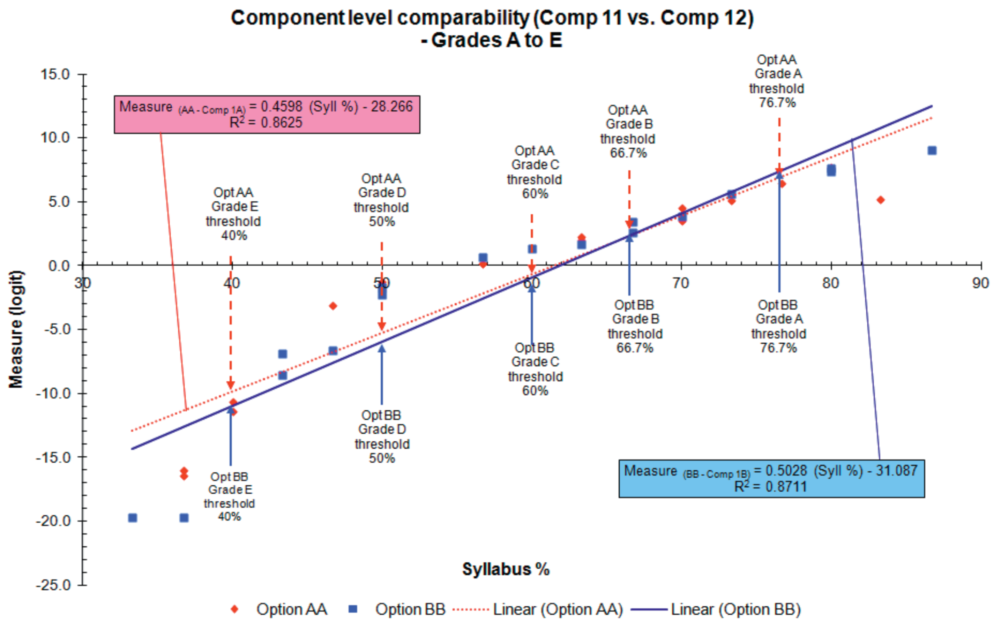

Figure 1: A comparability plot for Component 11 vs. Component 12 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

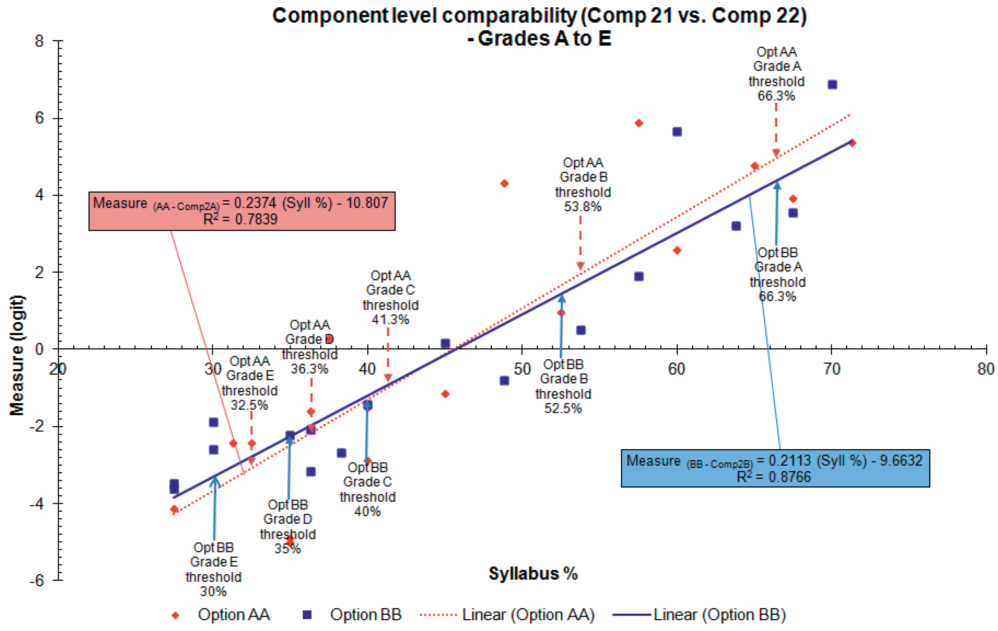

Figure 2: A comparability plot for Component 21 vs. Component 22 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

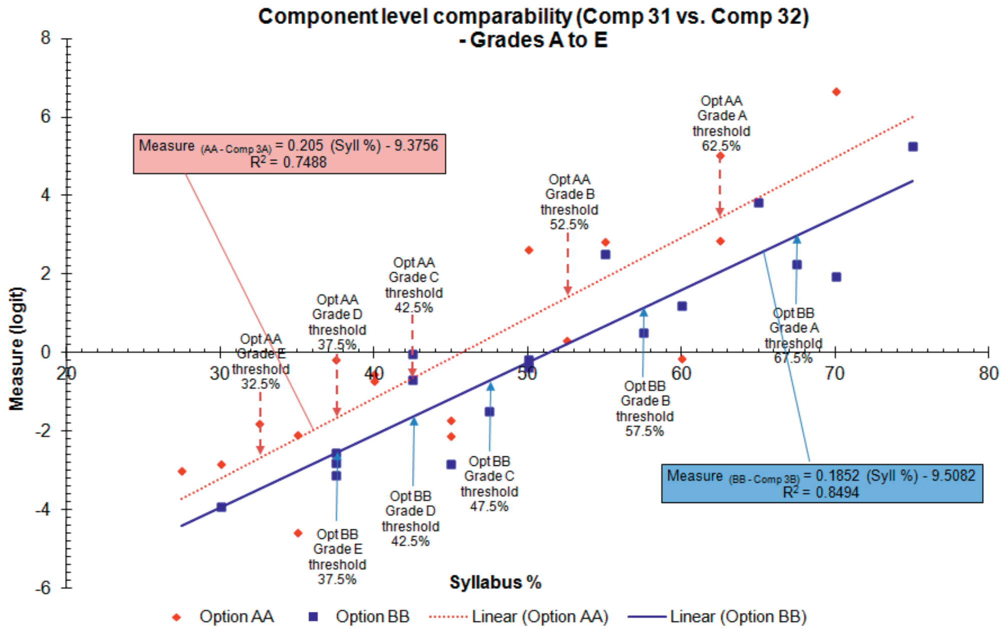

Figure 3: A comparability plot for Component 31 vs. Component 32 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

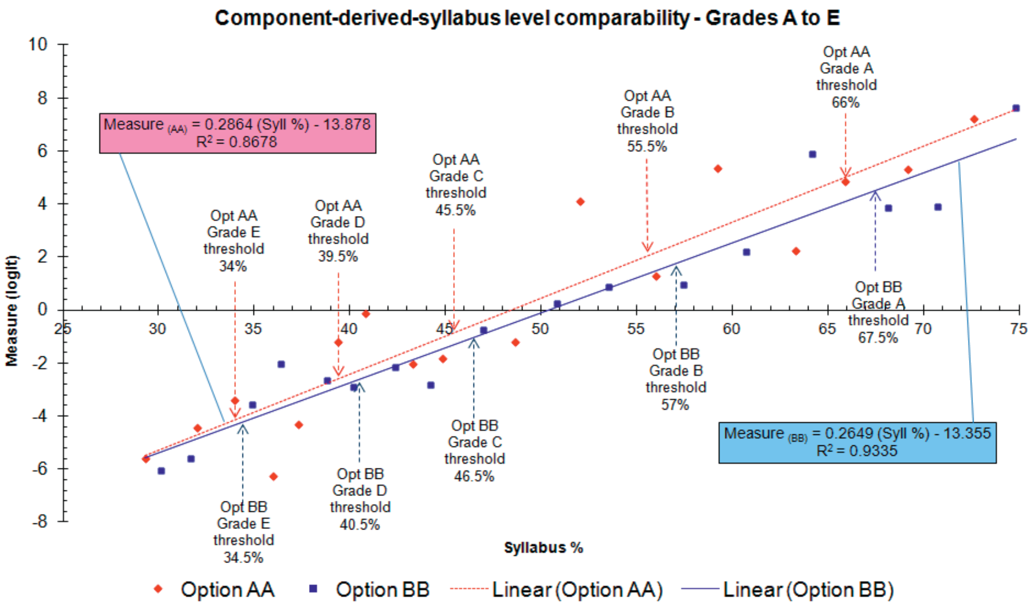

Figure 4: A comparability plot at syllabus level from grades A to E for the component-derived-syllabus approach between Options AA and BB

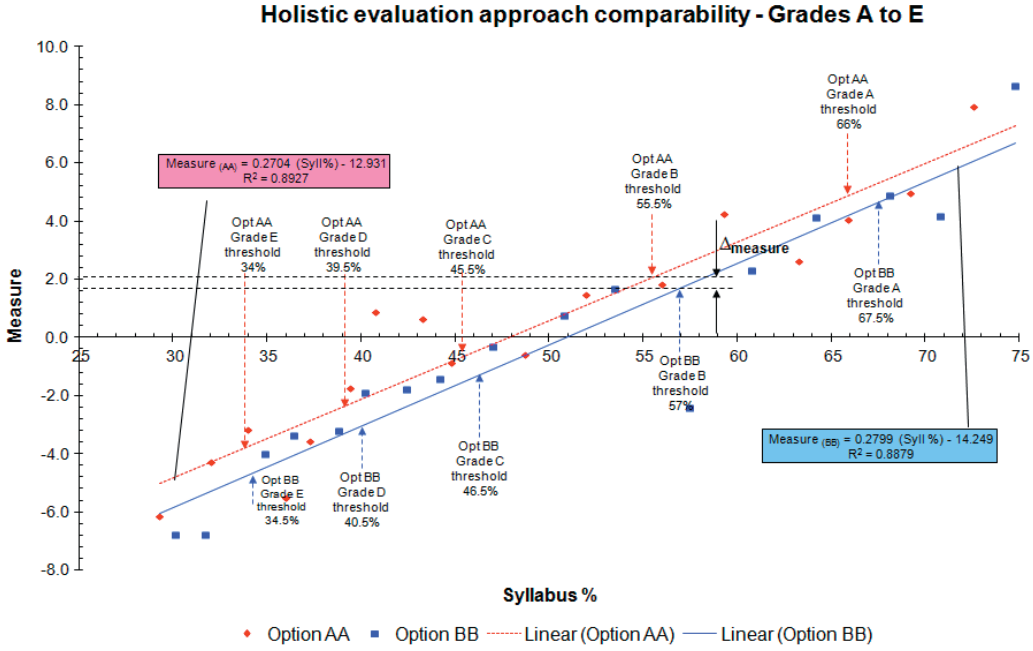

Figure 5: A comparability plot at syllabus level from grades A to E for the holistic evaluation approach between Options AA and BB (Yim, 2012).

Table 1 shows the differences in Measure (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both holistic evaluation and component-derived-syllabus approaches. In an ideal case the values of Δ*measure*, as shown in Figure 5, in both holistic evaluation and component-derived-syllabus approaches should be the same, but the differences in Table 1 suggest that there are disparities at all grades, albeit small, i.e. below or well below one logit. In other words, the recommendations for grade boundary adjustments at syllabus level to achieve the equivalence of standards between options are different depending on the methodology being used, which is understandable. The small differences between the two approaches at each grade are, in fact, rather encouraging as they demonstrate that the rank-ordering method could, to a certain extent, produce similar results when conducting two rank-ordering approaches.

Table 2 shows a comparison of the correlation coefficient (*R*) between 'Measure' and 'Syllabus %' for the holistic evaluation and component-

Table 1: Differences in '*Measure*' (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both holistic evaluation and component-derived-syllabus approaches.

Methodology	Δ*measure [logit]*				
	A	*B*	*C*	*D*	*E*
Holistic evaluation	0.27	0.37	0.61	0.66	0.86
Component-derived-syllabus	0.498	0.273	0.190	0.061	0.076

Table 2: A comparison of the correlation coefficient R between the holistic evaluation and component-derived-syllabus approaches.

Option	*Methodology*	*Correlation coefficient (R)*
AA	Component-derived-syllabus	0.93
	Holistic evaluation	0.94
BB	Component-derived-syllabus	0.97
	Holistic evaluation	0.94

derived-syllabus approaches in Options AA and BB. The correlations in both cases were very similar within the same assessment and across assessments. The strong correlations ($R \geq 0.93$) in all cases between the 'Measure' and the 'Syllabus %' show that the trait of quality as perceived by the judges was very similar to the trait of quality as rewarded by the mark scheme. It should be recalled that the only difference in terms of the research design between the previous comparability study and the current one was that in this study, the methodology of component-derived-syllabus approach was used rather than the holistic approach. Both assessments were from the same syllabus from the same examination board and assessed by the same group of judges.

Feedback from examiners

Responses on questionnaires were collected from five judges who carried out the evaluation to help understand the qualitative aspects of their rank-ordering experience relating to the overall difficulty of the task, the amount of time taken to rank order the scripts, difficulty compared with the holistic evaluation approach, what made some packs more or less difficult to rank, any differences in the task between papers, and the strategy they deployed.

Overall difficulty of the task

All five participants were senior examiners and had taken part in at least two rank-ordering exercises previously. Four of them found the overall task "fairly difficult" to execute; and one examiner found it "fairly easy". Reasons for difficulty are shown on the left-hand column in Table 3. Those from the previous holistic evaluation approach (Yim, 2012) are listed for reference. Judges tended to take an average of just under 30 minutes per pack during the evaluation as compared to between 40 and 90 minutes per pack in the holistic evaluation approach. It should be reminded that the amount of scripts between the holistic evaluation and component-derived-syllabus approaches were different, i.e. three components versus one respectively. The component-derived-syllabus approach probably took longer overall based on the number of packs being evaluated. Despite this, the judges were more confident about their rank-orders and there was generally no need to re-visit the design packs after the exercise, unlike the holistic evaluation approach. Three out of five examiners thought the length of time for the evaluation varied greatly from pack to pack when ranked by individual component.

Table 3: Overall difficulty of the task encountered by judges during the evaluation phase. Reasons from the holistic evaluation approach (Yim, 2012) are also included for reference.

Component-derived-syllabus approach	*Holistic evaluation approach*
Differences between questions in question papers from both options;	Differences between questions in question papers from both options;
Difficult to retain script information to make judgement on the rank-order;	Difficult to obtain an overview of papers with a number of parts;
Candidates' standards are very close within each pack.	Difficult to retain script information to make judgement on the rank-order;
	Candidates' standards are very close within each pack. (Yim, 2012)

Differences were also reported relating to the ease or difficulty of rank-ordering certain packs. Scripts from more able candidates were the most time-consuming to rank order although they were slightly less problematic as there was perceived to be a wider range of ability instantiated in performances. Scripts from less able candidates were more difficult to rank, and standards were perceived to be more closely grouped. Other factors included the mode of assessment. The MCQ[7] component was easier to rank compared to the written component.

All examiners concurred that the task of rank-ordering individual components was much easier compared with that of the holistic evaluation approach at syllabus level. Examiners articulated that they felt more confident at the end of the exercise with their rank order results when their focus was on the same assessment instrument rather than attempting to compare performance across a number of them. It was necessary to keep less script information 'in mind' in each pack and hence most of them were confident about their results.

All examiners felt that it was possible to carry out the judging for a pack of six candidates with one component paper, while three out of five examiners agreed that it was possible to carry out the judging for a pack of six candidates at syllabus level with three component papers holistically (Yim, 2012). It should be noted that examiners from both exercises managed to complete the research studies well, as suggested by the comparable analyses' results.

Rank-ordering strategy

Examiners were allowed to adopt their own rank-ordering strategy during the evaluation phase though they were not allowed to re-mark the scripts. A variety of strategies were identified as follows:

- Identification of common and indicative questions across question papers to evaluate candidates' ability.
- Identification of questions attempted by less able students: based on examiners' experience, some questions can act as an indicator to distinguish between able and less able candidates.
- Identification of the quality of answers given, e.g. correct terminology, accuracy of diagrams.

Overall judgement of depth and accuracy of answers

No examiner indicated a change of approach as the rank order task became increasingly more familiar. Three out of five examiners commented that they employed the same strategy as for the holistic evaluation approach exercise that they completed a year ago.

Examiners were uncertain as to whether more or less time on each script made any difference to the final rank order. However, in the main, they believed that a reduction or extension in the time taken to undertake the exercise would have little impact on the outcome.

Conclusions

A new component-derived-syllabus rank-ordering approach for intra-board comparability study has been reported in this paper. The aim of this approach is to enhance judges' experience and the quality of results from the evaluation exercise, and to generate quantitative evidence of comparison at component level for grading purposes, in addition to the usual practice of acquiring evidence only at syllabus level. The results showed that the component-derived-syllabus recommendations for grade boundary adjustments at syllabus level were close to the findings recommended by the holistic evaluation approach under the same

[7] By manually circling individual candidates' answers on the multiple choice question papers based on their answer strings (original m.c. responses), judges evaluated candidates' answers in relation to each question to rank their performances within each pack design.

boundary conditions (Yim, 2012). The small differences between the two approaches at each grade boundary are, in fact, rather encouraging as they, demonstrate comparable results even though different rank-ordering approaches were used.

In the current study the correlations between perceived quality and aggregate mark were very similar across the component-derived-syllabus approach and holistic evaluation approach. The implication of this finding is that the use of different rank-ordering approaches does not affect how the trait of quality is perceived. This contradicts the initial hypothesis that the application of a weighting factor to individual components could improve the correlation. The *prima facie* evidence of the current study suggests that there is no advantage in terms of using either type of approach in relation to the internal quality of the scale produced (separation reliability and fit), or its correlation with an external variable (aggregate Syllabus % mark). The qualitative feedback from all expert judges suggests that the component-derived-syllabus approach was made much easier by rank-ordering scripts by component rather than by the holistic evaluation approach. They felt confident in carrying out the tasks as well as their rank-order judgements.

Limitations of the study

If judges see the same two scripts in a consecutive design pack, there may be a memory effect which could affect the rank-order results during the evaluation of each component. In the light of this, an evaluation procedure of scrutinising alternate design packs was used. In other words, judges were strongly recommended to evaluate design packs according to the designated sequence: A→C→E→G→B→D→F→H in the instructions; and they were also reminded to complete the evaluation of a design pack fully before moving on the next one. Although this should, to a large extent, minimise the impact of the memory effect, it could not totally eliminate the possibility of memory effects. Since the same full set of scripts was presented to the same judges a year ago (but with different design pack arrangements) and the previous study used the holistic evaluation approach, there is a small chance that some judges could have remembered certain scripts. However, the rank-order data required this time was very different from that in the earlier study, so the impact should therefore be minimal.

Acknowledgements

The author would like to thank Tom Bramley for the discussion of the component-derived-syllabus level comparability methodology.

References

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement, 6*(2), 202–223.

Bramley, T. (2007). Chapter 7. Paired comparison methods. In Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (Eds.), *Techniques for monitoring the comparability of examination standards.* (pp.246–294). London: Qualifications and Curriculum Authority.

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment Publication,13*, 18–26.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In: G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives*. (pp.89–116). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J.M. (2008). FACETS Rasch measurement computer program. Chicago: Winsteps.com.

Lunz, M. E., & Wright, B. D. (1997). Latent trait models for performance examinations. In Jürgen Rost & Rolf Langeheine (Eds), *Applications of latent trait and latent class models in the social sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Newton, P (2007). Chapter 1. Contextualising the comparability of examination standards. In Newton, P., Baird, J., Goldstein, G., Patrick, H & Tymms, P (Eds.), *Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority.

Thurston, L.L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286. Chapter 3 in L.L. Thurston (1959), *The measure of values*. Chicago, Illinois: University of Chicago Press.

Yim, L.W.K., Shaw, S.D. & Lewis, M (2008). A science comparability study between two exam boards using a rank-ordering methodology at syllabus level. *9th AEA Europe Conference Proceeding, Hisar, Bulgaria*, 6–8 Nov 2008.

Yim, L.W.K and Shaw, D. S. (2009). A comparability study using a rank-ordering methodology at syllabus level between examination boards. *35th IAEA Annual Conference Proceedings, Brisbane, Australia,* 13–18 September 2009.

Yim, L.W.K and Forster, M. (2010). A comparison between the effect of using pseudo candidates' scripts and real candidates' scripts in a rank-ordering comparability methodology at syllabus level. *36th IAEA Annual Conference Proceedings (2010) – Assessment for the future generations, Bangkok, Thailand,* 22–27 August 2010.

Yim, L.W.K. (2012). An Intra-board comparison of the effect of using pseudo candidates' scripts and real candidates' scripts in a rank-ordering exercise at syllabus level. *Research Matters: A Cambridge Assessment Publication*, *14*, 2–9.

Appendix A – A pack design layout for Component 1. Components 2 and 3 follow the same pack design.

X = option AA
Y = option BB

Real-candidate - Component 1 comparison

		Grade level		E-2/3		E-1/3		E		E+1/3		E+2/3 or D-1/3		D		C-2/3		C-1/3		C		C+1/3		C+2/3 or B-1/3		B		A-2/3		A-1/3		A		A+1/3		A+2/3	
Designed by Lewis Kim		syll %		36.7	36.7	33.3	36.7	43.3	40.0	43.3	40.0	50.0	46.7	46.7	50.0	50.0	50.0	56.7	56.7	60.00	56.67	66.67	63.33	63.33	63.33	66.7	70	70	70	73.3	76.7	80	73.3	80	83.3	86.7	80
		Code		H5	H4	H6	H3	H1	H2			F4	F5	F6	F2	F3	F1			D5	D6	D1	D3	D4	D2			B6	B5	B4	B2	B1	B3				
Pack Range	Pack	Judge						G4	G6	G3	G5	G1	G2			E6	E5	E4	E3	E1	E2			C6	C5	C4	C2	C3	C1			A5	A6	A4	A2	A1	A3
A+2/3, A+1/3, A	A	1																														Y	X	Y	X	Y	X
	A	2																														Y	X	Y	X	Y	X
	A	3																														Y	X	Y	X	Y	X
	A	4																														Y	X	Y	X	Y	X
	A	5																														Y	X	Y	X	Y	X
A, A-1/3, A-2/3	B	1																										Y	X	Y	X	Y	X				
	B	2																										Y	X	Y	X	Y	X				
	B	3																										Y	X	Y	X	Y	X				
	B	4																										Y	X	Y	X	Y	X				
	B	5																										Y	X	Y	X	Y	X				
A-2/3, B, C+2/3 or B-1/3	C	1																						Y	X	Y	X	Y	X								
	C	2																						Y	X	Y	X	Y	X								
	C	3																						Y	X	Y	X	Y	X								
	C	4																						Y	X	Y	X	Y	X								
	C	5																						Y	X	Y	X	Y	X								
C+2/3 or B-1/3, C+1/3, C	D	1																		Y	X	Y	X	Y	X												
	D	2																		Y	X	Y	X	Y	X												
	D	3																		Y	X	Y	X	Y	X												
	D	4																		Y	X	Y	X	Y	X												
	D	5																		Y	X	Y	X	Y	X												
C, C-1/3, C-2/3	E	1														Y	X	Y	X	Y	X																
	E	2														Y	X	Y	X	Y	X																
	E	3														Y	X	Y	X	Y	X																
	E	4														Y	X	Y	X	Y	X																
	E	5														Y	X	Y	X	Y	X																
C-2/3, D, E+2/3 or D-1/3	F	1										Y	X	Y	X	Y	X																				
	F	2										Y	X	Y	X	Y	X																				
	F	3										Y	X	Y	X	Y	X																				
	F	4										Y	X	Y	X	Y	X																				
	F	5										Y	X	Y	X	Y	X																				
E+2/3 or D-1/3, E+1/3, E	G	1						Y	X	Y	X	Y	X																								
	G	2						Y	X	Y	X	Y	X																								
	G	3						Y	X	Y	X	Y	X																								
	G	4						Y	X	Y	X	Y	X																								
	G	5						Y	X	Y	X	Y	X																								
E, E-1/3, E-2/3	H	1		Y	X	Y	X	Y	X																												
	H	2		Y	X	Y	X	Y	X																												
	H	3		Y	X	Y	X	Y	X																												
	H	4		Y	X	Y	X	Y	X																												
	H	5		Y	X	Y	X	Y	X																												
No of copies needed			X		5		5		10		5		10		5		5		5		10		5		10		5		10		5		10		5		5
No of copies needed			Y	5		5		10		5		10		5		5		5		10		5		10		5		10		5		10		5		5	

Appendix B – FACETS output

Component 11 vs. Component 12

Table 7.1.1 Judge Measurement Report (arranged by mN)

```
+---------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                     |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Judge             |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|    60     120        .5    .50|    .00   .26 | 1.74  4.7  1.80  1.0| -.11 |   .50   .62 | 1 CHS               |
|    60     120        .5    .50|    .00   .26 |  .67 -2.8   .43   .4| 1.51 |   .68   .62 | 2 TC                |
|    60     120        .5    .50|    .00   .26 |  .87  -.9  1.08   .7| 1.13 |   .63   .62 | 3 PC                |
|    60     120        .5    .50|    .00   .26 |  .96  -.3   .76   .6| 1.09 |   .63   .62 | 4 NB                |
|    60     120        .5    .50|    .00   .26 |  .68 -2.7   .44   .4| 1.49 |   .67   .62 | 5 GM                |
|-------------------------------+--------------+---------------------+------+-------------+---------------------|
|    60.0   120.0       .5    .50|    .00   .26 |  .98  -.4   .90   .6|      |   .62       | Mean (Count: 5)     |
|      .0      .0       .0    .00|    .00   .00 |  .40  2.8   .51   .2|      |   .07       | S.D. (Population)   |
|      .0      .0       .0    .00|    .00   .00 |  .44  3.1   .57   .3|      |   .07       | S.D. (Sample)       |
+---------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .26  Adj (True) S.D. .00  Separation .00  Reliability 1.00
Model, Sample: RMSE .26  Adj (True) S.D. .00  Separation .00  Reliability .80
Model, Fixed (all same) chi-square: .0  d.f.: 4  significance (probability): 1.00
-----------------------------------------------------------------------------------------------------------------
```

Table 7.3.1 Script Measurement Report (arranged by mN)

```
+------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                        |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script              |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|    12.5    25        .5   1.00|   9.04   .64 | 1.07   .3  2.31  1.3|  .80 |   .71   .76 | 18 t2_A1 (mark 86.7)   |
|    12.5    25        .5   1.00|   7.73   .50 | 1.42  1.6  1.86  1.6|  .23 |   .39   .60 |  2 t1_A3 (mark 80)     |
|    25      50        .5   1.00|   7.61   .41 |  .68 -1.5   .53  -.9| 1.34 |   .79   .71 | 20 t2_A5B1 (mark 80)   |
|    12.5    25        .5   1.00|   7.35   .50 |  .89  -.3   .84  -.3| 1.18 |   .66   .60 | 19 t2_A4 (mark 80)     |
|    12.5    25        .5   1.00|   6.44   .55 |  .67 -1.0   .81  -.2| 1.30 |   .77   .67 |  4 t1_B2 (mark 76.7)   |
|    12.5    25        .5   1.00|   5.62   .51 |  .71 -1.0   .63 -1.0| 1.39 |   .76   .62 | 21 t2_B4 (mark 73.3)   |
|    12.5    25        .5    .99|   5.19   .63 | 1.59  1.4  2.95  1.9|  .36 |   .60   .76 |  1 t1_A2 (mark 83.3)   |
|    25      50        .5    .99|   5.10   .40 |  .94  -.2  1.01   .1| 1.04 |   .70   .69 |  3 t1_A6B3 (mark 73.3) |
|    12.5    25        .5    .99|   4.51   .52 | 1.07   .3  1.28   .7|  .85 |   .59   .64 |  6 t1_C2 (mark 70)     |
|    25      50        .5    .98|   3.88   .36 | 1.03   .2  1.07   .2|  .93 |   .59   .61 | 22 t2_B6C3 (mark 70)   |
|    25      50        .5    .97|   3.49   .37 |  .99   .0   .81  -.3| 1.05 |   .63   .62 |  5 t1_B5C1 (mark 70)   |
|    12.5    25        .5    .97|   3.44   .46 |  .89  -.5   .87  -.4| 1.25 |   .57   .48 | 23 t2_C4 (mark 66.7)   |
|    12.5    25        .5    .93|   2.55   .46 | 1.01   .1   .91  -.1| 1.03 |   .50   .49 | 25 t2_D1 (mark 66.7)   |
|    12.5    25        .5    .90|   2.23   .44 |  .90  -.6   .81  -.7| 1.39 |   .52   .41 |  8 t1_D3 (mark 63.3)   |
|    25      50        .5    .88|   1.97   .33 |  .84 -1.1   .76 -1.0| 1.34 |   .63   .53 |  7 t1_C5D2 (mark 63.3) |
|    25      50        .5    .85|   1.70   .34 | 1.11   .7  1.31  1.1|  .75 |   .48   .55 | 24 t2_C6D4 (mark 63.3) |
|    25      50        .5    .80|   1.37   .34 | 1.16  1.0  1.18   .6|  .70 |   .47   .55 | 26 t2_D5E1 (mark 60)   |
|    12.5    25        .5    .66|    .66   .49 | 1.05   .2   .93   .0|  .95 |   .55   .56 | 27 t2_E4 (mark 56.7)   |
|    25      50        .5    .54|    .17   .36 |  .82  -.9   .69 -1.1| 1.26 |   .72   .62 |  9 t1_D6E2 (mark 56.7) |
|    12.5    25        .5    .53|    .13   .47 |  .98   .0   .89  -.2| 1.07 |   .55   .53 | 10 t1_E3 (mark 56.7)   |
|    25      50        .5    .22|  -1.24   .37 | 1.19   .9  1.33  1.2|  .70 |   .54   .60 | 11 t1_E5F1 (mark 50)   |
|    25      50        .5    .16|  -1.68   .44 | 1.02   .1   .57  2.5|  .98 |   .71   .71 | 29 t2_F4G1 (mark 50)   |
|    12.5    25        .5    .14|  -1.85   .48 | 1.00   .0  1.00   .0| 1.00 |   .52   .51 | 12 t1_F2 (mark 50)     |
|    25      50        .5    .09|  -2.29   .40 |  .85  -.7   .59  -.9| 1.26 |   .73   .68 | 28 t2_E6F3 (mark 50)   |
|    25      50        .5    .04|  -3.14   .49 |  .99   .0   .59  1.1| 1.03 |   .75   .74 | 13 t1_F5G2 (mark 46.7) |
|    25      50        .5    .00|  -6.92   .96 |  .74  -.1   .16  2.6| 1.16 |   .86   .85 | 32 t2_G4H1 (mark 43.3) |
|    12.5    25        .5    .00|  -8.56   .83 | 1.15   .4   .61   .8|  .95 |   .81   .82 | 31 t2_G3 (mark 43.3)   |
|    25      50        .5    .00| -10.69   .73 | 1.12   .4   .37  1.7|  .98 |   .87   .87 | 15 t1_G6H2 (mark 40)   |
|    12.5    25        .5    .00| -11.38   .83 |  .86  -.1   .23  4.1| 1.21 |   .82   .81 | 14 t1_G5 (mark 40)     |
|    12.5    25        .5    .00| -16.01   .90 |  .98   .0   .33  4.6| 1.15 |   .83   .82 | 16 t1_H3 (mark 36.7)   |
|    12.5    25        .5    .00| -16.42   .91 |  .99   .0   .33  5.7| 1.13 |   .81   .79 | 17 t1_H4 (mark 36.7)   |
|    12.5    25        .5    .00|( -6.67  1.84)|Minimum              |      |   .00   .00 | 30 t2_F6 (mark 46.7)   |
|    12.5    25        .5    .00|(-19.70  1.85)|Minimum              |      |   .00   .00 | 33 t2_H5 (mark 36.7)   |
|    12.5    25        .5    .00|(-19.70  1.85)|Minimum              |      |   .00   .00 | 34 t2_H6 (mark 33.3)   |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                        |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script              |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|    17.6    35.3        .5    .55|  -1.35   .65 |  .99   .0   .92   .8|      |   .60       | Mean (Count: 34)       |
|     6.2    12.3        .0    .44|   7.92   .41 |  .19   .7   .58  1.7|      |   .22       | S.D. (Population)      |
|     6.2    12.5        .0    .44|   8.04   .42 |  .20   .7   .59  1.7|      |   .23       | S.D. (Sample)          |
+------------------------------------------------------------------------------------------------------------------+
   With extremes, Model, Populn: RMSE .77  Adj (True) S.D. 7.88  Separation 10.30  Reliability .99
   With extremes, Model, Sample: RMSE .77  Adj (True) S.D. 8.00  Separation 10.45  Reliability .99
Without extremes, Model, Populn: RMSE .56  Adj (True) S.D. 6.63  Separation 11.85  Reliability .99
Without extremes, Model, Sample: RMSE .56  Adj (True) S.D. 6.74  Separation 12.05  Reliability .99
With extremes, Model, Fixed (all same) chi-square: 3201.1  d.f.: 33  significance (probability): .00
With extremes, Model,  Random (normal) chi-square: 32.4  d.f.: 32  significance (probability): .44
--------------------------------------------------------------------------------------------------------------------
```

Appendix C – FACETS output

Component 21 vs. Component 22

Table 7.1.1 Judge Measurement Report (arranged by mN)

```
+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |                        |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Judge                |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|    60     120         .5    .50|    .00   .24 |  .82 -1.5   .58   -.8| 1.30 |   .69   .63 | 1 CHS                  |
|    60     120         .5    .50|    .00   .24 | 1.16  1.3  1.24    .6|  .72 |   .58   .63 | 2 TC                   |
|    60     120         .5    .50|    .00   .24 | 1.12  1.0   .93    .0|  .86 |   .60   .63 | 3 PC                   |
|    60     120         .5    .50|    .00   .24 | 1.02   .2  1.00    .1|  .94 |   .62   .63 | 4 NB                   |
|    60     120         .5    .50|    .00   .24 |  .87 -1.1   .61   -.7| 1.24 |   .68   .63 | 5 GM                   |
|-------------------------------+--------------+---------------------+------+-------------+------------------------|
|    60.0   120.0       .5    .50|    .00   .24 | 1.00   .0   .87   -.2|      |   .63       | Mean (Count: 5)        |
|      .0      .0       .0    .00|    .00   .00 |  .13  1.1   .25    .6|      |   .04       | S.D. (Population)      |
|      .0      .0       .0    .00|    .00   .00 |  .15  1.3   .28    .6|      |   .05       | S.D. (Sample)          |
+-------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .24  Adj (True) S.D. .00  Separation .00  Reliability 1.00
Model, Sample: RMSE .24  Adj (True) S.D. .00  Separation .00  Reliability .80
Model, Fixed (all same) chi-square: .0  d.f.: 4  significance (probability): 1.00
---------------------------------------------------------------------------------------------------------------------
```

Table 7.3.1 Script Measurement Report (arranged by mN)

```
+----------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |                           |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script                 |
|-------------------------------+--------------+---------------------+------+-------------+---------------------------|
|    12.5    25         .5   1.00|   6.87   .75 | 1.01   .2  1.03    .3|  .98 |   .83   .84 | 18 t2_J1 (mark 70)        |
|    25      50         .5   1.00|   5.90   .53 | 1.04   .2   .61    .2|  .99 |   .81   .81 |  4 t1_K1L3 (mark 57.5)    |
|    12.5    25         .5   1.00|   5.65   .58 |  .99   .0   .85    .0| 1.01 |   .72   .72 | 22 t2_K6 (mark 60)        |
|    12.5    25         .5   1.00|   5.37   .52 |  .99   .0   .93    .0| 1.03 |   .65   .63 |  3 t1_J5 (mark 71.3)      |
|    25      50         .5    .99|   4.78   .36 |  .83  -.8   .94   -.1| 1.19 |   .68   .62 |  2 t1_J3K3 (mark 65)      |
|    25      50         .5    .99|   4.32   .81 | 1.10   .3   .36    .0| 1.02 |   .92   .91 |  6 t1_L1M1 (mark 48.8)    |
|    12.5    25         .5    .98|   3.91   .48 | 1.28  1.2  1.68   1.5|  .35 |   .37   .54 |  1 t1_J2 (mark 67.5)      |
|    12.5    25         .5    .97|   3.55   .50 |  .90  -.3   .77   -.3| 1.20 |   .63   .58 | 19 t2_J4 (mark 67.5)      |
|    25      50         .5    .96|   3.20   .38 | 1.04   .2  1.10    .3|  .94 |   .64   .65 | 20 t2_J6K5 (mark 63.8)    |
|    12.5    25         .5    .93|   2.59   .57 |  .70 -1.0   .41   -.9| 1.38 |   .80   .70 |  5 t1_K4 (mark 60)        |
|    25      50         .5    .87|   1.90   .45 | 1.12   .5  1.01    .2|  .90 |   .74   .76 | 21 t2_K2L4 (mark 57.5)    |
|    12.5    25         .5    .73|    .97   .58 | 1.09   .4   .73    .1|  .94 |   .68   .69 |  7 t1_L5 (mark 52.5)      |
|    12.5    25         .5    .62|    .50   .58 |  .95   .0   .60    .1| 1.10 |   .71   .69 | 23 t2_L2 (mark 53.8)      |
|    25      50         .5    .57|    .29   .53 |  .98   .0   .74   -.1| 1.03 |   .83   .82 | 11 t1_N6O4 (mark 37.5)    |
|    12.5    25         .5    .54|    .16   .55 | 1.00   .0   .85    .0| 1.01 |   .66   .66 | 25 t2_M4 (mark 45)        |
|    25      50         .5    .31|   -.81   .40 |  .93  -.3   .61    .2| 1.15 |   .68   .66 | 24 t2_L6M6 (mark 48.8)    |
|    12.5    25         .5    .24|  -1.14   .47 |  .98  -.1   .78    .2| 1.12 |   .51   .50 |  9 t1_M3 (mark 45)        |
|    25      50         .5    .19|  -1.42   .33 |  .90  -.7   .80    .0| 1.26 |   .52   .48 | 26 t2_M5N5 (mark 40)      |
|    25      50         .5    .18|  -1.51   .33 | 1.18  1.3  1.13    .4|  .53 |   .41   .48 |  8 t1_M2N1 (mark 40)      |
|    12.5    25         .5    .17|  -1.59   .52 |  .80  -.7   .63   -.6| 1.31 |   .70   .62 | 12 t1_O1 (mark 36.3)      |
|    12.5    25         .5    .13|  -1.87   .51 |  .91  -.3   .63   -.5| 1.25 |   .66   .61 | 31 t2_P4 (mark 30)        |
|    25      50         .5    .11|  -2.09   .34 | 1.19  1.3  1.53   1.7|  .49 |   .40   .53 | 28 t2_N4O3 (mark 36.3)    |
|    25      50         .5    .10|  -2.21   .35 |  .88  -.8   .64  -1.1| 1.30 |   .65   .58 | 30 t2_O5P2 (mark 35)      |
|    12.5    25         .5    .08|  -2.42   .45 |  .95  -.2   .91   -.3| 1.15 |   .50   .45 | 16 t1_Q2 (mark 31.3)      |
|    25      50         .5    .08|  -2.42   .33 |  .97  -.2   .90   -.3| 1.09 |   .53   .50 | 15 t1_P3Q4 (mark 32.5)    |
|    25      50         .5    .07|  -2.59   .32 |  .99   .0  1.00    .0| 1.02 |   .49   .48 | 32 t2_P6Q5 (mark 30)      |
|    12.5    25         .5    .06|  -2.69   .48 |  .93  -.2   .78   -.4| 1.19 |   .59   .54 | 27 t2_N3 (mark 38.3)      |
|    12.5    25         .5    .05|  -2.88   .49 | 1.05   .2   .92    .0|  .95 |   .56   .58 | 10 t1_N2 (mark 40)        |
|    12.5    25         .5    .04|  -3.16   .54 | 1.45  1.3  1.29    .6|  .54 |   .52   .66 | 29 t2_O2 (mark 36.3)      |
|    12.5    25         .5    .03|  -3.48   .43 |  .86  -.8   .86   -.7| 1.44 |   .53   .39 | 33 t2_Q1 (mark 27.5)      |
|    12.5    25         .5    .03|  -3.63   .44 | 1.05   .3  1.06    .3|  .86 |   .37   .42 | 34 t2_Q3 (mark 27.5)      |
|    12.5    25         .5    .02|  -4.13   .48 | 1.05   .2  1.09    .3|  .91 |   .52   .55 | 17 t1_Q6 (mark 27.5)      |
|    12.5    25         .5    .01|  -4.91   .63 |  .85  -.2   .66   -.4| 1.15 |   .81   .77 | 14 t1_P1 (mark 35)        |
|    25      50         .5    .01|  -5.01   .52 | 1.03   .2   .78    .0| 1.01 |   .82   .82 | 13 t1_O6P5 (mark 35)      |
|-------------------------------+--------------+---------------------+------+-------------+---------------------------|
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |                           |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script                 |
|-------------------------------+--------------+---------------------+------+-------------+---------------------------|
|    17.6    35.3       .5    .44|    .00   .49 | 1.00   .0   .87    .0|      |   .63       | Mean (Count: 34)          |
|     6.2    12.3       .0    .40|   3.39   .11 |  .14   .6   .27    .6|      |   .14       | S.D. (Population)         |
|     6.2    12.5       .0    .41|   3.44   .11 |  .14   .6   .28    .6|      |   .14       | S.D. (Sample)             |
+----------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .50  Adj (True) S.D. 3.35  Separation 6.72  Reliability .98
Model, Sample: RMSE .50  Adj (True) S.D. 3.41  Separation 6.83  Reliability .98
Model, Fixed (all same) chi-square: 1567.3  d.f.: 33  significance (probability): .00
Model,  Random (normal) chi-square: 32.3  d.f.: 32  significance (probability): .45
```

Appendix D – FACETS output

Component 31 vs. Component 32

Table 7.1.1 Judge Measurement Report (arranged by mN)

```
+---------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |                |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Judge        |
|-------------------------------+--------------+---------------------+------+-------------+----------------|
|    60     120        .5    .50|   .00    .24 | 1.14  1.2  1.14   .5|  .79 |   .56   .61 | 1 CHS          |
|    60     120        .5    .50|   .00    .24 |  .88 -1.0   .84  -.4| 1.17 |   .65   .61 | 2 TC           |
|    60     120        .5    .50|   .00    .24 | 1.04   .4  1.00   .0|  .93 |   .59   .61 | 3 PC           |
|    60     120        .5    .50|   .00    .24 |  .86 -1.2   .72  -.9| 1.23 |   .67   .61 | 4 NB           |
|    60     120        .5    .50|   .00    .24 | 1.07   .6   .92  -.1|  .93 |   .60   .61 | 5 GM           |
|-------------------------------+--------------+---------------------+------+-------------+----------------|
|   60.0   120.0       .5    .50|   .00    .24 | 1.00   .0   .92  -.2|      |   .61       | Mean (Count: 5)  |
|     .0      .0       .0    .00|   .00    .00 |  .11  1.0   .14   .5|      |   .04       | S.D. (Population)|
|     .0      .0       .0    .00|   .00    .00 |  .12  1.1   .16   .6|      |   .05       | S.D. (Sample)    |
+---------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .24  Adj (True) S.D. .00  Separation .00  Reliability 1.00
Model, Sample: RMSE .24  Adj (True) S.D. .00  Separation .00  Reliability .80
Model, Fixed (all same) chi-square: .0  d.f.: 4  significance (probability): 1.00
-----------------------------------------------------------------------------------------------------------
```

Table 7.3.1 Script Measurement Report (arranged by mN)

```
+-----------------------------------------------------------------------------------------------------------
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script
|-------------------------------+--------------+---------------------+------+-------------+------------------
|   12.5    25         .5   1.00|  6.67    .77 |  .96   .1   .45   .0| 1.08 |   .85   .83 |  2 t1_R2 (mark 70)
|   12.5    25         .5    .99|  5.27    .60 |  .78  -.5   .45  -.8| 1.27 |   .81   .74 | 18 t2_R3 (mark 75)
|   12.5    25         .5    .99|  5.03    .59 | 1.37  1.0  1.71  1.1|  .61 |   .62   .73 |  1 t1_R1 (mark 62.5)
|   12.5    25         .5    .98|  3.83    .57 |  .95   .0   .67  -.2| 1.10 |   .71   .69 | 22 t2_S3 (mark 65)
|   25      50         .5    .95|  2.86    .37 |  .84  -.8   .68  -.7| 1.25 |   .69   .63 |  3 t1_R5S5 (mark 62.5)
|   25      50         .5    .94|  2.83    .35 | 1.10   .6  1.16   .6|  .81 |   .53   .58 |  4 t1_S4T3 (mark 55)
|   25      50         .5    .93|  2.63    .41 | 1.03   .2   .76  -.1| 1.00 |   .70   .70 |  6 t1_T4U6 (mark 50)
|   25      50         .5    .93|  2.52    .40 | 1.01   .1   .87   .0| 1.01 |   .69   .69 | 23 t2_T2U3 (mark 55)
|   25      50         .5    .91|  2.25    .37 | 1.07   .4   .90   .0|  .93 |   .61   .63 | 19 t2_R4S2 (mark 67.5)
|   12.5    25         .5    .88|  1.95    .61 | 1.01   .1  1.19   .5|  .90 |   .73   .74 | 20 t2_R6 (mark 70)
|   25      50         .5    .77|  1.20    .36 |  .84  -.7   .78  -.8| 1.22 |   .69   .60 | 21 t2_S1T1 (mark 60)
|   12.5    25         .5    .62|   .50    .52 | 1.27  1.0  1.51  1.2|  .58 |   .49   .63 | 24 t2_T6 (mark 57.5)
|   12.5    25         .5    .57|   .30    .54 |  .90  -.2   .93   .0| 1.10 |   .70   .67 |  7 t1_T5 (mark 52.5)
|   12.5    25         .5    .49|  -.03    .75 |  .95   .1   .63   .0| 1.06 |   .84   .82 | 31 t2_X1 (mark 42.5)
|   12.5    25         .5    .46|  -.16    .76 |  .98   .1   .68  -.1| 1.04 |   .85   .84 |  5 t1_S6 (mark 60)
|   12.5    25         .5    .46|  -.17    .49 | 1.00   .0  1.08   .3|  .97 |   .57   .58 | 12 t1_W5 (mark 37.5)
|   25      50         .5    .46|  -.17    .38 |  .87  -.5   .74  -.7| 1.18 |   .71   .65 | 26 t2_U2V3 (mark 50)
|   12.5    25         .5    .41|  -.37    .56 | 1.19   .6  1.39   .7|  .76 |   .63   .70 | 25 t2_U1 (mark 50)
|   25      50         .5    .37|  -.54    .33 | 1.13   .9  1.22  1.0|  .70 |   .41   .51 | 11 t1_V6W4 (mark 40)
|   12.5    25         .5    .33|  -.69    .45 |  .94  -.2   .87  -.4| 1.17 |   .53   .47 | 30 t2_W3 (mark 42.5)
|   12.5    25         .5    .33|  -.72    .46 |  .82  -.9   .74  -.9| 1.43 |   .62   .49 | 10 t1_V5 (mark 40)
|   25      50         .5    .18| -1.51    .32 | 1.00   .0  1.00   .0| 1.00 |   .46   .46 | 27 t2_V1W1 (mark 47.5)
|   12.5    25         .5    .15| -1.73    .59 |  .68 -1.0   .41  -.3| 1.40 |   .78   .71 |  9 t1_U5 (mark 45)
|   25      50         .5    .14| -1.80    .35 | 1.01   .1   .99   .0|  .99 |   .57   .57 | 14 t1_X4Y4 (mark 32.5)
|   25      50         .5    .11| -2.10    .34 | 1.04   .2   .97   .0|  .96 |   .53   .54 | 13 t1_W6X6 (mark 35)
|   25      50         .5    .11| -2.13    .39 | 1.12   .6   .95   .1|  .86 |   .61   .64 |  8 t1_U4V4 (mark 45)
|   25      50         .5    .07| -2.56    .35 | 1.04   .3  1.08   .3|  .93 |   .58   .60 | 29 t2_W2X3 (mark 37.5)
|   25      50         .5    .06| -2.81    .32 |  .99   .0   .93  -.2| 1.05 |   .47   .46 | 32 t2_X2Y2 (mark 37.5)
|   12.5    25         .5    .06| -2.84    .57 |  .95   .0   .87  -.1| 1.06 |   .73   .70 | 28 t2_V2 (mark 45)
|   12.5    25         .5    .05| -2.85    .42 |  .86 -1.1   .84 -1.1| 1.75 |   .52   .32 | 17 t1_Y6 (mark 30)
|   12.5    25         .5    .05| -3.00    .42 | 1.11   .9  1.14   .9|  .37 |   .15   .31 | 16 t1_Y3 (mark 27.5)
|   12.5    25         .5    .04| -3.14    .42 |  .95  -.3   .95  -.3| 1.25 |   .39   .33 | 34 t2_Y5 (mark 37.5)
|   12.5    25         .5    .02| -3.93    .47 | 1.04   .2  1.11   .4|  .91 |   .50   .53 | 33 t2_Y1 (mark 30)
|   12.5    25         .5    .01| -4.60    .75 |  .95   .1   .63   .0| 1.06 |   .84   .82 | 15 t1_X5 (mark 35)
|-------------------------------+--------------+---------------------+------+-------------+------------------
|  Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Correlation |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | Nu Script
|-------------------------------+--------------+---------------------+------+-------------+------------------
|   17.6    35.3       .5    .47|   .00    .48 |  .99   .0   .92   .0|      |   .62       | Mean (Count: 34)
|    6.2    12.3       .0    .36|  2.75    .13 |  .13   .6   .28   .6|      |   .15       | S.D. (Population)
|    6.2    12.5       .0    .37|  2.79    .14 |  .14   .6   .29   .6|      |   .15       | S.D. (Sample)
+-----------------------------------------------------------------------------------------------------------
Model, Populn: RMSE .50  Adj (True) S.D. 2.71  Separation 5.42  Reliability .97
Model, Sample: RMSE .50  Adj (True) S.D. 2.75  Separation 5.50  Reliability .97
Model, Fixed (all same) chi-square: 1068.2  d.f.: 33  significance (probability): .00
Model,  Random (normal) chi-square: 31.9  d.f.: 32  significance (probability): .47
-----------------------------------------------------------------------------------------------------------
```