Comparing the Score Distribution of a Trial Computer-Based Examination Cohort with that of the Standard Paper-Based Examination Cohort

Nathan Zoanetti
Victorian Curriculum and Assessment Authority
<zoanetti.nathan.p@edumail.vic.gov.au>

Magdalena Les
Victorian Curriculum and Assessment Authority
<les.magdalena.m@edumail.vic.gov.au>

David Leigh- Lancaster
Victorian Curriculum and Assessment Authority
<leigh-lancaster.david.d@edumail.vic.gov.au>

From 2011 – 2013 the VCAA conducted a trial aligning the use of computers in curriculum, pedagogy and assessment culminating in a group of 62 volunteer students sitting their end of Year 12 technology-active Mathematical Methods (CAS) Examination 2 as a computer-based examination. This paper reports on statistical modelling undertaken to compare the distribution of results for this group with the standard cohort, and any differences in student response between the two groups at the item level.

Computer-based assessment has been trialled or introduced in a range of domains, such as the PISA 2006 computer based test for science assessment (OECD, 2010) and the current Certified Practising Accountants (CPA) Australia Practice Management examination (CPA Australia, 2014). On-line linear and adaptive testing for numeracy and mathematics has been used in Victoria since 1998 and adaptive on-line testing is being researched by the Australian Curriculum Assessment and Reporting Authority (ACARA) for possible implementation in NAPLAN from 2016 (ACARA, 2014). System wide *e*-assessment in mathematics is a major area of education discourse in the United States, with a range of states implementing, or planning for implementation, of this in the near future. However, none of these contexts involve high stakes senior secondary certificate mathematics examinations.

From 2011 – 2013 the Victorian Curriculum and Assessment Authority ((VCAA) conducted a trial based on aligning the use of a computer algebra system software, (*Mathematica* Version 8 in this case) as enabling technology for pedagogical practice, curriculum delivery and assessment, including examinations. The trial involved volunteer students from five regional and metropolitan schools, the Computer Based Examination (CBE) group, who completed their 2013 Mathematical Methods (CAS) Examination 2 in computer-based mode. The trial cohort of students studied the same curriculum and sat the same Examination 2 under the same general conditions (apart from the mode of delivery and response) as the standard cohort of over 15 000 students, who did the examination in traditional pen and paper mode, with CAS as a computational tool only. There are two end-of-year examinations for Mathematical Methods (CAS). Examination 1 is a one hour short answer and some extended-answer examination designed to assess student's knowledge of mathematical concepts, their skills in carrying out mathematical algorithms and their ability to apply concepts and skills in standard ways without the use of technology. Examination 2 is a two hour multiple-choice and extended-answer technology active examination

2014. In J. Anderson, M. Cavanagh & A. Prescott (Eds.). Curriculum in focus: Research guided practice *(Proceedings of the 37th annual conference of the Mathematics Education Research Group of Australasia)* pp. 685–692. Sydney: MERGA.

designed to assess student's ability to understand and communicate mathematical ideas, and to interpret, analyse and solve both routine and non-routine problems.

The VCAA trial was unique in that these students completed a high-stake end of secondary school certificate mathematics examination in computer-based mode. Throughout the trial these students developed familiarity with the use of *Mathematica* as a tool for working mathematically, in school-based assessment, and in trial tests and practice examinations provided by the VCAA. The preparatory tests and practice examinations, as well as the final examinations were carried out using a model developed by Wolfram Research according to VCAA specifications, and refined by feedback from schools.

Mathematica files are called *notebooks*, and can incorporate text, graphics and computations (Wolfram Research, 2014). The model developed has three components: a production palette that enables the examination document to be produced as a notebook file; a student palette that enables the examination to be run according to VCAA examination processes (log in, reading time, writing time, ending the examination); and a uniquely identified notebook that is the student's digital examination 'paper'. Once the examination is running, for the student this notebook operates like a standard *Mathematica* notebook with respect to working mathematically. It has some additional functionality that enables students to readily insert new computation and text cells and open and close the associated formula sheet as applicable.

Apart from the inclusion of radio buttons for responding to multiple choice items (for the standard pen and paper version students enter responses on a mark sense sheet) as shown in Figure 1:

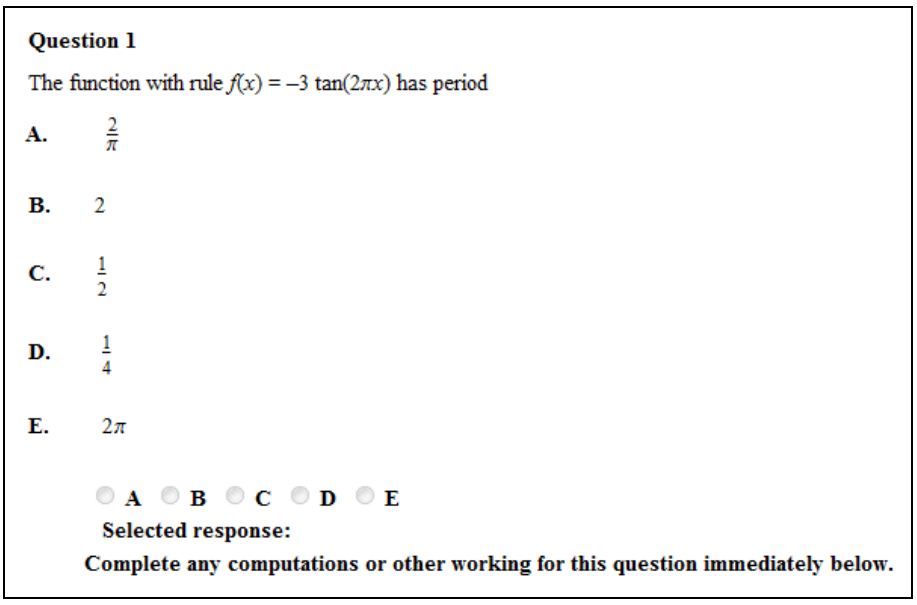

Figure 1: Multiple choice item showing radio buttons for response

and designated cells for computation and comment/discussion (for the standard pen and paper version students work/write in a lined section following a question or part of a question) as shown in Figure 2, the formatting and structure of the pen and paper and digital Examination 2's was the same.

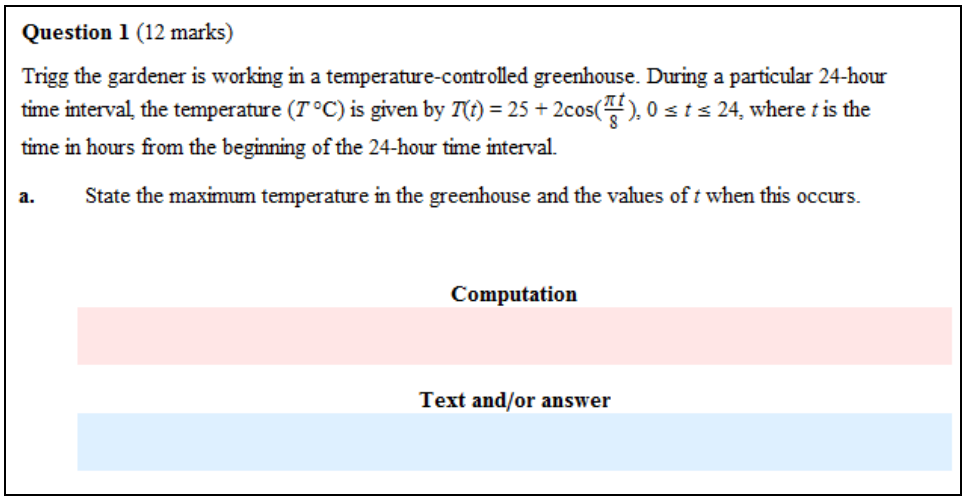

Figure 2: Extended-answer excerpt showing cells for response

Responses and working were auto-saved every two minutes, and students could also save at any time at their discretion. Answers and working could be amended, edited developed further, or deleted at any stage. Stimulus material was locked, and while selectable, not editable or able to be deleted.

A key distinctive feature of the digital mode is that the formulation of computations, results of these computations, and related analysis, discussion, commentary and the like all 'count' as material available to assessors as working and responses. In the traditional pen and paper mode, while student have access to a CAS calculator or software *as a computational tool only*, any working or results from these needs to be transcribed and suitably embedded in their written working and responses on paper. It is planned that a meta-analysis of student solutions in the digital mode will be reported on in future research.

Comparing the Score Distributions of the Two Cohorts

The following describes a series of analyses comparing the score distributions of students across the two delivery and response modes for the 2013 VCE Mathematical Methods (CAS) Examination 2 (VCAA, 2013). The purpose of these analyses was to evaluate whether the facility afforded, or difficulty presented, by the respective examination modes could be considered comparable. The evidence from this evaluation would then inform how the CBE group should be treated for subsequent scoring and reporting processes.

All students undertook the same paper-based technology free pen and paper Examination 1. All CBE and almost all of the standard cohort students also undertook the 2013 General Achievement Test (GAT). In any given VCE study, students receive a study score which reflects their relative rank within the study cohort based on the aggregation of their weighted, standardised scores typically across three graded assessments. Study scores are normally distributed, with mean 30, standard deviation 7, and maximum set at 50. This kind of scoring and reporting framework is generally regarded as being a norm-referenced framework, although it could more accurately be described as a cohort-referenced framework (Baird, Creswell, & Newton, 2000). In VCE Mathematical Methods (CAS) two of the graded assessments are external examinations, while the other graded assessment is school-based. In the case of the school-based graded assessment, a process of statistical

moderation is applied using the available external examination scores to establish the external reference score for the moderation process.

Both statistical moderation and study score calculations require that students have been assessed against a common, statewide scale. This requirement is addressed in a straightforward manner when a single standardised external assessment is undertaken by all students in the cohort. However, in the case of the CBE trial, this requirement could not automatically be assumed to have been met at the risk of adversely biasing scores for one group of students if one of the Examination 2 modes was inadvertently easier or more difficult than the other. This consideration was based on equivocal findings from hundreds of mode-effect studies spanning several decades in the education research literature (e.g. Bennett et al., 2008, Bunderson, Inouye, & Olsen, 1989). Variation in reported mode effects across these studies suggests that it is an issue that should be addressed within each assessment context (Bugbee, 1996). Familiarity and 'comfortableness' with context and medium seem to be a key factor in whether a significant mode effect is observed or not. Earlier studies tend to indicate an effect in favour of pen and paper mode, while more recent studies tend to indicate no significant mode effect. This seems to be in line with general familiarity with computers and other digital technologies in contemporary society.

A number of analysis models were specified a priori to test the hypothesis of whether the two external assessment modes could be treated as equivalent, thereby meeting the requirement that a common, statewide scale could be assumed for all students irrespective of the mode. If the evaluation described here identified that the mode had no statistically significant impact on overall score differences across the student cohorts, then all students could be considered as part of a single cohort for subsequent scoring and reporting processes. If, on the other hand, the mode appeared to have a statistically significant impact on the resultant score distributions, then the groups would need to be treated separately for a range of scoring and reporting processes.

The CBE students were enrolled under a distinct code so that they could be scaled separately if there was evidence that the computer-based mode was not of comparable facility/difficulty with the pen and paper mode. If facility/difficulty was found to be comparable, then the students could be treated as a single cohort for the purposes of grade scaling, statistical moderation and study score calculation. The analyses described in this section set out to identify the most appropriate approach. A baseline measure for the two groups being compared was required, so that any differences in final Examination 2 performance distributions controlled as much as possible for underlying mathematical ability. Examination 1 scores were included in the model to establish this baseline. GAT mathematics related component scores were also incorporated. Regression models and related Analysis of Variance (ANOVA) models from the Generalised Linear Model (GLM) class of statistical models were applied to reveal whether students of similar levels of ability on average achieved acceptably comparable Examination 2 scores irrespective of the examination mode after controlling for prior achievement in other, related, assessments.

Several regression and GLM ANOVA models were applied in the first instance, and two are reported here: *Model 1* predicted Examination 2 scores from Examination 1 scores and the examination mode and did not reveal a significant mode effect, $F(1, 15498) = 0.441$, $p = 0.506$. *Model 2* predicted Examination 2 scores from Examination 1 scores, GAT Mathematics, Science and Technology (MST) component scores and the examination mode and did not reveal a significant mode effect, $F(1, 15440) = 1.216$, $p = 0.27$. A series of mixed-effects multilevel regression models were also applied using

MLwiN (Rabash et al., 2009) and R (R Core Team, 2013) to verify the robustness of the initial findings. A linear mixed-effects regression model, with students nested within schools, was applied using the same covariate as Model 1. Using the R package, this model revealed that the coefficient for fixed mode effect was not significant (β = -0.02, $p = 0.995$), supporting the results from the GLM analyses.

These results indicate that the CBE mode did not impart a statistically significant effect on the facility/difficulty of Examination 2 after controlling for prior achievement on a range of other assessments. Irrespective of the covariates included in the model, and irrespective of whether a mixed-effects model was implemented to account for the nesting of students within schools, none of the models suggested that the null hypothesis of a zero mode effect should be rejected at conventional significance levels.

Differences at the Item Level

The results from the overall mode effect evaluation provided evidence that, on balance, examination delivery and response mode did not impart any statistically significant influence on student performance. This was an important result, however it did not provide information about whether *particular items* may have been easier or more difficult depending on the mode of delivery and response. To address this, additional analyses were undertaken. Noting that the sample size for the CBE group is small, methods that VCAA routinely applies as part of quality assurance for larger cohorts were applied to see if there were any indications of the kinds of items that might merit further inspection in larger samples. One of these methods is Differential Item Functioning (DIF) analysis (Holland & Wainer, 1993), implemented here as an extension of the Rasch (1980) model. DIF analyses seek to quantify whether test taker subgroups perform differently on specific items after controlling for estimated ability on the latent trait of interest. DIF can present as a uniform effect, where the facility/difficulty of an item is uniformly different between the groups being compared across the full ability range; or, it may present as non-uniform DIF, where the facility/difficulty of an item differs non-uniformly across the ability range. An earlier application of DIF has previously been described by two of the authors (Evans, Jones, Leigh-Lancaster, Les, Norton, & Wu, 2008).

In this case DIF analysis was applied to the set of 22 Examination 2 multiple-choice items (questions). As in a typical study of DIF, the individual item responses of the two groups were examined: a reference group (non-CBE), which was the majority, and a focus group (CBE group), which was the minority. The data used in this study were item-level responses from over 15 000 students. ConQuest 2.0 software was used for the DIF analysis (Adams, Haldane, Wilson, & Wu, 2007). The analysis indicates that one item *may* show uniform DIF, and while the sample size precludes considering this result as highly reliable, this sort of analysis and item review process is illustrative of the kind of quality assurance that is applied whenever the relative performance of different groups of students is of interest. In general, when an item exhibits indications of DIF, it is prudent for content experts to qualitatively review the item, to consider what features of the item, if any, may have contributed to the differential performance. This can then inform subsequent item design.

In this paper Question 22, which was found to be easier for the CBE group, is discussed in more detail. Figure 3 shows a comparison between the CBE and non-CBE groups average scores on this item at each mathematical ability level. In the following graph, the lower dotted curve shows the observed average score of students from the non-CBE group,

while the upper dotted curve shows the observed average score of students from the CBE group. The solid curve shows the expected score of students as a function of ability.

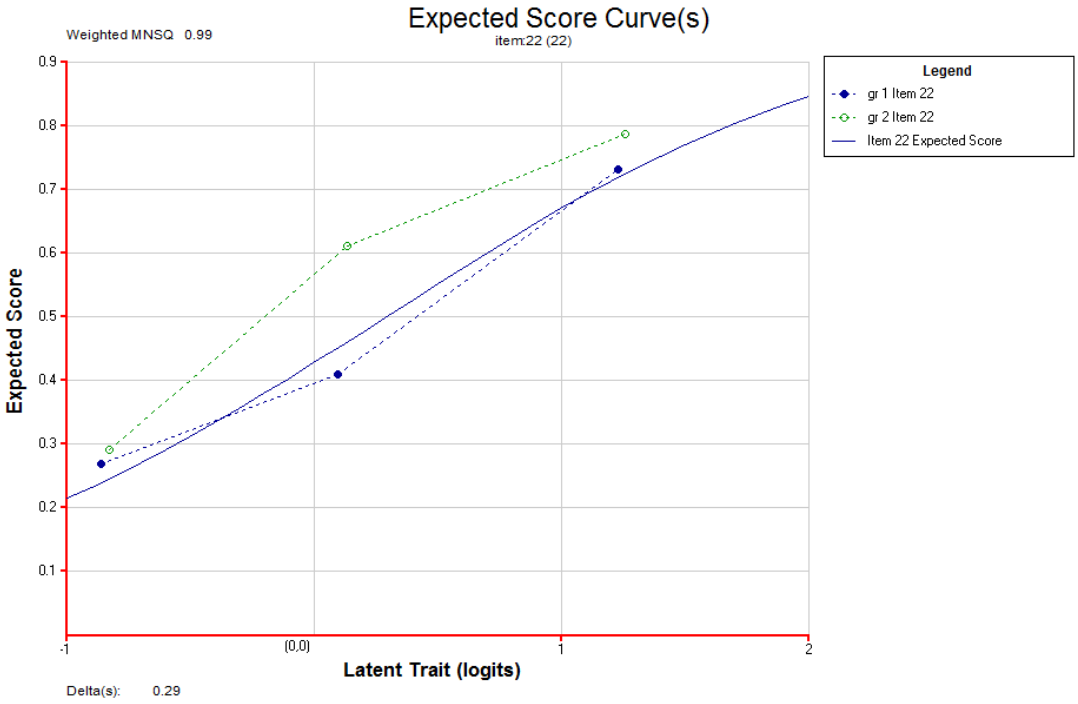

Figure 3: An item exhibiting DIF: the CBE group scores higher than the non-CBE group

Question 22, shown in Figure 4 involved a normally distributed random variable where the standard deviation was to be determined from given information.

Question 22

Butterflies of a particular species die T days after hatching, where T is a normally distributed random variable with a mean of 120 days and a standard deviation of σ days.

If, from a population of 2000 newly hatched butterflies, 150 are expected to die in the first 90 days, then the value of σ is closest to

A. 7 days
B. 13 days
C. 17 days
D. 21 days
E. 37 days

Figure 4: Multiple-choice Question22 from the 2013 Mathematical Methods (CAS) Examination 2

To answer this question students need to identify that $T \sim N\ (120,\ \sigma^2)$ and that $\Pr(T \leq 90) = 150/2000$. From this several approaches are possible, and each of these is similarly accessible for the CAS students. Like questions have been asked in previous years, so students who had been diligent in past paper practice would have familiarity with the nature and style of question. There is no apparent a priori reason why this item should be more or less amenable to correct response on the basis of the *mode* of examination delivery and response; however other explanatory factors may be relevant. For example, there may a *pedagogical* basis for students in the CBE group potentially using a given

approach more uniformly than in the general cohort, given the close working relationship of teachers involved in the trial, and sharing of common resources, including approaches to tackling particular types of questions. If this was the case, then why it might impact on some questions and not others would require further consideration.

Conclusions

The investigation of mode parity indicates that for 2013 the computer-based examination was of comparable facility/difficulty with the paper-based examination. A practical consequence was that the two student groups could reasonably be combined for all subsequent operational scoring and reporting processes.

The DIF analysis revealed that the 2013 multiple-choice items appeared to have comparable facility/difficulty for students from the non-CBE and CBE groups, with perhaps one item showing some indications of uniform DIF. However, these observations need to be confirmed by further research using larger samples. Analyses of this kind have the potential to reveal design features which may impact on item facility/difficulty differently across modes, an important aspect to be addressed as systems move to incorporate provision of computer-based and on-line assessment in mathematics and other studies.

References

Adams, R. J., Haldane, S., Wilson, M. R., & Wu, M. L. (2007). ACER ConQuest Version 2.0. Mulgrave. https://shop.acer.edu.au/acer-shop/group/CON3

Australian Curriculum, Assessment and Reporting Authority. (2014). *NAPLAN Online.* Retrieved 20/03/2014 from the World Wide Web: http://www.nap.edu.au/online-assessment/naplan-online/naplan-online.html.

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

Bates, D., Bolker, B., Maechler, M., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6.* http://CRAN.R-project.org/package=lme4

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in *NAEP. Journal of Technology, Learning, and Assessment,* 6(9). Retrieved 18/03/2014 from http://www.jtla.org.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education,* 28, 282-299.

Bunderson, C.V., Inouye D.K., & Olsen J.B. (1989). The four generations of computerized educational measurement in R. L Linn (ed) *Educational measurement.* Washington DC. American Council on Education 367-407.

CPA Australia (2014). *Practice Management Computer Based Exam*. Retrieved 19/03/2014 from the World Wide Web: http://www.cpaaustralia.com.au/professional-resources/practice-management/public-practice-certificate/public-practice-learning/exam.

Evans, M., Jones, P., Leigh-Lancaster, D., Les, M., Norton, P., & Wu, M. (2008). The 2007 Common Technology Free Examination for Victorian Certificate of Education (VCE) Mathematical Methods and Mathematical Methods Computer Algebra System (CAS). In M. Goos, R. Brown & K. Makar (Eds.) *Navigating currents and charting directions* (Proceedings of the 31st annual conference of the Mathematics Education Research Group of Australasia, pp. 331-336). Brisbane: MERGA.

Holland P. W., & Wainer H. (1993). *Differential Item Functioning*. Hillsdale. NJ: Lawrence Erlbaum.

OECD. (2010). PISA Computer-Based Assessment of Student Skills in Science. OECD Publishing. Paris.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009) *MLwiN Version 2.1.* Centre for Multilevel Modelling, University of Bristol.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

R Core Team (2013). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Retrieved 18/03/2014 from http://www.R-project.org/

Victorian Curriculum and Assessment Authority. (2013). *Mathematical Methods(CAS) Examination 2*. Retrieved 20/03/2014 from the World Wide Web: http://www.vcaa.vic.edu.au/Pages/vce/studies/mathematics/cas/casexams.aspx

Wolfram Research. (2014). *Computation + Knowledge*. Retrieved 20/03/2014 from the World Wide Web: http://www.wolfram.com/