Research Article https://doi.org/10.12973/ejmse.3.2.79

European Journal of Mathematics and Science Education
Volume 3, Issue 2, 79 - 90.
ISSN: 2694-2003
http://www.ejmse.com/

Comparing Examination Standards without Graded Candidate Scripts

Ian Jones* Loughborough University, UK
Colin Foster Loughborough University, UK
Jodie Hunter Loughborough University, UK

Received: January 24, 2022 ▪ Revised: April 13, 2022 ▪ Accepted: August 3, 2022

Abstract: Comparative judgement methods are commonly used to explore standards in examination papers over time. However, studies are limited by a paucity of graded candidate scripts from previous years, as well as the expense and time required to standardise scripts. We present three studies that attempted, without the use of graded candidate scripts, to replicate and extend previous results about standards in mathematics examination papers. We found that re-typesetting examination papers into a consistent format was necessary, but that comparative judgement of examination papers without an archive of graded candidate scripts offered a reliable and efficient method for revealing relative demand over time. Our approach enables standards comparison where previously this was not possible. We found a reasonable correlation between judgments of actual student scripts and judgments of the items only, meaning that conclusions may be drawn about the demand of examination papers even when graded candidate scripts are not available.

Keywords: *Comparative judgement, comparing demand, mathematics, student scripts, re-typesetting.*

To cite this article: Jones, I., Foster, C., & Hunter, J. (2022). Comparing examination standards without graded candidate scripts. *European Journal of Mathematics and Science Education, 3*(2), 79-90. https://doi.org/10.12973/ejmse.3.2.79

Introduction

A-level mathematics is a school-level qualification in parts of the United Kingdom taken at age 18 that is commonly expected by universities for new-entrants taking mathematics, science, engineering and similar degree programmes (Croft et al., 2009). Over the years many concerns have been raised that university entrants have insufficient mathematical knowledge and skills for university study (Noyes & Dalby, 2020), and that this is getting worse over time (e.g. Coe, 2007; Lawson, 2003). For example, Jones et al. (2016) reported that standards in A-level mathematics have declined since the 1960s. Their conclusion rested on experts, specifically mathematics Ph.D. students, making relative judgements about the quality of student responses to A-level examination questions. A scale was produced using comparative judgement (Bramley et al., 1998), a widely used method that we describe below.

Jones et al. were able to conduct the study because they acquired a historic archive of graded candidate scripts. However, the archive was sparse, containing scripts from only four time points between 1964 and 2012, as shown in Figure 1. Moreover, a complete set of scripts at these four time points was available only at grades B and E. Nevertheless, the existence of such an archive enabled Jones et al. to draw concrete and readily communicable conclusions, such as that "a candidate who achieved a grade B in 1996 or 2012 appears to have … performed approximately at the level of a candidate who achieved a grade E in 1964" (p. 555). The authors commented that "further graded scripts filling the gaps are unlikely to be found" (p. 557) and stressed the importance of archiving future scripts in a systematic manner (cf. Robinson, 2007).

* **Corresponding author:**
Ian Jones, Loughborough University, UK. ✉ I.Jones@lboro.ac.uk

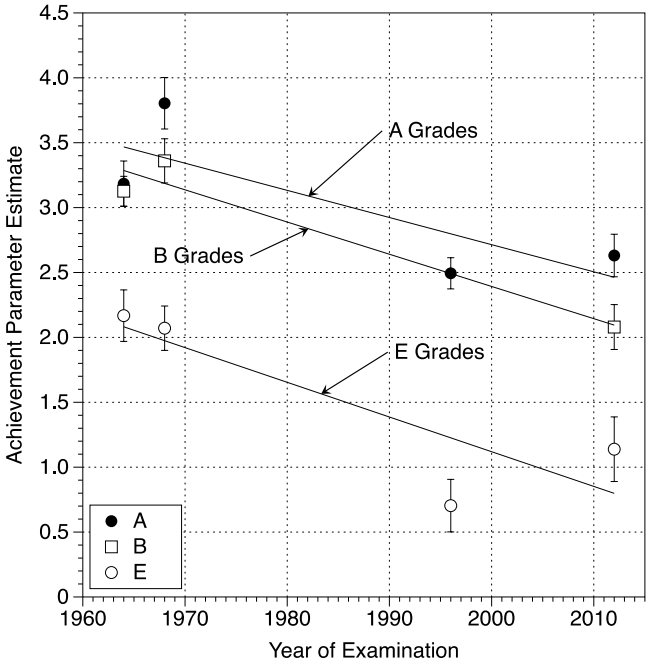

Figure 1. The perceived difficulty of A level Mathematics examinations from 1964 to 2012 (Reproduced from Jones et al., 2016

In the present three studies, we explored whether the possibility of evaluating item demand, sometimes called question difficulty (Ofqual, 2015), is indeed limited, as Jones et al. assumed, to situations in which graded historic scripts can be obtained. We drew inspiration from two sources. First, was a secondary study included in Jones et al., in which experts (mathematics PhD students) made relative judgements about the quality of *model* solutions to the items in the examination papers. This secondary study was presented as a validity check on the main study in which experts judged candidate scripts. Jones et al. reported that the correlation between outcomes from the main study and the secondary study was high, r = .68. Moreover, linear regression with year as a predictor of judgement outcomes of model solutions replicated the overall pattern of results found for year as a predictor of graded candidate scripts. This provides support for using model answers rather than actual scripts to investigate standards in qualifications.

The second inspiration for the present work was a comparative judgement study by Holmes et al. (2017), in which judges did not have access to either candidate scripts or model solutions. Holmes et al. investigated the demand of items in mathematics examinations (General Certificate of Secondary Education, GCSE) sat by most 16-year olds in England, Wales and Northern Ireland. In contrast to Jones et al., the experts made relative judgements of item difficulty, rather than of the quality of candidates' responses. Subsequently, Holmes et al. administered the items to students in a low-stakes assessment context. Holmes et al. reported a high correlation, disattenuated r = .76, between the experts' perceived difficulty of items and students' actual performance on the same items. This high correlation provided support for the validity of investigating standards based solely on judgements of assessment items.

In this paper, we first summarise the comparative judgement technique that underpins much contemporary standards comparison research. We then present three studies that explored the extent to which previous results about standards in school mathematics qualifications over time could be replicated and extended without using graded candidate scripts. The first study replicated the study by Holmes et al. (2017), but using the A-level items from the examination papers investigated by Jones et al. (2016). Consistent with Holmes et al., and contrary to the approach taken by Jones et al., the question items were not re-typeset, but were presented to expert judges in their original font and layout. In the second study, a different group of experts judged the same items, but this time the items were re-typeset, for consistency of presentation. In the third study, another group of experts judged entire examinations, rather than individual examination items, and the papers were re-typeset. An independent group of experts also judged entire examinations with model solutions that were scribed by a single person, again to ensure consistency of layout.

Comparative Judgement

Comparative judgement involves measuring the difficulty of an assessment item or the quality of a candidate response by using relative rather than absolute judgments (Jones & Sirl, 2017; Pollitt, 2012; Tarricone & Newhouse, 2017). It has long been known that human beings are more reliable when comparing one object to another than they are when evaluating a single object by itself (Thurstone, 1927). In comparative judgement, experts are repeatedly presented with pairs of items or candidate responses and asked each time to decide which item is more difficult or which candidate response is better (Davies et al., 2021). Judgments are quick and intuitive, and the outcomes of multiple judgments

from a pool of experts are modelled statistically to give a parameter estimate of, say, the difficulty of each item, allowing a scale of item difficulty to be constructed (Wheadon et al., 2020). Comparative judgment can offer reliable methods of determining item difficulty for two reasons. First, as mentioned above, relative judgments are more accurate and precise than absolute ones. Second, the fast speed with which judgments may be made allows those from multiple experts to be averaged, reducing the effects of any idiosyncratic judgments (Pollitt, 2012).

Comparative judgment methods are particularly appropriate for standards comparison research, such as the present study, where evaluating items separately would be very difficult, due to the many differences of examination style and content in the papers under consideration. For these reasons, comparative judgement has been deployed across several comparison-of-standards studies in England and Wales (Bramley, 2007; Bramley & Gill, 2010; Jones et al., 2016).

Study 1: Items Only (Not Re-typeset)

In Study 1, we sought to establish whether the comparative judgment outcomes of examination paper difficulty depend on whether items are presented in their original formats or are re-typeset for consistency.

Study 1 Methodology

We adapted the methods used by Holmes et al. (2017) in the comparative judgement component of their study, but for the case of the A-level items used by Jones et al. (2016). As with Holmes et al., the items were not re-typeset, as they were in Jones et al., but were presented to the judges in their original font and layout.

The examination papers corresponding to the archive of candidate scripts studied by Jones et al. (2016) were used in Study 1 (see Appendix 1). The papers were spliced into individual question items, totalling 42 items across the four papers. The items were presented in their original font and format, but with question numbers and the number of marks omitted. Two example items, one from the oldest examination paper and one from the most recent examination paper used in the study, are shown in Figure 2. Several differences in font and presentation are evident in these items. For example, the more recent item contains more white space on the page.

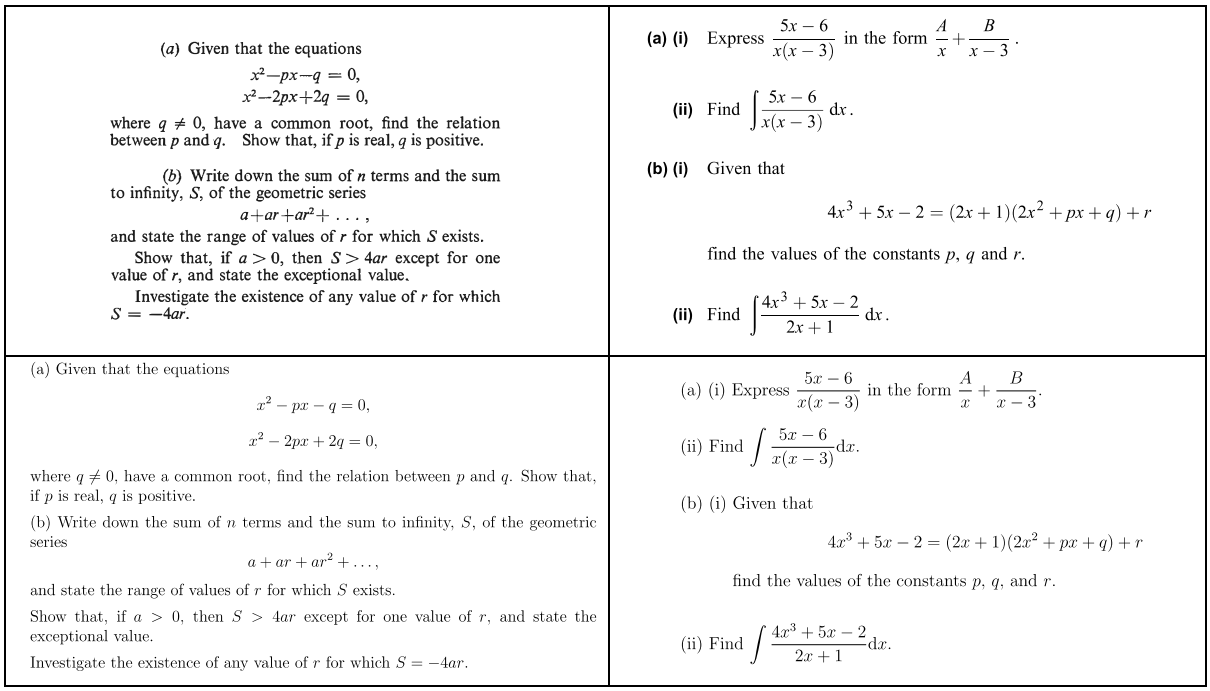

Figure 2. Example items in their original font and format (top) and in standardised font and format (bottom). Left: Q1 from JMB Paper 1 in Pure Mathematics for 1964. Right: Q1 from AQA Unit Pure Core 4 for 2012.

The 42 items were uploaded to the online comparative judgement platform nomoremarking.com. They were judged by eight experts (current or recent mathematics Ph.D. students studying, or having studied, in England), who were paid for their time. Following Ofqual (2015, p. 7), experts were asked to decide for each pairing "Which question is the more mathematically difficult to answer fully?" Each expert completed 83 or 84 pairwise judgements, and the median time per judgement was 47 seconds. The total number of judgements was 670 across the 42 items, which was just shy of the recommended 17 comparisons per object needed to achieve an internal consistency of .80 (Verhavert et al., 2019). The binary decision data were fitted to the Bradley-Terry model (Bradley & Terry, 1952) to produce a unique score of the perceived difficulty of each item. The internal consistency of the outcome was satisfactory (Verhavert et al., 2019), *SSR* = .91, as was the inter-rater reliability (split-halves, 100 iterations), r = .79.

Study 1 Results

The purpose of the analysis was to compare the outcomes of Study 1 with the results reported in Jones et al. (2016), where the items had been re-typeset into a consistent format. We calculated the Pearson Product-Moment correlation coefficient between the item scores and those reported in Jones et al. for the case of graded-candidate scripts and model solutions. Scores from Jones et al. were available for 38 of the 42 item scores. The missing four items were those which Jones et al. omitted from their study because no candidates in the historic archive of scripts had attempted them.

The correlation coefficient between the item scores from the present study and the item scores of the graded candidate scripts as reported in Jones et al. was positive, $r = .63$, $p < .001$. This was higher than the correlation coefficient between the item scores from the present study and the item scores of the model solutions reported in Jones et al., $r = .49$, although this difference did not reach significance ($p = .19$). Scatterplots are shown in Figure 3.

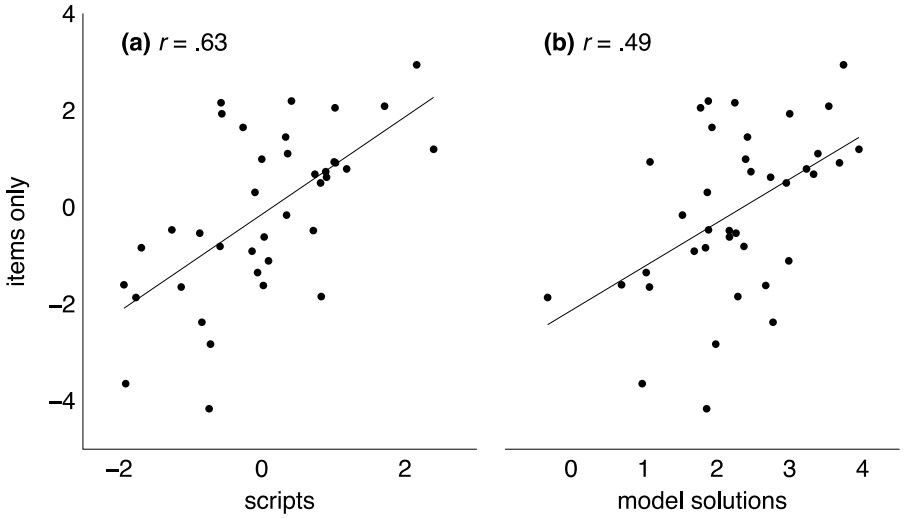

Figure 3. Correlations between Study 1 outcomes ("items only") and Jones et al. (2016) outcomes using (a) archived candidate scripts and (b) model solutions.

We also conducted a linear regression analysis with year as a predictor of item scores. Year was a significant predictor and explained 50% of the variance in item scores, $F(1,36) = 35.83$, $p < .001$, $b = -0.06$, $R^2 = .50$. This finding is consistent with the similar analysis reported by Jones et al. (2016, p. 551–552), but, to our surprise, year explained more variance in item scores than it did in model solution scores (28%) or in candidate script scores (27%).

Study 1 Discussion

The results reported by Jones et al. (2016), in which experts judged graded candidate scripts and model solutions, were broadly replicated when experts judged only examination items. We found positive and significant correlations of the item scores with candidate script and model solution scores, and we replicated the finding that year is a predictor of item scores. However, year explained almost twice as much variance in item scores as it did in candidate script or model solution scores. This finding was unexpected, given that the experts who made pairwise comparisons of the difficulty of items only might be thought to have had less information on which to base their judgements than did those with access to scripts or model solutions.

However, in the present study, the items were presented to the judges in their original font and formatting, whereas Jones et al. (2016) presented items and scripts or model solutions in a standardised font (items) and handwriting (scripts/solutions). It seems unlikely that the absence of scripts and model solutions would result in year explaining *more* variance than when they are included, and we therefore conjectured that the judges in Study 1 were influenced by font and format. Indeed, there is a considerable body of research examining the effects on the reader of different typographic variables (e.g., typeface/font, size, spacing, colour and alignment). Studies generally divide into those focused on communication, in areas such as political science or marketing (e.g., Haenschen & Tamul, 2020; Henderson et al., 2004), and psychological studies of reading fluency, often in the context of students with dyslexia or visual impairment (e.g., Krivec et al., 2020). Research examining subjective impressions associated with typographic features, both in printed form and on screen, indicates that different typefaces are perceived to have different personas (Brumberger, 2003). Although these effects may be small, they might be cumulative over time for extended exposure. In the context of education, effects on examination papers of different typesetting choices have rarely been explored (for an example, see Crisp et al., 2012). Further research is needed to establish whether students (or judges in comparative judgement tasks) perceive mathematics questions to be easier or harder depending on typographic variables.

To investigate the potential impact of varied fonts on the results of Study 1, we conducted a second study in which experts again judged items only, but, unlike in Study 1, these were presented in a standardised font.

Study 2: Items Only (Re-typeset)

In Study 2, we repeated Study 1 with an independent group of judges. We used the same items as in Study 1, except that they were re-typeset for presentation in a standardised font (Figure 2).

Study 2 Methodology

The 42 re-typeset items were uploaded to the online comparative judgement platform. They were judged by 10 experts (current mathematics PhD students studying, or having studied, in New Zealand), who were paid for their time. Each expert completed 84 pairwise judgements and the median time per judgement was 59 seconds. The total number of judgements was 840, which was 20 judgements per re-typeset item. This was two experts and 168 judgements more than was the case for Study 1, because using eight experts (672 judgements) produced an unacceptably low inter-rater reliability, r = .61 (internal consistency was acceptable, *SSR* = .86). The reliability of comparative judgement outcomes can be increased by adding further judgements (Verhavert et al., 2019), which we did to achieve an acceptable inter-rater reliability, r = .71 (*SSR* = .89). The additional experts and judgements did not substantively affect the item scores; the correlation coefficient between the scores using only the original eight experts and all 10 experts was high, r = .98. The analysis below uses the scores produced by all 10 judges, and we replicated it using the scores produced by the original eight judges only, which made no substantive difference to the results.

Study 2 Results

We first correlated the comparative judgement outcomes from Study 1 and Study 2, and found this was high, r = .92. We then replicated the analysis in Study 1, using the scores for the re-typeset items.

The correlation coefficient between the re-typeset item scores and the scores of the graded candidate scripts as reported in Jones et al. (2016) was positive, r = .55, p < .001. This was lower than the analogous correlation reported in Study 1, r = .63, although the difference fell short of significance (p = .11). The correlation coefficient was also higher than that reported by Jones et al. between the re-typeset item scores and the scores of the model solutions, r = .48, although this difference also did not reach significance (p = .67). Scatterplots are shown in Figure 4.

As in Study 1, we conducted a linear regression analysis with year as a predictor of item scores. Year was a significant predictor of item scores, $F(1,36) = 17.48$, p < .001, b = −0.04, R^2 = .33. The variance explained, 33%, was lower than for Study 1 (50%) and more in line with the variance explained in model solution scores (28%) and candidate script scores (27%).

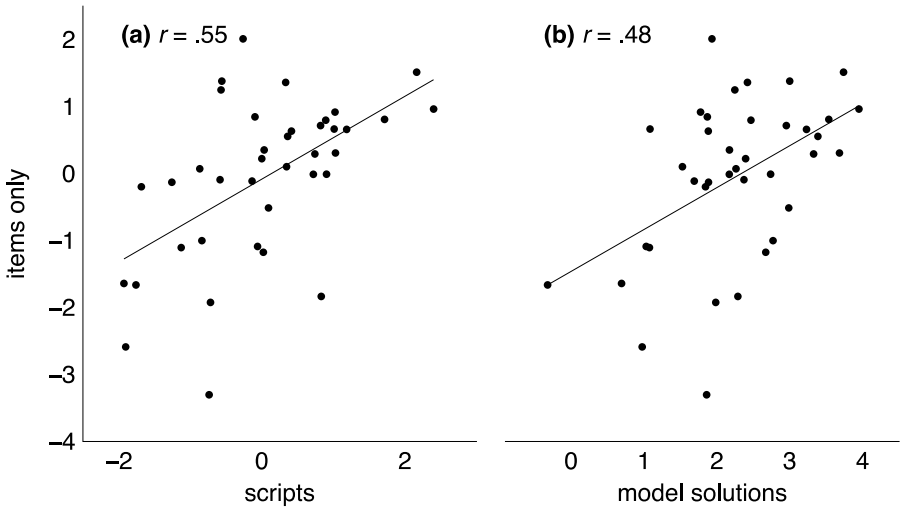

Figure 4. Correlations between Study 2 outcomes ("re-typeset items only") and Jones et al. (2016) outcomes using (a) archived candidate scripts and (b) model solutions.

Study 2 Discussion

Study 2 broadly replicated the findings of Study 1. Importantly, and in contrast to Study 1, in Study 2 year explained a similar amount of variance in re-typeset item scores as it did in candidate script or model solution scores. Moreover, the experts were less consistent judging the re-typeset items; hence, the need for more judgements to achieve an acceptable inter-rater reliability estimate than was the case for the non-re-typeset items This suggests that the font

differences of the items in Study 1 may have biased the experts' judgement decisions, and that standards research should use re-typeset and not original items.

Nevertheless, the correlation coefficients between the outcomes of Study 2 and those reported in Jones et al. were modest rather than large, and fell below the common reliability threshold of .70. From this we conclude that judging items alone is not adequate for establishing reliable estimates of the demand of mathematics examination papers. We therefore conducted a third study to investigate whether judging re-typeset items with model solutions would provide outcomes more consistent with those reported by Jones et al.

Study 3: Papers Only and Papers with Solutions

In Study 3 we obtained handwritten model solutions from a teacher, as explained below. This time, whole papers were judged, rather than individual items, in order to investigate the extent to which reliable judgments could be obtained in this way. As a validity check, we included the papers sampled by Jones et al. (2016).

Study 3 Methodology

The examination papers were those from studies 1 and 2, with seven additional papers from the summers of 1990, 1996, 2000, 2006, 2012 and 2017. The additional papers were sampled opportunistically after one of the authors discovered an archive of model solutions in standardised handwriting going back to 1990, constructed by a mathematics teacher over the duration of her career. We commissioned this teacher to produce further model solutions for the four examination papers used in Study 1, and this provided us with a complete set of re-typeset papers and model solutions in standardised handwriting for all the papers shown in Appendix 1.

The number of papers available each year changed as specifications changed. In order to attempt to compare like with like as well as possible, we included all papers covering the content of pure mathematics at Mathematics A-level, excluding Further Mathematics A-level. For the years 1964-1990, and 1996 AEB, this meant Paper 1 of two papers, and for 1996 ULEAC, 1996 UCLES and 2000 Edexcel, this meant Pure papers 1 and 2. Where occasional questions covered probability and statistics content, these questions were omitted from our analysis, as detailed in Appendix 1. With the advent of modular specifications, from 2006 we used the Core 1, 2, 3 and 4 papers, except that the 'comprehension' element of MEI Core 4 was not included. Where more than one examination paper was used, these were spliced together into one file for judging, which was then treated as one examination paper in our analysis.

Papers only. The 11 re-typeset examination papers were uploaded to the online comparative judgement platform nomoremarking.com. They were judged by five experts (mathematics PhD students based in New Zealand) who were not involved in studies 1 or 2 and who were paid for their time. Similar to Study 1, the experts were asked to decide for each pairing "Which exam is the more mathematically difficult to answer fully?" Each expert contributed 50 pairwise judgements, and the median time per judgement was 62 seconds. The total number of judgements was 250 across the 11 examinations, which exceeds the recommended minimum of 17 comparisons per object to achieve *SSR* = .80 (Verhavert et al., 2019). The decision data were fitted to the Bradley-Terry model to produce a score of the perceived overall difficulty of each paper. The internal consistency of outcome was satisfactory, *SSR* = .84. There were too few judges to estimate inter-rater reliability, although we note that many authors use Scale Separation Reliability as a proxy for inter-rater reliability (Verhavert et al., 2018).

Papers and model solutions. The re-typeset examination papers with model solutions were uploaded to the comparative judgement platform. They were judged by five independent experts (mathematics PhD students based in New Zealand) who were not involved in studies 1 or 2, or the judging of papers only, and who were paid for their time. The experts were asked to decide for each pairing "Which is the better mathematician?" Four experts contributed 70 pairwise judgements each, and one expert contributed 50, and the median time per judgement was 85 seconds. The total number of judgements was 330 judgements. There were more judgements than for papers only (250), because one paper and model solution was accidentally omitted when the experts were first assigned 50 judgements each. After the missing paper and solution were added, four of the judges contributed an additional 20 judgements each. The data were fitted to the Bradley-Terry model and the internal consistency of outcome was satisfactory, *SSR* = .87. Again, there were too few judges to estimate inter-rater reliability.

All sets of scores were transformed to *z* scores prior to the following analysis, for consistency of presentation and ease of interpretation.

Study 3 Results

There were three analytic goals for Study 3. First, we compared the outcomes of judging papers only with judging papers and model solutions. Second, we compared these outcomes with those of Jones et al. (2016). Third, we extended the picture presented in Jones et al. of standards over time in A-level mathematics.

Comparison of Study 3 outcomes. The Pearson Product-Moment correlation coefficient between the papers-only scores and model-solutions scores was high, $r = .74$, $p = .009$, suggesting good agreement across the two sets of judgements. A scatterplot is shown in Figure 5.

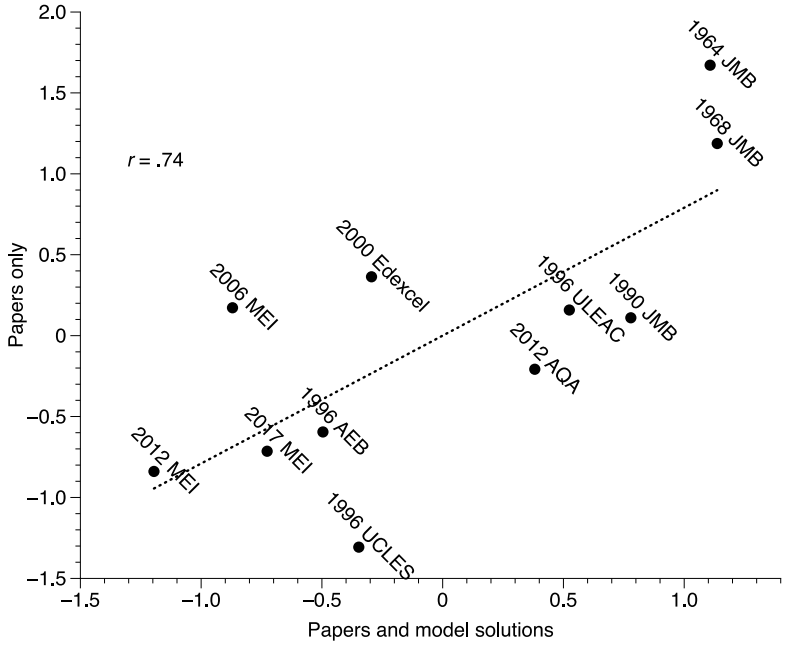

Figure 5. Scatter plot of scores from judgements of papers only and papers with model solutions.

Study 3 vs Jones et al. outcomes. Comparing our outcomes with those of Jones et al. required a different approach from that used in Studies 1 and 2, because there are only four common data points between Study 3 and Jones et al. In the absence of suitable statistical tests, we compared the sets of results graphically. Figure 6 plots the outcomes for papers only and model solutions for the case of the four examination papers used in studies 1 and 2 and Jones et al. For comparison, the outcomes from Jones et al. for the cases of graded candidate scripts and model solutions are also shown.

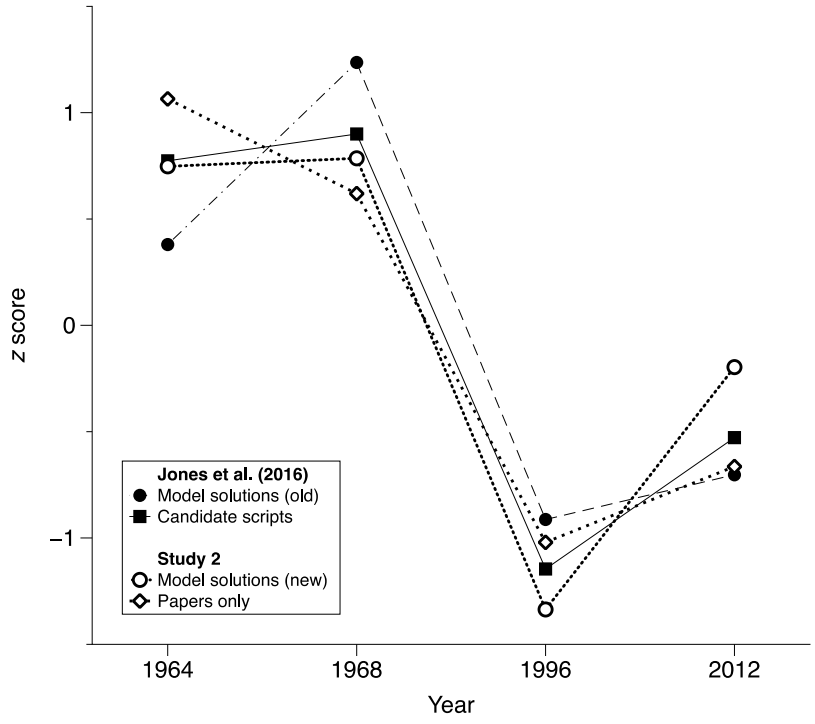

Figure 6. Comparison of results from Study 3 ("model solutions (new)" and "papers only") with results from Jones et al. (2016) ("model solutions (old)" and "candidate scripts").

The four sets of judging outcomes in Figure 6 follow the same broad pattern. With the exception of the papers-only judging outcomes, 1968 has a higher *z* score than 1964, meaning that the examination paper in 1968 was collectively perceived by experts as more difficult. For the remaining years, all four sets of outcomes are consistent in placing 1996 bottom, 2012 in the middle, and then the two 1960s papers top. This is in line with Jones et al. (2016), who found no difference between the perceived difficulty of the papers from 1964 and 1968 with the judging outcomes from either the model solutions or the candidate scripts. Similarly, they found no difference between the papers from 1996 and 2012. Figure 6 appears to follow this same pattern for the Study 3 judging outcomes: closely aligned scores across the four time points.

Extended picture of standards over time in A-level mathematics. This analysis gave us confidence to proceed with the third analytic goal for Study 3, which was to extend the picture presented in Jones et al. of standards over time in A-level mathematics. The scores for both papers-only and model-solutions are shown in Figure 7, where we can see reasonable agreement between the patterns of the two sets of scores.

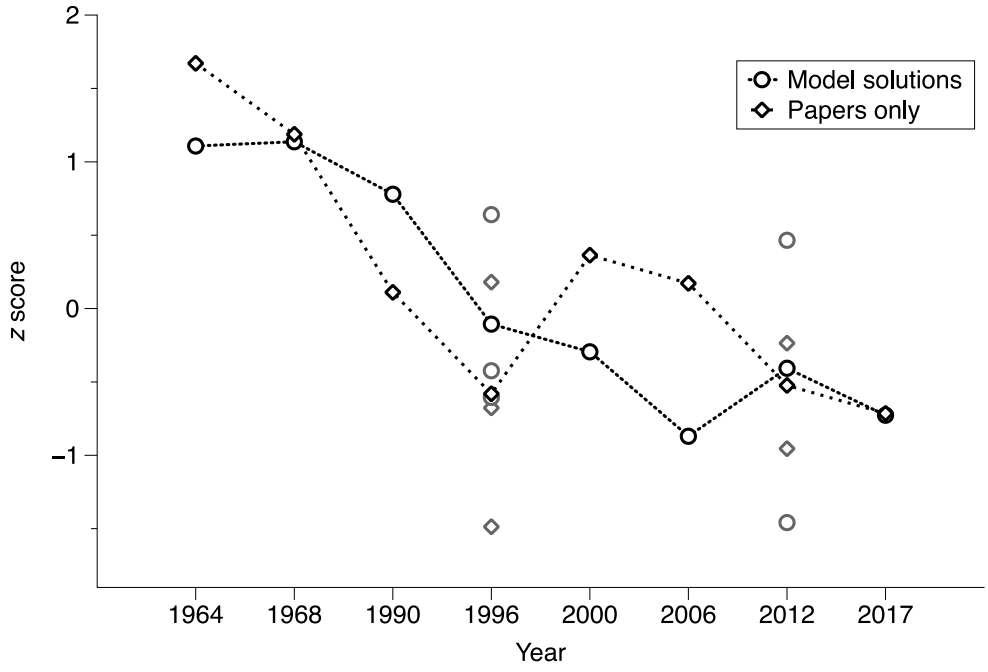

Figure 7. Full picture of standards over time from Study 3. Mean z score is shown for 1996 and 2012, with individual scores shown in grey for those years.

To further compare the outcomes, we conducted a linear regression with year as a predictor of comparative judgement scores. For papers-only scores, year was a significant predictor, and explained 60% of the variance, $F(1,9) = 13.38$, $p = .005$, $b = -0.05$, $R^2 = .60$. Similarly, for the model solution scores, year was again a significant predictor, this time explaining 62% of the variance, $F(1,9) = 14.68$, $p = .004$, $b = -0.05$, $R^2 = .62$. These almost identical outcomes provide evidence that, despite the imperfect correlation between papers-only and model-solutions scores, both methods would lead to the same general conclusions in an investigation into qualification standards. Given that the papers-only method requires considerably less resources, since model solutions do not need to be transcribed in standardised handwriting, we suggest that it should generally be the preferred method.

Study 3 Discussion

In Study 3 we presented three findings. First, we reported that judging papers only and judging papers and model solutions together produced broadly the same results, as assessed using the Pearson Product-Moment correlation coefficient.

Second, we found that the pattern of results for the examination papers included in the study by Jones et al. (2016) matched their results.

Third, we considered a fuller picture of changes in examination paper demand over time of A-level mathematics by including examination papers for which no archived scripts were available. How might we amend our view of A-level mathematics examination paper demand over the past six decades in light of this additional evidence? In Figure 7 we can see a decline from the 1960s to the mid-1990s, as reported by Jones et al. (2016). We have only added one additional timepoint to this picture, at 1990, which suggests that some of the decline took place between 1990 and 1996. We have added two additional timepoints between 1996 and 2012. Based on the model solutions outcomes, we

might conclude that demand stayed approximately constant over this duration, with a possible dip around 2006. However, based on the papers-only solutions, we might conclude that examination difficulty rose and fell throughout the late 90s and the 2000s. The demand in 2017 appears to be relatively low according to both sets of outcomes.

General Discussion

Comparative judgement techniques have been used to monitor standards in mathematics qualifications for decades (e.g., Bramley et al., 1998), but have traditionally required a historic archive of graded candidate scripts. However such archives are rare and so research into the mathematical backgrounds of undergraduates has been limited (Hodgen et al., 2020). In the research reported here, we investigated whether results based on graded scripts can be replicated and extended using only examination papers and model solutions. An affirmative answer would greatly enhance the scope of standards research because historic archives of graded candidate scripts are difficult and often impossible to acquire.

In Study 1, we found that when experts judged examination items the outcomes correlated positively with outcomes based on judgements of candidate responses to those same items. However, linear regression with year as a predictor of mean examination paper score showed that the item-only judging explained almost twice as much variance as candidate-response judging. This disparity appeared to arise because the items in the current study were presented in their original font and format.

In Study 2, we repeated Study 1 using the independent experts and the same items, except that they were re-typeset into a standard format prior to judging. The results from Study 1 were replicated, except that item-only judging explained about the same amount of variance as candidate-response judging. However, while the correlation coefficient between item-only outcomes and candidate-response outcomes was positive and significant, it was modest by the standards required to establish reliability in assessment research. Therefore we concluded that standards research studies in mathematics education should not be based on items only.

In Study 3, independent experts judged entire examination papers only, or entire examination papers with model solutions in standardised handwriting. We found that the two sets of outcomes correlated highly, and that both broadly replicated the findings reported in Jones et al. (2016). Moreover, we were able to include more examination papers than was possible in Jones et al., because we were no longer limited to only those papers for which graded candidate scripts were available. This enabled us not only to replicate but to extend the findings of Jones et al. to include additional years and examination boards.

Conclusion

We have argued that a papers-only approach might enable investigations of qualification demand where previously this was impossible due to a lack of archived scripts. Another advantage of using papers only is that it is less resource-intensive and time-consuming than the method used by Jones et al. (2016), which required transcribing 546 candidate responses to items to ensure consistency of handwriting. Even the model solutions method, used here in Study 3, and to a limited extent in Jones et al., requires obtaining or creating perfect answers to every item in standardised handwriting. By contrast, all that was required here was to re-typeset the examination papers, although even this step might not be necessary when comparing standards across contemporaneously published examination materials (as in Ofqual, 2015, 2017).

Another efficiency saving is for experts to judge entire examination papers rather than individual items, an approach that has been reported before (e.g. Bramley et al., 1998; Jones et al., 2014). This reduced the number of objects that needed to be judged in Study 3 from 256 (the sum of total items in Appendix 1) to 11 (the number of examinations). The median time per judgement for the model-solutions method (85 seconds) was longer than the median time per judgement for the papers-only method (62 seconds), which in turn was longer than the median time per judgement for the items-only method (47 to 59 seconds across studies 1 and 2, respectively). However, clearly, the time saved by having fewer objects needing judging more than makes up for the increased time required per judgement. Taking these values as typical, this would represent a potential time saving of around 94%.

Recommendations

Jones et al. were able to report their results in terms of equivalent grades at different timepoints, the headline finding being that a grade B in 2012 was roughly equivalent to a grade E in the 1960s. Here, we could only draw relative conclusions about examination paper demand being lower or higher than in the past. Therefore, the method reported here is only applicable where a relative comparison without reference to awarded grades is informative. For example, we might be interested in whether a major curriculum innovation has resulted in higher standards. We could sample papers from before and after the innovation to draw conclusions about the relative quality of candidates' responses. With adequate data points (examination papers), we might be able to report the findings in terms of effect sizes or Bayes factors, but not in terms of grade boundaries.

Limitations

In this paper we believe that we have offered a method that enables the investigation of qualifications standards without the need for an archive of graded scripts, at least for the case of mathematics examinations. However, all approaches to investigating standards are imperfect. We will not rehearse the general arguments and cautions regarding the limitations of standards-based assessment research here (see Goldstein, 1979; Newton, 1997). Instead we highlight limitations specific to the methods that apply comparative judgement in the absence of graded scripts, as advocated here.

First, the paucity of archived, graded scripts that motivated the research reported here also makes it difficult to validate our findings. There were only four data points provided by Jones et al. (2016) that could be directly matched with the present research. Consequently, statistical tests of the similarity of the present results with those reported by Jones et al. could not be conducted. Instead, we had to make graphical comparisons by eye. We have argued that a study using the methods reported here would have drawn the same overall conclusions as Jones et al., but we cannot estimate the probability that this replication happened by chance.

Second, examination papers vary in numerous dimensions other than difficulty of the construct which is intended to be assessed. An important dimension is surface appearance, including font and layout, and, as we found in Study 1, this can affect the judgements made by subject experts. In Appendix 1, we can see that the examination papers presented to judges in Study 2 varied widely in length, ranging from eight to 40 pages (for administrative purposes we re-typeset every item onto a single page in the present study). An expert's judgement of the difficulty of an examination paper might plausibly be affected by its length. Indeed, the Spearman Rank-Order correlation coefficient between comparative judgement scores for the papers-only method and total number of pages was moderate, $\rho = -.47$, although the confidence intervals, 95% CI [–.84, .20], show that ρ is unlikely to be positive and large, suggesting that judges were not simply associating longer examination papers with greater difficulty. Further research is required to identify and so minimise other construct-irrelevant features that might influence expert judges.

Third, our findings relate only to the case of A-level Mathematics. We offer no data or generalisations to other subjects, although we see no reason to assume that the method could not be applied beyond mathematics. Indeed, the difficulty of science examinations has been investigated by applying comparative judgement to items only (Ofqual, 2017). However, an examination-papers-only method may be relevant for only a limited range of subjects, such as mathematics and the sciences, where much of the assessment content is held within the questions. For example, it might not be valid to apply a papers-only – or even a model-solutions – method to investigate subjects assessed using essay prompts. In such cases the need for graded candidate scripts might be unavoidable. Further research is needed to ascertain whether this is the case.

Authorship Contribution Statement

Jones: Conceptualisation, design, analysis, writing. Foster: Design, analysis, writing and critical revision of manuscript. Hunter: Securing funding, data collection.

References

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*(3/4), 324–345. https://doi.org/10.2307/2334029

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards* (pp. 264–294). QCA. https://bit.ly/3vBiUmA

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education, 25*(3), 293–317. https://doi.org/10.1080/02671522.2010.498147

Bramley, T., Bell, J., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives, 25*(2), 1–24. https://bit.ly/3SlfgqH

Brumberger, E. R. (2003). The rhetoric of typography: the persona of typeface and text. *Technical Communication, 50*(2), 206–223. https://bit.ly/3ztXoBf

Coe, R. (2007). *Changes in standards at GCSE and A-level: Evidence from ALIS and YELLIS.* Durham, Centre for Curriculum, Evaluation and Management, Durham University. https://bit.ly/3JpTTAu

Croft, A. C., Harrison, M. C. & Robinson, C. L. (2009). Recruitment and retention of students: an integrated and holistic vision of mathematics support. *International Journal of Mathematical Education in Science and Technology, 40*(1), 109–125. https://doi.org/10.1080/00207390802542395

Crisp, V., Johnson, M., & Novaković, N. (2012). The effects of features of examination questions on the performance of students with dyslexia. *British Educational Research Journal, 38*(5), 813–839. https://doi.org/10.1080/01411926.2011.584964

Davies, B., Alcock, L., & Jones, I. (2021). What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views. *The Journal of Mathematical Behavior, 61*, 1-10. https://doi.org/10.1016/j.jmathb.2020.100824

Goldstein, H. (1979). Changing educational standards: a fruitless search. *Journal of the National Association of Inspectors and Educational Advisers, 11*, 18–19.

Haenschen, K., & Tamul, D. J. (2020). What's in a font?: Ideological perceptions of typography. *Communication Studies, 71*(2), 244–261. https://doi.org/10.1080/10510974.2019.1692884

Henderson, P. W., Giese, J. L., & Cote, J. A. (2004). Impression management using typeface design. *Journal of Marketing, 68*(4), 60–72. https://doi.org/10.1509/jmkg.68.4.60.42736

Hodgen, J., Adkins, M., & Tomei, A. (2020). The mathematical backgrounds of undergraduates from England. *Teaching Mathematics and its Applications. 39*(1), 38–60. https://doi.org/10.1093/teamat/hry017

Holmes, S., He, Q., & Meadows, M. (2017). An investigation of construct relevant and irrelevant features of mathematics problem-solving questions using comparative judgement and Kelly's repertory grid. *Research in Mathematics Education, 19*(2), 112–129. https://doi.org/10.1080/14794802.2017.1334576

Jones, I., & Sirl, D. (2017). Peer assessment of mathematical understanding using comparative judgement. *Nordic Studies in Mathematics Education, 22*(4), 147–164.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-Level Mathematics: Have standards changed? *British Educational Research Journal, 42*(4), 543–560. https://doi.org/10.1002/berj.3224

Krivec, T., Košak Babuder, M., Godec, P., Weingerl, P., & Stankovič Elesini, U. (2020). Impact of digital text variables on legibility for persons with dyslexia. *Dyslexia, 26*(1), 87–103. https://doi.org/10.1002/dys.1646

Lawson, D. (2003). Changes in student entry competencies 1991–2001. *Teaching Mathematics and its Applications, 22*(4), 171–175. https://doi.org/10.1093/teamat/22.4.171

Newton, P. (1997). Examining standards over time. *Research Papers in Education, 12*(3), 227–247. https://doi.org/10.1080/0267152970120302

Noyes, A. & Dalby, D. (2020). *Mathematics in England's Further Education Colleges: an analysis of policy enactment and practice*. The Mathematics in Further Education Colleges Project: Interim report 2. The University of Nottingham and The Nuffield Foundation. https://bit.ly/3oP5EqA

Ofqual. (2015). *A comparison of expected difficulty, actual difficulty and assessment of problem solving across GCSE Maths sample assessment materials* (No. Ofqual/15/5679). Ofqual.

Ofqual. (2017). *GCSE Science: An Evaluation of the Expected Difficulty of Items* (No. Ofqual/17/6163). Ofqual.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In P. Newton, J-A Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 97–123). QCA. https://bit.ly/3d42hcP

Tarricone, P., & Newhouse, C. P. (2017). An investigation of the reliability of using comparative judgment to score creative products. *Educational Assessment, 22*(4), 220–230. https://doi.org/10.1080/10627197.2017.1381553

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. D. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 1–22. https://doi.org/10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement, 42*(6), 428–445. https://doi.org/10.1177/0146621617748321

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212

Appendix

Re-typeset examination papers used in the two studies

Year	Exam board	Papers used	Total items	Total marks	Comments
1964	JMB	1	11	98	Q7 omitted due to missing model solution. Used in Jones et al. (2016) and Study 1.
1968	JMB	1	8	98	Q5 omitted due to missing model solution. Used in Jones et al. (2016) and Study 1.
1990	JMB	1	14	110	Q7 omitted due to missing model solution.
1996	AEB	1	12	95	Q4 (statistics) omitted. Used in Jones et al. (2016) and Study 1.
1996	ULEAC	1, 2	18	189	Q4 (probability) from Paper 2 omitted.
1996	UCLES	1, 2	18	103	Q10 (probability) from Paper 1 and Q8 (statistics) from Paper 2 omitted.
2000	Edexcel	1, 2	19	191	Q7 (statistics) from Paper 2 omitted.
2006	MEI	1, 2, 3, 4	40	288	'Comprehension' section of Core 4 omitted.
2012	AQA	1, 2, 3, 4	33	300	Paper 4 was used in Jones et al. (2016) and Study 1.
2012	MEI	1, 2, 3, 4	40	288	'Comprehension' section of Core 4 omitted.
2017	MEI	1, 2, 3, 4	40	288	'Comprehension' section of Core 4 omitted.