Judging the efficacy of artificial exemplar material

Mithra Vijayathasan, Anne Pinot de Moira and Neil Stringer

Summary

Before any examiner can mark national examinations in England they must be trained to use the mark scheme in a process called standardisation. The selection of exemplar material from live scripts for use in standardisation is time consuming and often fails to unearth a full range of candidate responses. This study investigates the possibility of improving the process by generating artificial exemplars in advance of the time-critical period. As a first step, it investigates whether there is any detectable difference in the quality of an exemplar dependent upon how it is created. Four different conditions are included: artificial exemplars written by the Principal Examiner in advance of the examination being sat; artificial exemplars written by the Principal Examiner after the examination is sat; standardisation exemplars selected in the traditional manner; and randomly selected live exemplars.

The study concludes that there is no apparent difference in the perceived quality of response dependent upon condition. However, exemplars randomly selected from the live scripts may be harder to judge. It is recommended that future research should focus on whether artificial exemplars can give rise to comparable, or higher, levels of marking reliability.

Keywords: comparative judgement, standardisation, examiner training, exemplars

Introduction

The time between the date of an examination and the awarding of grades is known as the marking period. During this period, responses are received from centres, examiners are trained (or standardised) in the use of the mark scheme, and the marking is undertaken. This must all be completed in time for the grade boundaries to be determined in an award meeting, and for the data from each candidate to be processed. Each stage of the marking period is subject to extreme time pressure. Improvements are continually sought to improve the process.

The standardisation of examiners has undergone many changes in the past few years. On the whole, it is no longer a face-to-face training exercise. It is delivered online and is a prerequisite for marking. Part of the process involves exemplification of aspects of the mark scheme by use of live responses. These live responses – or exemplars, as they are called – are augmented by commentaries. The selection of exemplars is a time-consuming process. It takes place in the first few days after the examination is sat and must be completed as soon as possible to maximise the time available for the marking itself.

This study considers the proposal that creating artificial exemplars in advance of the date of the examination might allow online standardisation to take place earlier; thus increasing the time for marking, while exemplifying a fuller range of candidate responses. Artificial exemplars would be responses written by the Principal Examiner to illustrate given qualities within an answer and to help with application of the mark scheme. In theory, if artificial exemplars could be shown to elicit similar marking behaviours and levels of marking reliability, they could be used to replace the current time-consuming and time-constrained script selection process. As a first step, the study investigates whether the artificial exemplars are distinguishable from exemplars selected

from live scripts. In other words, whether there are any systematic differences between artificial and live exemplars judged to be of the same quality by the Principal Examiner.

Data

Questions from a foundation tier GCSE Geography paper (90351F) sat in June 2015 were used to test the null hypothesis:

> *Ho: There is no detectable difference in the quality of response dependent upon the source of the exemplar material.*

This Geography paper is marked on screen at item level and thus the study could be easily restricted to just two questions. Both the questions, which are listed below, had a tariff of four marks and were chosen because of the potential for subjectivity in the evaluation of a response.

> *Question 1d (iii) Describe the formation of a coastal spit. You may use a diagram.*
>
> *Question 2e Study Figure 10, on the insert. Figure 10 gives information about the new eco-town at Rackheath. Describe some of the features that can help to make urban areas sustainable. Use Figure 10 and your own knowledge.*

Study design

The four conditions

The study included exemplar material created under four different conditions.

1. **Artificial-pre**: before the examination was sat, for each question, the Principal Examiner created ten artificial exemplar responses sufficient to cover each available mark (examples of handwritten artificial responses in situ are given in Appendix A & B).
2. **Artificial-post**: after the examination was sat, but before live marking, the Principal Examiner created a further ten artificial exemplar responses. These were sufficient to cover each available mark for each question. These responses were created after the Principal Examiner had seen candidate responses and gained a feeling for how the examination had panned out.
3. **Standardisation**: for each question, five exemplar responses were selected according to the normal standardisation procedures[1]. Necessarily, these responses were chosen from the live scripts available at the time of standardisation. They were included as the control in this study.
4. **Live**: for each question, ten exemplar responses were pseudo-randomly selected (to represent the full range of available marks) from live June 2015 scripts that had been marked by the Principal Examiner.

The responses from all four conditions were typed up and saved electronically. This ensured that all conditions were viewed in the same format and that artificial responses did not look as if they were purposefully created for the study by the Principal Examiner.

[1] The selection should cover the attainment range and include as many different types of response as possible.

Comparative judgement

In the past, comparative judgement has been used by the regulator and awarding organisations in comparability studies involving candidates' work and specification materials (Whitehouse, 2013). It has also been suggested that comparative judgement may offer an alternative to marking (Whitehouse & Pollitt, 2012). Here, it is used as a research tool to judge the quality of the responses in each of the four conditions. Rather than applying the mark based on a comparison with a mark scheme mediated through an examiner, comparative judgement places the responses in a subtler order of quality. Using Item Response Theory (IRT), a 'true score' is generated based on the outcomes of paired judgements made by a set of judges.

This method is used because of its sensitivity. When determining the mark for a response, it is often debatable whether that response is worth, say, three or four marks. However, when using comparative judgement, the task of choosing the better of two responses is simpler. After a number of judgements, one can expect an acceptable rank order to emerge. One potential issue with using comparative judgement may be that the resultant order does not correlate with the marks awarded using the mark scheme. A response may be judged to be better than the one it is paired with, without the judging criteria matching the criteria in the mark scheme. This can be a problem, especially with longer responses. However, as this study uses responses to questions that are worth four marks, it is felt that there is little scope for variation in the criteria used. In this study, the judges were not given the mark scheme. This would have worked against the use of holistic, instinctual subject expertise upon which the method of comparative judgement relies.

Judges were asked to compare pairs of responses and choose the better of the two. The aim was to see if the artificial exemplars matched the live responses in terms of quality. If, after analysis, there was no significant difference between the quality of work in each condition, it would indicate that the Principal Examiner was able to create realistic exemplar responses. However, if the artificial exemplars were judged to be significantly better or worse than the live responses, then the value of artificial exemplars in future standardisation training would be compromised.

Judges and judging

Initially, examiners who marked on the foundation tier of unit 1 of the GCSE were asked to partake in the study. This produced a sample of only half the required size. Mark allocations and mode of marking for the other question papers in the specification were checked to find more examiners who might be suitable participants. Consequently, examiners marking the higher tier were also asked to take part in the comparative judgement exercise. A total of 10 examiners were recruited as judges for the study.

Comparative judgements were made for each of the two questions separately. In other words, for each question, the 35 responses across the four conditions were paired randomly against each other. Each judge was required to make 70 comparisons per question and therefore, for each question, there were a total of 700 comparisons. A pilot test suggested that a judge's allocation of 70 comparisons would take approximately an hour.

Initially, for question 2e, the figure on the insert was not sent with the emailed instructions or provided when doing the judging on the website (Appendix C). Some judges managed to complete the comparative judgement exercise without this information, while others either requested it or accessed it themselves. To balance the study as far as possible, the insert was sent to all those judges who had not completed the exercise, and those who had already completed it were asked whether or not they used the figure. Only three of the ten examiners had completed the exercise before the insert was sent. Of these three, two had access to the insert and used it, while one felt the responses were assessable without the insert. One judge

also mentioned making reference to the mark scheme. These variations in the amount of information used by the judges could have led to differences in the criteria used while making judgements.

Results

The data from the comparative judgements were collected and analysed using the *NoMoreMarking* software (Wheadon, Henderson, & Jones, 2015). By fitting a Bradley-Terry model (Bertoli Barsotti & Punzo, 2012; Hunter, 2004), the software takes the comparisons and estimates a measurement scale, ordering the responses from weakest to strongest. It also provides details of the model fit in the form of infit statistics for both judges and responses. For a sample of 700 comparisons, infit statistics greater than 1.2 are said to indicate that judgements are not consistent with the overall measurement scale (Smith, Schumacker, & Bush, 1998). In the case of the current study, judges with high infit statistics would have views that are at odds with the majority. On the other hand, responses with high infit statistics would be those where the approach taken by the student was such that there was no common view on the quality of work.

The results are presented in two sections. Firstly, the performance of the judges is considered with a view to identifying misfits to the data. Secondly, the quality of the responses is analysed.

Judges

For each judge, the software recorded the total time taken for the exercise and median time per comparison. Every judge considered every available response at least once.

Table 1 shows that, for question 1d (iii), all but one of the judges' infit statistics were below the threshold deemed acceptable for a well-fitting model. The median time and the total time taken to make judgements differed between judges. Although there was little evidence to suggest any relationship between infit and time taken to make judgements, the judge who made judgements with the greatest alacrity was also the judge with the highest infit (Judge 3).

Table 1 Judge statistics for question 1d (iii)

		Total time taken			Median time taken
Judge	**Infit**	**Hours**	**Minutes**	**Seconds**	**Seconds**
1	0.91	0	36	23	23
2	0.89	0	20	38	13
3	***1.28***	***0***	***16***	***10***	***10***
4	0.74	0	42	17	25
5	0.93	0	37	26	25
6	0.95	0	26	54	18
7	0.86	0	30	42	22
8	0.65	0	35	21	24
9	0.92	0	28	53	17
10	0.83	1	54	55	35

For question 2e, Judge 3 also had an infit above the acceptable threshold, which suggests that the decisions made by this judge were inconsistent with the model (Table 2). Judge 3 had the lowest median judgement time of 10 seconds. It is possible that Judge 3 was too hasty in making the decisions. On the other hand, given that Judges 2 and 9 had median decision times of 12 seconds and 15 seconds, respectively, perhaps Judge 3 was simply not well suited to the

task. Given that the judgements made by Judge 3 might bias the findings of the study, this judge was removed from further analyses.

Table 2 Judge statistics for question 2e

		Total time taken			Median time taken
Judge	**Infit**	**Hours**	**Minutes**	**Seconds**	**Seconds**
1	0.73	0	28	27	21
2	0.93	0	20	35	12
3	***1.85***	***0***	***23***	***25***	***10***
4	0.87	0	35	08	23
5	0.82	0	50	40	36
6	0.96	1	30	10	23
7	0.69	0	57	21	34
8	0.64	0	33	10	19
9	0.65	0	31	23	15
10	0.79	0	52	14	38

Responses

For each response in the sample, the Bradley-Terry model produced an estimate of the true score and an infit statistic. The infit statistic is a measure of how consistently a response was chosen as the better of a pair. For question 1d (iii), there were three responses with an infit statistic greater than 1.2, which suggested that they were difficult to judge. The responses were retained in the analyses because it is anomalies such as these that provide a greater understanding of the effect of the four conditions. Two of the responses with high infit statistics were from the live condition and one was from the artificial-post condition.

The scattered points in Figure 1 show the relationship between true score, as calculated from the software, and the raw mark. The regression lines show the relationship between raw mark and the predicted true score based on the following linear model[2]:

$$True\ Score = ß_0 + ß_1\ Raw\ Mark + ß_2\ Condition + ß_3\ Raw\ Mark * Condition$$

Equation 1

On the graph, pink refers to artificially generated exemplars (conditions 1 & 2) and blue relates to exemplars taken from live responses (conditions 3 & 4). There is a strong, positive correlation between the raw mark and the true score for all the conditions in the study, as evidenced by the coefficient of determination, R^2, for each group. The adjusted multiple R^2 for the full model described by Equation 1 was 0.773. While raw mark was a significant independent variable in the model, neither condition nor the interaction of condition with raw mark had a significant influence over true score.

[2] Condition was included as a set of three orthogonal binary contrasts.

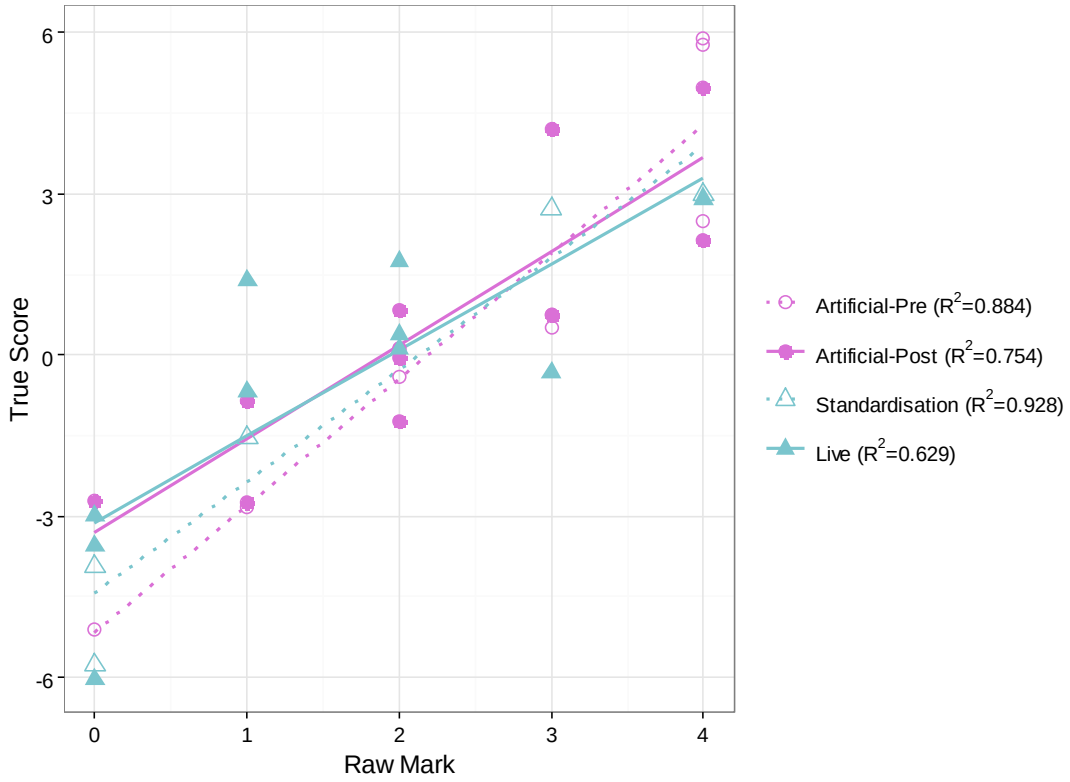


Figure 1 Raw mark versus true score for question 1d (iii)

A similar picture emerged for question 2e (Figure 2). Once again, there were three responses with high infit statistics. They came from three different conditions: artificial-pre, standardisation and live. The model fitted to the data had an adjusted multiple R^2 of 0.800 and the only significant independent variable was raw mark.

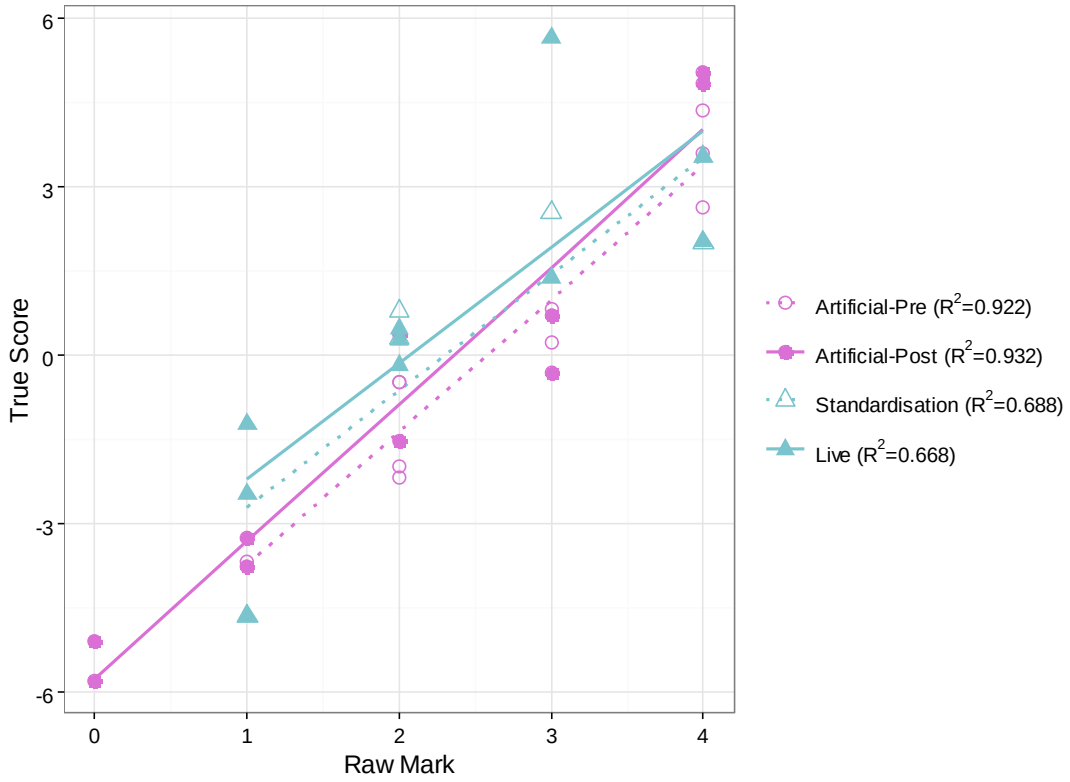


Figure 2 Raw mark versus true score for question 2e

Finally, the data for both questions were combined to produce an overall picture of the relationship between raw mark and true score under the four conditions (Figure 3). There was no significant effect of condition but what is interesting is that the R^2 for the live condition was considerably lower than that for the other conditions.

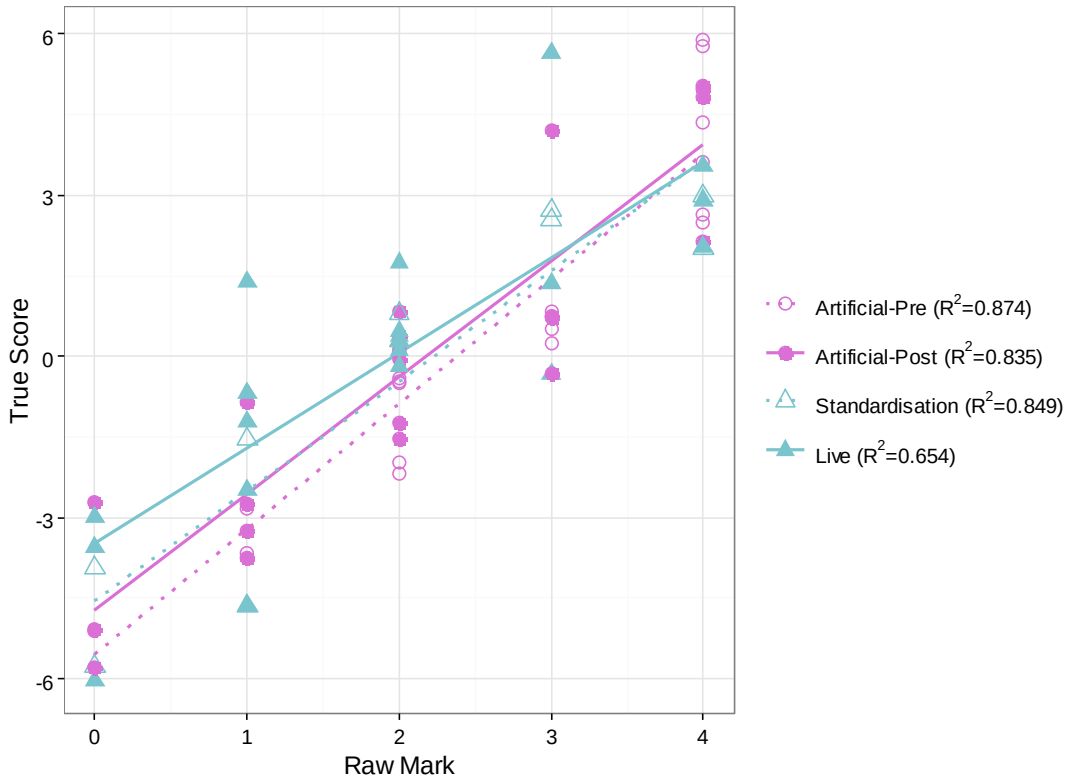


Figure 3 **Raw mark versus true score for question 1d (iii) and question 2e combined**

A residual plot shows that the linear model (Equation 1) fits less well for the live responses than for responses included from the other conditions (Figure 4). The solid blue triangles represent the live condition. In a well-fitting model, it is not unusual for a small proportion of standardised residuals to be outside the range -2 to +2. However, it is undesirable for patterns to be observed in a residual plot. There are four relatively large residuals and they are all for responses in the live condition. Interestingly, of these four responses, only one had a high infit statistic; the high infit suggesting disagreement between judges on the quality of work. The other three elicited no such disagreement, yet, in the view of the judges, the quality of work was different from that originally determined by the Principal Examiner.

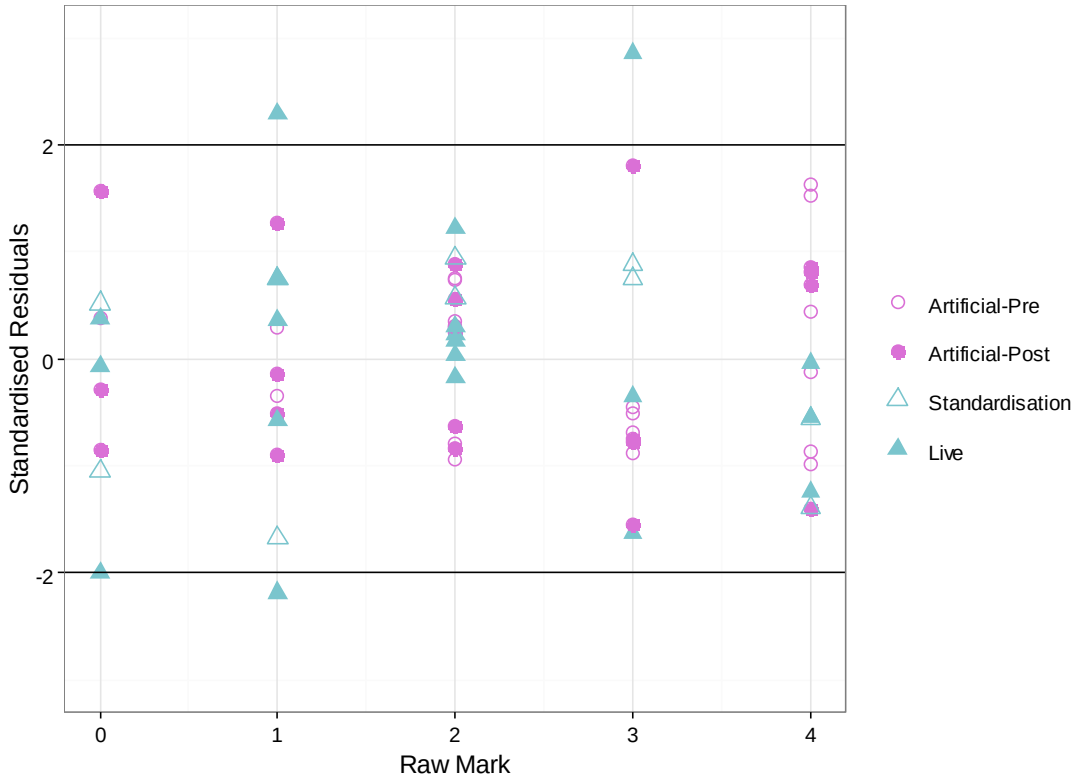


Figure 4 **A residual plot for the model of data from question 1d (iii) and 2e combined**

Therefore, it seems that the act of 'choosing' exemplar material, be it by active selection or by artificial creation, enables the Principal Examiner to focus on key aspects of a response as they relate to the mark scheme. Randomly selected exemplars do not have such clearly distinguished features and so, it would appear, are harder to judge in terms of quality.

Discussion

Limitations

Conclusions drawn from the findings of this study are not generalisable in the sense that they pertain to two low-tariff questions on one foundation tier GCSE Geography paper. They were also typed to remove bias from the study; a process that would not be viable in a live examination series. Furthermore, the artificial exemplars were generated by one willing Principal Examiner. However, they allow reflection on what is understood by the term exemplar, they provide the opportunity for a cost-benefit analysis, and they highlight areas where further research might be beneficial.

Conclusions

Exemplar material is provided to examiners at standardisation to help in embedding the mark scheme. Standardisation itself takes place to facilitate reliable marking. Time limitations place enormous pressure on Principal Examiners when preparing for standardisation, hence the constant efforts to improve the process. It is acknowledged that a common understanding of what is creditworthy is desirable but it is not always clear how this understanding might best be imparted. This study was set up to look at the effect of artificial exemplar material in supporting that common understanding. It tested the null hypothesis:

> *Ho: There is no detectable difference in the quality of response dependent upon the source of the exemplar material.*

It has shown that, once responses have been typed, there is little difference in the perceived quality of the exemplars, no matter how they are selected. On the whole, judges were able to rank responses reasonably accurately.

However, what is evident is that there is a slight difference in the ease with which exemplars can be ranked dependent on condition. In the study, we saw that the correlation between raw mark and true score was lower for live responses than for other responses. Live responses were the only exemplars randomly selected from the pool. Standardisation responses are carefully selected to exemplify the mark scheme, and the artificial responses were written to fulfil the same remit. So what is special about the live responses? Two possibilities emerge. The first is that, in the population of responses, there are very few responses that perfectly exemplify the mark scheme. In the act of choosing responses, the Principal Examiner limits the pool of choice to those that suit his or her need. Ironically, this act undermines the purpose of standardisation because examiners ultimately need to be able to mark responses of all descriptions. The second possibility is that, when the Principal Examiner chooses or writes exemplars, he or she selects those with unambiguous wording. While this lack of ambiguity is not enough to make the exemplars appear better than live responses, it is enough to simplify the job of rank ordering. In other words, exemplars in conditions 1, 2 and 3 are just a little bit too 'perfect'.

Therefore, as far as the current standardisation process stands, this study suggests that the creation of artificial exemplars would not introduce bias. After the judges made comparisons between the exemplars in the four different conditions, no significant differences in the quality of work were found. However, the study also suggests that the provision of exemplar material might not be a panacea for reliable marking; idiosyncratic responses will always divide opinion.

Cost and time benefits

Any change to the standardisation process needs to be supported by sound research evidence but also a sensible cost-benefit analysis. For this study, the Principal Examiner wrote 10 complete scripts for condition 1 and for condition 2. This process took approximately five working days. The same Principal Examiner spent approximately two days creating the standardisation material for the live examination in summer 2015. For this process, he only needed to select five exemplars (Table 3). Therefore, the traditional method of selecting exemplars appears less time and cost efficient.

Table 3 Time taken to create or select exemplars for standardisation

	Number of scripts	Days	Scripts per day
Artificial exemplars	20	5	4.0
Standardisation	5	2	2.5

However, while writing artificial exemplar material is more time efficient, it might still be regarded as a more onerous task. Indeed, after taking part in the study, the Principal Examiner described his experience:

> *'There is no doubt that looking at a lot of scripts and considering the range of ideas that candidates are coming up with and then writing a range of answers covering all of the possibilities is quite a challenging job'*

Nevertheless, the idea behind creating artificial exemplars was not just to increase time efficiency at the pre-standardisation stage but also to increase the effectiveness of standardisation. By providing a fuller range of candidate responses, the artificial exemplars might improve the consistency and accuracy of marking and also increase examiner confidence levels. In turn, this might reduce the number of team leader interventions and examiner attrition.

Recommendations

At the start, it was suggested that if artificial exemplars could be shown to elicit similar marking behaviours and levels of marking reliability, they could be used to replace the current time-consuming and time-constrained script selection process. This study has shown that artificial exemplars seem, at least, to pass for live responses. However, there is still work to be done to be content that they are a feasible alternative. To make a move towards the use of artificial exemplars, further research is recommended to be sure that they give rise to comparable, or higher, levels of marking reliability. The awarding organisation would also need to determine whether all Principal Examiners are equally capable of creating credible artificial exemplar material. Finally, it is recommended that further work is done to explore the differing effects of using live, atypical and specifically chosen exemplars in the establishment of a reliable workforce of examiners. After all, subject to gaining a greater understanding of the effect of 'choosing' exemplars, it is possible that simply using randomly selected live responses might be a prudent method of expediting, or even improving, the marking process.

Acknowledgements

We would like to thank the Principal Examiner for his help with this study. It was he who suggested the use of artificial exemplars in standardisation and who created the exemplars for use in the study. Our thanks also go to Claire Whitehouse who helped with an early draft of this paper.

References

Bertoli Barsotti, L., & Punzo, A. (2012). Comparison of two bias reduction techniques for the Rasch model. *Electronic Journal of Applied Statistical Analysis*.

Hunter, D. R. (2004). MM Algorithms for Generalized Bradley-Terry Models. *The Annals of Statistics*, *32*(1), 384–406.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using Item Mean Squares to Evaluate Fit to the Rasch Model. *Journal of Outcome Measurement*, *2*(1), 66–78.

Wheadon, C., Henderson, B., & Jones, I. (2015). No More Marking: An online platform for Comparative Judgement. Retrieved from https://nomoremarking.com/

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester, UK: AQA Centre for Education Research and Practice.

Whitehouse, C. & Pollitt, A (2012). *Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment.* Manchester, UK: AQA Centre for Education Research and Practice.

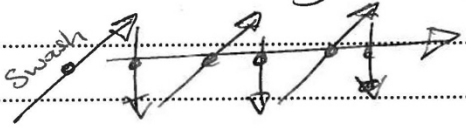
Appendix A: Question 1d (iii) in situ with Principal Examiner's artificial response

1 (d) (iii) Describe the formation of a coastal spit.

You may use a diagram.

[4 marks]

A coastal spit is formed by material being moved along the beach by longshore drift and then deposited when the land curves inland. Longshore drift is when the material is moved up the beach as swash at the angle of the waves and back down at 90°. Because of this it is moved along the beach.



Extra Space

wind/
waves

Question 1 continues on page 10

Appendix B: Question 2 (e) in situ with Principal Examiner's artificial response

2 (e) Study **Figure 10**, on the insert. **Figure 10** gives information about the new eco-town at Rackheath.

Describe some of the features that can help to make urban areas sustainable.

Use **Figure 10** and your own knowledge.

[4 marks]

The town has carbon neutral buildings so there will be less pollution. The environment will be protected, it looks like there are quite a lot of trees and grass. There is recycling which means that there will be less stuff thrown away.

Extra Space

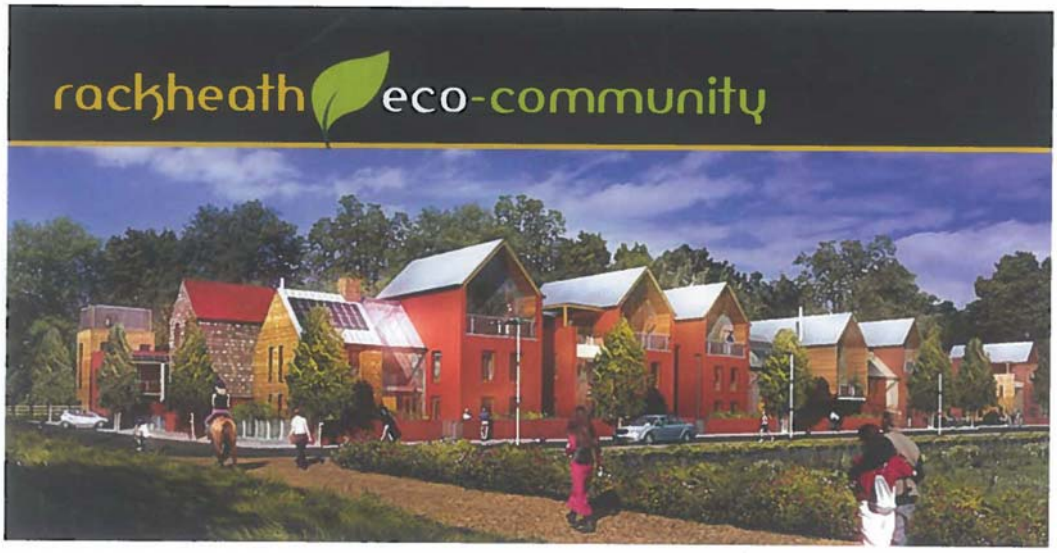
END OF QUESTIONS

Appendix C: Insert

Figure 10

For use with Question 2 (e) – Foundation Tier

For use with Question 2 (f) – Higher Tier



What is an eco-town?

In 2007, the UK government announced proposals for a number of sustainable eco-towns to be built. This initiative came about because of the need to develop more residential settlements during a time of housing shortages. The newly built eco-towns would also be used as examples for future residential developments.

Rackheath, near the city of Norwich, was one of twelve eco-towns proposed by the government. It is designed to be a sustainable and self-sufficient settlement of 5000 homes.