CAMBRIDGE ASSESSMENT

Investigating the reliability of Adaptive Comparative Judgment

Tom Bramley

Cambridge Assessment Research Report

23rd March 2015

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

Author contact details:

ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

Bramley.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

How to cite this publication:

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Contents

Introduction

The aim of this study was to investigate by simulation the reliability of comparative judgment (CJ) and adaptive comparative judgment (ACJ) methods as a means of creating a scale for use in educational assessment contexts. Of particular interest was whether adaptivity inflates the apparent reliability.

The motivation and rationale for CJ and ACJ, the theoretical background, and several practical applications have been described in numerous articles (e.g. Pollitt, 2004, 2012; Bramley, 2007; those listed in Table 1) and will not be repeated here. In CJ studies, experts are asked to place two or more[1] objects into rank order according to some attribute. The 'objects' can be examination scripts, portfolios, individual essays, recordings of oral examinations or musical performances, videos etc. The attribute is usually 'perceived overall quality'. Analysis of all the judgments creates a scale with each object represented by a number – its 'measure'. The greater the distance between two objects on the scale, the greater the probability that the one with the higher measure would be ranked above the one with the lower measure. The A for 'adaptivity' in ACJ means that the choice of which objects are presented to the judges depends on the outcomes of judgments made so far – the idea being to use judge time efficiently by not presenting objects to compare where the result of the comparison is almost certain. The analogy is with computerised adaptive testing (CAT) where the choice of the next item to present is based on the examinee's success or failure on items presented so far. This allows for better targeting of tests to examinees and for an equivalent level of precision in the estimate of an examinee's ability to be obtained from fewer items than in a fixed length test.

Cambridge Assessment's position on the use of CJ/ACJ was given in Bramley & Oates (2011), who noted the distinction between using CJ/ACJ as a tool in comparability research or standard maintaining exercises; and its more radical use as an alternative to marking. It is in its latter use where claims about reliability have been more prominent, so studies where CJ/ACJ was used as an alternative to marking were the focus of this investigation.

Although achieving high reliability is by no means the only motivation for using CJ/ACJ, it is certainly promoted as an advantage of the method, as the following quotes show:

> The judges are asked only to make a valid decision about quality, yet ACJ achieves extremely high levels of reliability, often considerably higher than practicable operational marking can achieve. It therefore offers a radical alternative to the pursuit of reliability through detailed marking schemes. (Pollitt, 2012, p281).
>
> After round 6 … the reliability rose sharply, reaching an alpha of 0.96 after 16 rounds. This is an extraordinarily high value for the reliability of a writing test; for comparison Royal-Dawson reported values ranging from 0.75 to 0.80 in a study of marking reliability in English national curriculum writing at age 14. (Pollitt, ibid. p286-7)
>
> In the primary writing study, there were 499,500 possible comparisons, but only 8161 decisions were made, or 0.016 per possible pairing. To create a consistent measurement scale with so little data shows the remarkable robustness of ACJ. (Pollitt, ibid. p288)
>
> It is worth noting that the average standard error of measurement for a portfolio was 0.668, which is just less than half the width of one of these "GCSE" grades [1.385 logits]. It is unlikely that many GCSE

[1] In online applications of CJ/ACJ, as far as I am aware, it is only pairs of objects that have been compared to date.

> components – or any that are judged rather than scored quite objectively – could match this level of measurement accuracy. (Kimbell et al. 2009, p80).

All the published studies cited in this report (see Table 1) used the same statistic as the indicator of reliability of the scale produced by a CJ/ACJ exercise, but did not always use the same name for it. In this report it will be called the Scale Separation Reliability (SSR). It is the same as the 'index of person separation' described in Andrich (1982) and the 'test reliability of person separation' in Wright & Masters (1982), but these terms are not as apt for CJ/ACJ since the scaled items are usually scripts or portfolios or works of art etc. It is calculated as follows:

The software estimates a 'measure' for each object being compared (henceforth 'script'), and an associated standard error. First, the 'true SD' of the script measures is calculated from:

$(\text{True SD})^2 = (\text{Observed SD})^2 - \text{MSE}$.
where MSE is the mean squared standard error across the scripts.

The SSR is defined like reliability coefficients in traditional test theory, as the ratio of true variance to observed variance, i.e.:

$\text{SSR} = (\text{True SD})^2 / (\text{Observed SD})^2$.

Sometimes another separation index G is calculated as the ratio of the 'true' spread of the measures to their average error, i.e.:

$G = (\text{True SD}) / \text{RMSE}$
where RMSE is the square root of the MSE. Then the SSR can also be calculated as

$\text{SSR} = G^2 / (1+G^2)$.

It is somewhat frustrating to note the difficulty that CJ/ACJ practitioners have had in getting these formulas reproduced correctly in their published work[2]. In Heldsinger & Humphry (2010, p9) the formula is the wrong way up, and in Pollitt (2012, p286) G was defined as (Observed SD) / RMSE instead of (True SD) / RMSE, and then SSR as $G^2 / (1+G)^2$ instead of $G^2 / (1+G^2)$. A later corrigendum (Pollitt, 2012b) kept the new definition of G and gave the correct formula for SSR as $(G^2-1) / G^2$ with this new definition. Since much CJ/ACJ work has involved or been inspired by Pollitt, it is worth bearing in mind that reported values for SSR are likely to be comparable with other Rasch literature, but values for G may not be. For high reliabilities this may not matter, e.g. for an SSR of 0.8 the Pollitt G would be 2.24 compared to a value of 2 on the earlier definition.

[2] I know from personal experience that errors in formulas can be introduced by the publisher in the typesetting process. I am not criticising these authors, but trying to clear up some confusion in the literature.

Table 1. Design features* and SSR reliability results from some published CJ/ACJ studies.

Study	Adaptive?	What was judged	#scripts	#judges	#comps	%max	#rounds	Av. # comps per script	SSR
Kimbell et al (2009)	Yes	Design & Tech. portfolios	352	28	3067	4.96%		14 or 20 bimodal	0.95
Heldsinger & Humphry (2010)	No	Y1-Y7 narrative texts	30	20	~2000?			~69	0.98
Pollitt (2012)	Yes	2 English essays (9-11 year olds)	1000	54	8161	1.6%	16	~16	0.96
Pollitt (2012)	Yes	English critical writing	110	4	(495)	(8.3%)	9	~9	0.93
Whitehouse & Pollitt (2012)	Yes	15-mark Geography essay	564	23	3519	2.2%	(12-13)	~12.5	0.97
Jones & Alcock (2014)	Yes	Maths question, by peers	168	100,93	1217	8.7%	N/A?	~14.5	0.73 0.86
Jones & Alcock (2014)	Yes	Maths question, by experts	168	11,11	1217	8.7%	N/A?	~14.5	0.93 0.89
Jones & Alcock (2014)	Yes	Maths question, by novices	168	9	1217	8.7%	N/A?	~14.5	0.97
Newhouse (2014)	Yes	Visual Arts portfolio	75	14	?	?	?	13	0.95
Newhouse (2014)	Yes	Design portfolio	82	9	?	?	?	13	0.95
Jones, Swan & Pollitt (2015)	No	Maths GCSE scripts	18	12,11	151,150	100%	N/A	~16.7	0.80 0.93
Jones, Swan & Pollitt (2015)	No	Maths task	18	12,11	173,177	114%	N/A	~19.5	0.85 0.93
McMahon & Jones (2014)	No	Chemistry task	154	5	1550	13.2%		~20	0.87

*The values in the table for numbers of scripts, judges, comparisons and rounds have either been taken from the listed articles or calculated based on information provided in the article. The latter calculations may have involved some assumptions.

Values for SSR have generally been high or very high in published CJ/ACJ work where the method was being used as an alternative to marking. For example, in the 13 studies listed in Table 1, the majority were above 0.9 and only one was below 0.8. The purpose of the simulation studies reported in the next section was to investigate how the SSR statistic compared with the true reliability in various conditions. In particular, it was of interest to find out whether adaptivity produced biased estimates of the reliability – i.e. whether the SSR was higher than the true value.

Simulations of CJ/ACJ exercises

Study 1 – SD of 'true quality' 1.7 logits

This study used two simulated data sets, one containing the results of all possible paired comparisons among 100 scripts N(N-1)/2 = 4950; the other from among 1000 scripts (499500). These simulations were intended to be baseline simulations, kept as simple as possible. All comparisons were statistically independent, fitting the Rasch formulation of Thurstone's case V paired comparisons model[3] (see equation 1 below) and thus treatable as coming from a single 'ideal' judge.

$$p(A > B) = \frac{e^{(\beta_A - \beta_B)}}{1 + e^{(\beta_A - \beta_B)}} \qquad (1)$$

where p(A>B) is the probability of script A being judged to be better than script B, β_A is the measure (scale value) of script A and β_B is the measure of script B.

It was not entirely straightforward to choose a distribution of 'true scores' for the simulated scripts. The method of analysis constructs a scale from the judgments with a pre-constrained mean of zero, but the SD appears to be free to vary. This is potentially misleading, because the SD of the scale is not separable from the implied average discrimination of the judges. The analysis effectively sets this discrimination at 1 unit. (In the Rasch model for dichotomous items this is the much disputed assumption/requirement that all items discriminate equally – see Humphry & Andrich (2008) for further discussion of this issue.) The scale unit created by the analysis – the logit (log-odds unit) – is not a unit of measurement like a metre or degree Kelvin, but rather a pure number interpretable as the log of the odds that a script with a given measure would win a paired comparison against a script with a measure of zero (i.e. an average script). To choose a value for the 'true' SD of a set of script measures thus requires an assumption about what is a reasonable value for the probability that a script 1 SD above the mean would 'beat' an average script. Using the well-known fact that the cdf[4] of the logistic distribution is very similar to that of the normal distribution with a multiplier of 1.7, first 100 and then 1000 random numbers were sampled from a unit normal distribution and rescaled to set the mean to 0 and the SD to 1.7. This corresponds to a probability of ~0.84 for a script 1 SD above the mean winning a comparison with the average script, and puts the 5th and 95th percentiles at ± 3.33 logits. Choosing a higher or lower value than 1.7 would therefore correspond *either* to assuming a higher 'true' spread of scripts, *or* higher discrimination for the judges (perhaps experts or trained judges can discriminate better than novices). In fact, these are just different sides of the same coin, because the true values are only defined in terms of what the judges can discriminate.

[3] The model is usually described in the wider psychology and psychometrics literature as the Bradley-Terry-Luce model but in the CJ/ACJ assessment context it has been derived from the work of Thurstone and Rasch.

[4] Cumulative distribution function.

A win or loss for each simulated comparison was produced by comparing the modelled probability of a win from equation (1) above with a random number from the uniform distribution in the range 0-1 and assigning a win (score of 1) if the modelled probability was greater than the random number and 0 otherwise.

It should be noted that an outcome for every possible paired comparison was simulated once only for simplicity. It is of course possible in a CJ/ACJ study for the same pair to be judged more than once – either by different judges or the same judge. It is not always clear from the published studies whether repeated judgments of the same pair of scripts were allowed.

The different analyses were based on the same two underlying datasets. First of all, observations from these datasets were sampled in order to mimic either random allocation of pairs of scripts to judges, or adaptive allocation (where the probability of being sampled depends on the outcome of earlier comparisons). In all cases the selection of observations was carefully planned to ensure that i) each script was compared the same number of times; and ii) the complete set of judgments formed a fully linked design whereby each script was compared directly or indirectly with every other script[5]. For the 'random only' analyses, this could be achieved by systematically sampling the required observations. For analyses involving adaptivity, however, the temporal nature of the process needed to be recognised. This was done by dividing the comparisons into rounds where half the scripts were compared with the other half. To achieve random pairings by this method, the standard 'round-robin' allocation algorithm[6] for tournaments where all players/teams have to play all other teams over a series of rounds was used. To achieve adaptive pairings, a version of the 'Swiss' method (alluded to in Pollitt, 2012) was used. Scripts were ranked according to total number of wins so far (i.e. total from previous rounds[7]) and then, starting at the top of the list, each unpaired script was paired with the first script below it in the list that it had not been compared with before. This algorithm is not optimal because it means scripts at the bottom end of the list may end up unpaired. However, the effect of losing a few comparisons was assumed to be negligible. Finally, a third method of allocation was simulated – that of comparing all scripts with the same fixed set of scripts. Whilst this might not be a practical method in the real world because the judges would repeatedly encounter the same scripts (the reference set), it still provides a linked design and is statistically identical to the Rasch analysis of a N-item dichotomous test (where N is the number of scripts in the reference set). Therefore it could shed light on both the conception of reliability, and on interpreting the SSR.

To summarise, the three allocation methods are somewhat analogous to standard tests of dichotomous items constructed from an item bank in the following way:
Random – a test of N items is randomly sampled from a bank for each examinee;
Adaptive – a test of N items is sampled adaptively from a bank for each examinee;
Fixed – a test of the same (random) N items from the bank is given to each examinee.

It is important to note at this point that the method of adaptivity used here was less sophisticated than the adaptivity used in the published studies in Table 1, because there was no estimation of parameters after each round. Pollitt (2012, p286) implies that the Swiss method is rather crude: "…providing *very rough* sorting by quality." (emphasis added). This assessment is supported by his Figure 2 (ibid.) which shows the SSR increasing rapidly after Swiss rounds are replaced by truly adaptive rounds. However, without specifically comparing them on a like-for-like basis after

[5] This might not be the case if there are 'disjoint subsets': sets of scripts whose members are only compared with each other and not with scripts in other subsets.
[6] As described at http://en.wikipedia.org/wiki/Round-robin_tournament
[7] Even a fully adaptive allocation method requires one random round to get started.

the same number of rounds it is not possible to conclude from this how much a Swiss adaptive method differs from a truly adaptive method in this context.

None of the works cited in Table 1 explains exactly how scripts were paired adaptively. Possibly this is because the software underlying the technology had proprietary elements, possibly because there was not space in the journal or it was not deemed sufficiently interesting. Recently the NoMoreMarking team has announced on its website[8] that its algorithms for CJ/ACJ are publicly available, and that they are based on the progressive adaptive algorithm of Barrada, Olea, Ponsoda & Abad (2008). This algorithm was developed for selecting the next item to present in a normal computer adaptive test (CAT). It gives a higher weight to random allocation at the start of the test and a lower weight to the statistical information[9]. As the test progresses, the relative weights of these two components change at a rate that can be controlled by an 'acceleration parameter'.

It is not quite clear what constraints have been applied in practice to selection of possible pairs – my guess is that because this is real-time software which judges can access at any time, a number of ad hoc but sensible constraints are used to make sure that all scripts are compared at roughly the same rate, no judge sees the same script too often etc. What is not clear is whether there are rules preventing repeats of the comparison between the same two scripts, whether or not it is by a different judge. This is potentially important because in situations with a relatively small pool of scripts and a relatively large number of rounds, it is much easier to find pairs if you don't have to worry about whether they have been matched against each other before.

The main point of all this is that while the simulations should provide useful insight about the size of the SSR statistic in CJ/ACJ studies, it is not clear how closely they resemble 'live' studies, or whether there are differences that could potentially affect the interpretation.

In all cases the script measures from the simulated paired comparison data were estimated with FACETS software (Linacre, 1987). The true reliability was taken to be the square of the correlation between the generating script measures (i.e. those used in the simulation to generate the data) and the estimated measures (i.e. those estimated by FACETS from the data). This is in accordance with the definition of reliability in classical test theory where the square of the correlation between true and observed scores is the ratio of true to observed variance, i.e. reliability.

[8] See https://www.nomoremarking.com/ and https://github.com/NoMoreMarking/cj

[9] For the Rasch model, the information is at a maximum when the current estimate of person ability and item difficulty are the same. By extension to ACJ the information is at a maximum when two scripts with the same current estimate of quality are compared.

Results of simulation study 1

Table 2 shows the main results of the different studies and Figure 1 plots the estimated reliability (SSR) against the true reliability. It is clear from Figure 1 that adaptivity (or at least, Swiss adaptivity) tends to inflate the SSR coefficient above its true value. The effect is more dramatic for lower numbers of comparisons per script, but still detectable with 30 comparisons per script. Since most of the published CJ/ACJ studies have used between 10 and 20 comparisons per script, this suggests that some of the high values of SSR reported may be overestimates of the true reliability. It is also interesting to note that even random allocation slightly overestimated reliability for 30 or fewer comparisons per script, and that the fixed allocation method was where the SSR was closest to the true reliability.

Note that in a proper simulation study, each analysis would be repeated a large number of times with different randomly generated data and/or different samples from the same data. This is possible in future work but since the aim of this study was not to generate bootstrap estimates of the variability of the SSR it seemed reasonable to take the results from each single run at face value – i.e. as being generally representative of other runs with those ‘design parameters’.

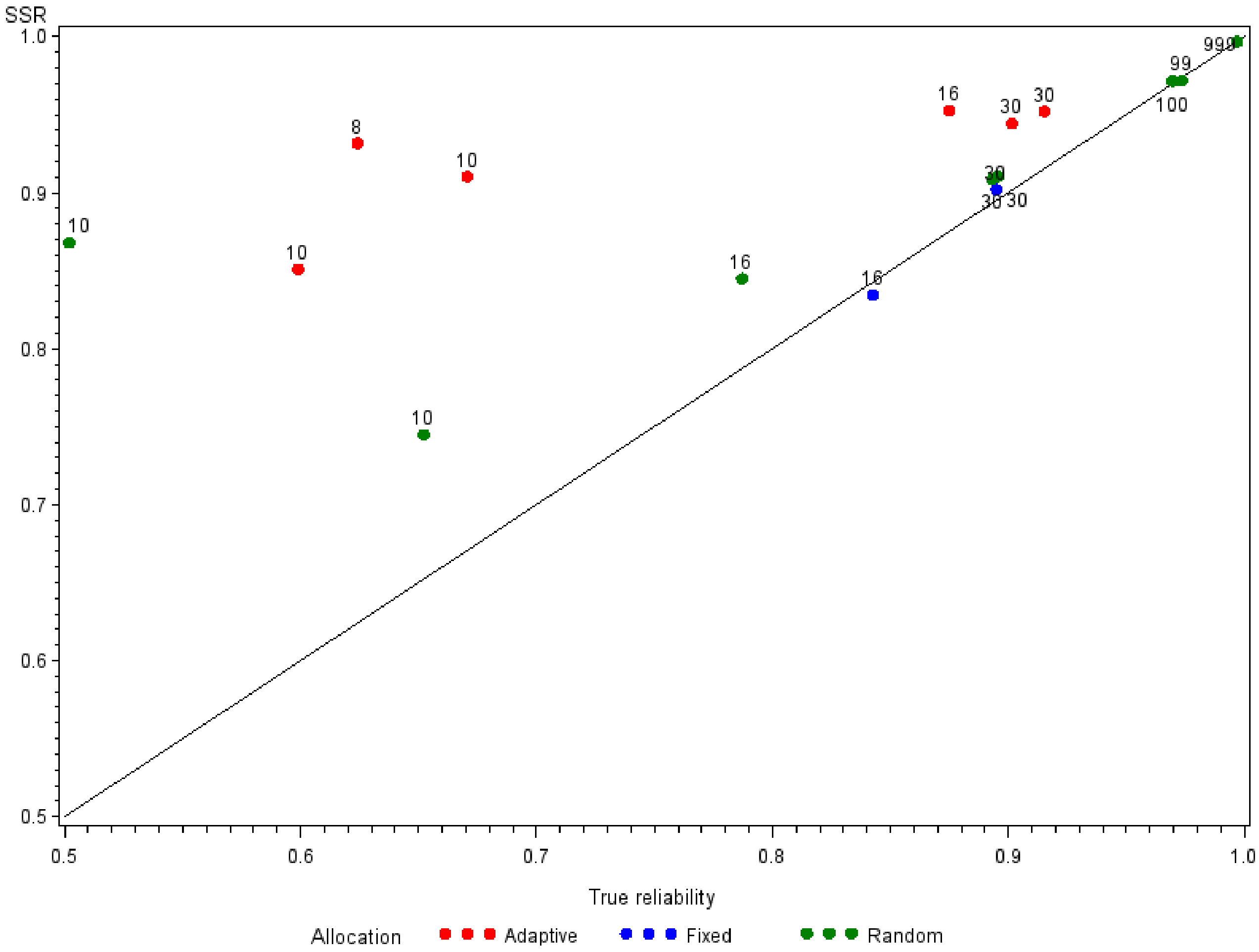

Figure 1. SSR plotted against true reliability for each simulation. The numbers by each point indicate the number of comparisons per script. The black line shows where SSR=True reliability.

Table 2. Results of simulation study 1.

							Generating (true) measures				Estimated measures			
Source	Allocation	N scripts	Comps Per Script	Random Rounds	Adaptive Rounds	N estimated[10]	SD	Min	Max	Reliability	SD	Min	Max	SSR
sim1	Random	100	99	99	0	100	1.70	-3.49	4.81	0.97	1.81	-3.43	4.76	0.97
sim1b	Random	100	10	10	0	85	1.44	-3.49	3.17	0.50	2.92	-7.16	6.09	0.87
sim1c	Random	100	30	30	0	97	1.61	-3.49	4.81	0.89	1.92	-4.36	4.74	0.91
sim1e	Adaptive	100	8	1	7	94	1.52	-3.49	3.17	0.62	5.25	-14.58	7.44	0.93
sim1h	Adaptive	100	10	5	5	90	1.50	-3.49	4.81	0.67	3.51	-7.36	7.17	0.91
sim1i	Adaptive	100	30	20	10	100	1.70	-3.49	4.81	0.90	2.24	-6.37	6.04	0.94
sim2a	Random	1000	999	999	0	1000	1.70	-5.39	5.25	1.00	1.71	-5.38	5.58	1.00
sim2b	Random	1000	10	10	0	871	1.46	-3.96	4.50	0.65	1.98	-5.56	4.64	0.75
sim2c	Random	1000	30	30	0	989	1.66	-4.95	4.54	0.90	1.87	-5.03	5.17	0.91
sim2d	Random	1000	100	100	0	1000	1.70	-5.39	5.25	0.97	1.77	-6.09	5.29	0.97
sim2e	Adaptive	1000	10	5	5	947	1.56	-4.33	4.11	0.60	4.39	-23.35	20.26	0.85
sim2f	Adaptive	1000	30	20	10	1000	1.70	-5.39	5.25	0.92	2.23	-8.62	7.04	0.95
sim2g	Random	1000	16	16	0	960	1.60	-4.45	4.50	0.79	1.99	-5.07	5.35	0.85
sim2h	Adaptive	1000	16	1	15	998	1.69	-5.39	5.25	0.87	3.05	-9.59	8.34	0.95
sim2i	Fixed	116	16*	0	0	98	1.55	-2.45	4.02	0.84	1.77	-3.86	3.78	0.83
sim2j	Fixed	130	30*	0	0	99	1.56	-2.45	4.02	0.90	1.63	-2.92	4.14	0.90

[10] Measures cannot be estimated for scripts that won or lost all their comparisons (or that only won or lost to such a script). FACETS extrapolates a value for such scripts but indicates that it was an 'extreme' measure. Such scripts have been excluded from Table 2 and Figure 1. Their inclusion does not affect the interpretation.

* The 100 scripts outside the reference set were compared 16 or 30 times. The scripts in the reference set were compared many more times, but excluded from the analysis.

Study 2 – SD of 'true quality' 0 logits (i.e. random judgments)

The second set of simulations used the same features as simulation study 1, but replaced every paired comparison outcome in the two starting datasets (all possible paired comparisons among 100 or 1000 scripts) with a random outcome (i.e. 1 or 0 with 0.5 probability for every comparison). This is equivalent to assuming every script has the same underlying 'true quality', or that the judges have no discriminating power whatsoever. In such a scenario the true reliability is zero, because there is no true score variance.

Figure 2 shows the resulting values for the SSR. To facilitate comparison with simulation study 1, the source of the data (see Table 2 above) has 's0' appended, to indicate the equivalent comparison but with a true SD of zero. The results are startling – the cases involving adaptivity produced drastically inflated values for SSR, rising as high as 0.89 for the simulation with 1 random round and 15 adaptive rounds.

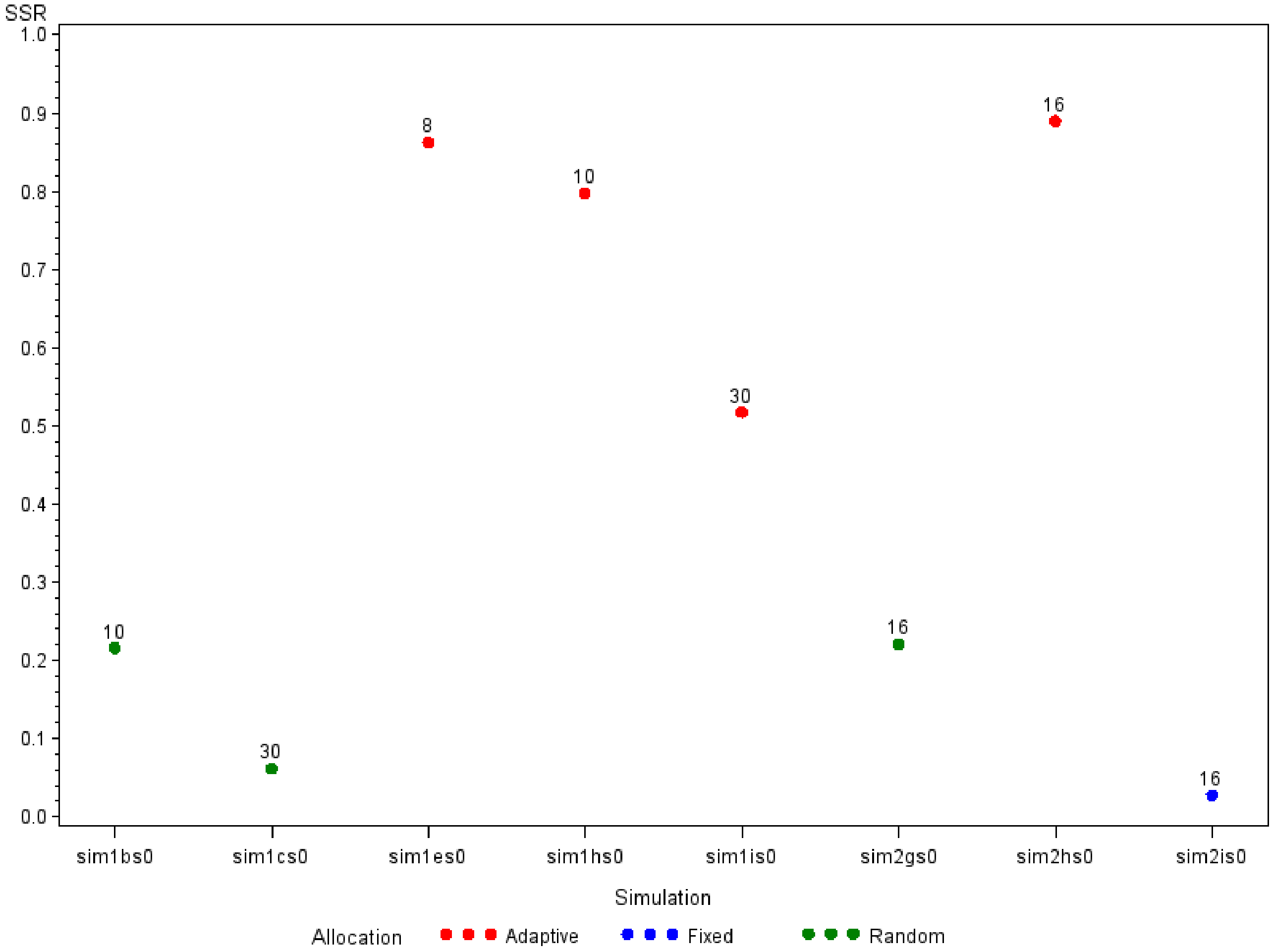

Figure 2. SSR values from simulation study 2 (random data).

To check whether these worrying results were somehow an artefact of using the Swiss method for adaptivity rather than true adaptivity (where parameters are estimated after each round) I contacted Chris Wheadon and Ian Jones at NoMoreMarking, who used a simulation workshop they were running at Antwerp University to explore the issue with their own algorithms. Their results[11] essentially agreed with those above.

[11] See 'The opposite of adaptivity' on the NoMoreMarking website blog: https://www.nomoremarking.com/blog

The implication is that the SSR statistic is worthless as an indicator of the quality / consistency / reliability of a scale constructed from an ACJ study that has involved a significant number of adaptive rounds. With random or fixed rounds, the SSR values were all below 0.25, which would generally be taken as an indicator of failure to create a meaningful scale. To put it another way, a low value of SSR almost certainly indicates low reliability, but a high value of SSR does not necessarily indicate high reliability.

This is not to say that the ACJ studies listed in Table 1 (or indeed other studies not listed) were themselves worthless, just that not much should be made of claims about very high reliability in ACJ studies on the basis of high values of the SSR statistic. In fact, many of those studies also showed moderate to high correlations with variables external to the analysis, which were rightly interpreted as evidence of concurrent validity.

Discussion

The effect of adaptivity on reliability

The adaptivity in ACJ is not quite analogous to the adaptivity in normal computerised adaptive testing (CAT) because in CAT, the items are administered from a pre-calibrated item bank. That is, the item difficulties are known beforehand and treated as fixed – it is only the ability of the person taking them that is being estimated. With ACJ as it has been applied so far, nothing is known about the scripts before judging starts. Therefore all parameters are being estimated ‘on the fly’ on the basis of the judgment data coming in.

The IRT literature (e.g. Tsutakawa & Johnson, 1990; Mislevy, Wingersky & Sheehan, 1994) shows that considerable attention has been given to the problem of how uncertainty in the item calibrations from pre-testing can affect ability estimates in live testing, the typical finding being that uncertainty in the item calibrations (resulting from small calibration samples) leads to bias and underestimation of uncertainty in ability estimates.

In the CAT literature too the issue has been discussed (e.g. Parshall, 2002). Here the problem has been framed in terms of whether or in what circumstances it is acceptable to field-test new items in the course of administering a CAT, where the issues are i) the sparseness of the dataset (most examinees have ‘missing data’ for most items); and ii) the non-randomness of the missing data (items are presented to examinees based on their performance). Ito & Sykes (1994, p1), using the Rasch model, found that existing item bank difficulties were “not well replicated when difficult items were calibrated using responses from able examinees and easy items were calibrated using responses from less able examinees”.

These considerations show why adaptivity is causing a problem for ACJ as currently implemented – if adaptivity is ‘switched on’ at an early stage then the uncertainty in the estimates is large, being based, in the worst case, on a single random comparison. This would be equivalent to estimating ability based on an item that been calibrated in a pre-test sample containing one person. When there are no ‘true’ differences among the scripts a random half of them lose in the first round. In the second round half of these losers will be paired against the other half (and likewise for the winners) and again a random half of them will lose, and a random half of the winners will win, and thus the estimates of script quality become spread out. However, because of the adaptivity, the scripts that have lost twice will not have the chance to show that they are just as likely to beat scripts that have won twice as they are to beat any others because they will not be paired against scripts that have won twice in the next round. In the context of CAT, Styles & Andrich (1993, p915) note that when there are no off-target comparisons (encounters of a high ability person with an easy item and vice versa) the effect “is to sharpen the scale, that is to stretch the item difficulties because responses that are very unexpected tend to regress the estimates.” The way this is phrased does not make it clear whether they considered this effect to be a good or a bad thing, but in the context of ACJ, where there is so little data per script, it seems clear that it creates spurious separation among the scripts.

Is there a way to retain adaptivity in CJ studies?

The above considerations suggest that adaptivity in CJ studies may only be a viable option if one of the scripts in each paired comparison already has an estimate with low uncertainty. This could be achieved if an ACJ study involving N scripts began by selecting a reference subset of scripts of size R, and making all R(R-1)/2 comparisons among them. Then the remaining N-R scripts could be compared adaptively against scripts in the reference set, starting with an average one and then presenting better or worse scripts from the reference set according to whether the new script was judged better or worse than the average one. This is basically equivalent to attempting to form a well-calibrated item bank from a subset of scripts, and then administering an adaptive test based on it.

For example, with 100 scripts a reference set of 30 would require 435 comparisons for each to be compared with every other script once. Then if the remaining 70 scripts were compared adaptively with scripts from that set of 30 (anchoring each reference script at its pre-calibrated value) we might find that 15 adaptive comparisons gave a reasonable estimate of the quality of the remaining 70 scripts. This would involve a further 1050 comparisons, giving 1485 in total, or 30% of the total number of possible comparisons. Simply comparing each of the 70 scripts with each of the reference set of 30 would require 2100 comparisons, or 42.4% of the total.

Clearly further work is needed to explore how many comparisons would be needed to calibrate the scripts in the reference set properly. The above example assumes 29 comparisons, which seems very optimistic given the guidance in the CAT literature for 1,000 persons to calibrate an item with the 3-parameter model (Wainer, 2000). However, it is recognised that the Rasch model does not generally require such large samples (e.g. Lord, 1983), and it does not seem unreasonable to hope that expert judges comparing relatively substantial pieces of work might be able to produce more discrimination than dichotomous test items.

What is reliability in CJ/ACJ studies?

Reliability theory is well established in Classical Test Theory (CTT), Generalisability Theory (GT) and Item Response Theory (IRT) (see for example Haertel, 2006; Brennan, 2011; Mellenburgh, 1996). The general idea is that reliability is about what would happen if an assessment process were repeated, varying one or more aspects of it. In the CJ/ACJ context, the SSR clearly cannot give any information about what would happen if the exercise was repeated with different work from the same examinees (e.g. if the test was made longer or shorter, or contained different questions). It is therefore not comparable to Cronbach’s Alpha.

However, another way a CJ/ACJ study could be repeated would be to present exactly the same comparisons to a different set of judges, and then correlate the script measures estimated in each case. As far as I am aware, this has never been tried. Along similar lines, a study could be repeated with a different set of judges using the same set of scripts and allocation method (but not necessarily repeating the exact comparisons). Although the SSR statistic for each set of judges would be biased upwards if there had been adaptive allocation, the correlation of estimated measures between each set of judges would be an unbiased estimate of reliability regardless of whether adaptivity had been involved. This approach was taken by Jones & Alcock (2014) and Jones, Swann & Pollitt (2015) and yielded estimates of reliability ranging from 0.72 to 0.87. Given that both studies involved judgments of responses to maths tasks it can be seen that the reliability was not as high as the inter-marker reliability that would be obtained from conventional marking of this type of task, which is usually above 0.95 (e.g. Bramley & Dhawan, 2012, p276 & p283). However, if similar values could be obtained for other types of task (e.g. judgments of portfolios; essays; works of art or musical performances) this would be evidence that CJ/ACJ could yield similar levels of reliability to conventional marking.

A different interpretation of repeating a CJ/ACJ study would be obtain a set of estimated measures for the same set of scripts by getting the same judges to compare them with a different set of scripts from the same population. For example, if each of the N scripts had been compared on average with P scripts (with the P being a subset of the N) in the original study, the replication could compare each of the N with R=P further scripts (with the R not being among the original N) in a fixed allocation design. The correlation between the estimated measures of the N scripts in each setting would also be an indication of reliability.

In summary, when there is no adaptivity the SSR statistic estimates the true reliability (defined as the squared correlation between ‘true’ (generating) quality and estimated quality when the data fits the model). In these non-adaptive CJ contexts, the size of the SSR statistic will mainly depend on the number of scripts each script is compared with, and the ‘true’ spread of measures (=discrimination power of the judges) in the sample of scripts. If there is adaptivity, then it is necessary to repeat the comparison process in some way to get an unbiased estimate of the true reliability.

Conclusion

The motivation for using adaptivity in CJ studies – namely to avoid wasting time and resource getting judges to make comparisons whose outcome is a foregone conclusion – is entirely understandable. But both theoretical considerations from the IRT and CAT literature and the results of the simulation study in this report show that adaptivity produces spurious scale separation reliability, as indicated by values of the SSR coefficient that are substantially biased upwards from their true value. The higher the proportion of adaptive rounds, the greater the bias. SSR values above 0.70 and even as high as 0.89 can be obtained from random judgments (or equivalently, scripts that all have the same true quality) when adaptivity is ‘switched on’ after the first random pairing.

The conclusion is therefore that the SSR statistic is at best misleading and at worst worthless as an indicator of scale reliability *whenever a CJ study has involved a significant amount of adaptivity*. Other indicators of reliability, such as correlations with measures obtained from comparisons among a different group of judges, or correlations with relevant external variables, should be used instead. ACJ studies that have used high values of the SSR coefficient alone to justify claims that ACJ produces a more reliable scale than conventional marking need to be re-evaluated.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9*(1), 95-104.

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

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

Bramley, T., & Dhawan, V. (2012). Estimates of reliability of qualifications. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp. 217-319). Coventry: Office of Qualifications and Examinations Regulation.

Bramley, T., & Oates, T. (2011). Rank ordering and paired comparisons - the way Cambridge Assessment is using them in operational and experimental work. *Research Matters: A Cambridge Assessment Publication, 11*, 32-35.

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education, 24*, 1-21.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 65-110): ACE/Praeger series on higher education.

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1-20.

Humphry, S., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement, 9*(3), 249-264.

Ito, K., & Sykes, R. C. (1994). *The Effect of Restricting Ability Distributions in the Estimation of Item Difficulties: Implications for a CAT Implementation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgment. *International Journal of Science and Mathematics Education, 13*(1), 151-177.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., Pollitt, A., Whitehouse, G. (2009). *E-scape portfolio assessment phase 3 report*. London: Goldsmiths, University of London.

Linacre, J. M. (1987). FACETS (Version 3.67.1): www.winsteps.com.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 51-61). New York: Academic Press.

McMahon, S., & Jones, I. (2014). A comparative judgment approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, 1-22. doi: 10.1080/0969594x.2014.978839

Mellenburgh, G. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293-299.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: expected response functions. *ETS Research Report Series, 1994*(1), i-20.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice, 21*(2), 205-220.

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 119-142): Routledge.

Pollitt, A. (2004). *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, June 2004.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.

Pollitt, A. (2012b). The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy and Practice 19(3), 387.

Styles, I., & Andrich, D. (1993). Linking the Standard and Advanced forms of the Raven's Progressive Matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53*(4), 905-925.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390.

Wainer, H. (Ed.). (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Whitehouse, C. and Pollitt, A (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment.* Manchester: AQA Centre for Education Research and Policy.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.