Learning in practice

Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study

I C McManus, Tom Lissauer, S E Williams

Department of Psychology, University College London, London WC1E 6BT
I C McManus
professor of psychology and medical education

Examinations Department, Royal College of Paediatrics and Child Health, London W1W 6DE
Tom Lissauer
officer for examinations
S E Williams
psychometrician

Correspondence to:
I C McManus
i.mcmanus@ucl.ac.uk

BMJ 2005;330:1064–6

Abstract

Objective To assess whether a computer program using a variant of Angoff's method can detect anomalous behaviour indicative of cheating in multiple choice medical examinations.
Design Statistical analysis of 11 examinations held by the Royal College of Paediatrics and Child Health.
Setting UK postgraduate medical examination.
Participants Examination candidates.
Main outcome measures Detection of anomalous candidate pairs by regression of similarity of correct answers in all possible pairs of candidates on the overall proportion of correct answers. Anomalous pairs were subsequently assessed in terms of examination centres and the seating plan of candidates, to assess adjacency.
Results The 11 examinations were taken by a total of 11 518 candidates, and Acinonyx examined 6 178 628 pairs of candidates. Two examinations showed no anomalies, and one examination found an anomaly resulting from a scanning error. The other eight examinations showed 13 anomalies compatible with cheating, and in each pair the two candidates had sat the examination at the same centre, and for six examinations with seating plans, the candidates in the anomalous pairs had been seated side by side. The raw probabilities of the anomalies varied from 3.9×10^{-11} to 9.3×10^{-30} (median = 1.1×10^{-17}), with Bonferroni-corrected probabilities in the range 2.4×10^{-5} to 4.1×10^{-24} (median = 1.6×10^{-11}). This suggests that one anomalous pair is found for every 1000 or so candidates taking this postgraduate examination.
Conclusions This statistical technique identified a small proportion of candidates who had very similar patterns of correctly answered questions. The likelihood is that one candidate has copied from the other, or that there was collusion, or that a technical error occurred in the exams department (as happened in a single case). Analysis of similarities can be used to identify cheating and as part of the quality assurance process of postgraduate medical examinations.

Introduction

> "Ninety-two coins spun consecutively have come down heads ninety-two consecutive times . . . One, probability is a factor which operates within natural forces. Two, probability is not operating as a force. Three, we are now within un-, sub- or super-natural forces. Discuss."
> Tom Stoppard, *Rosencrantz and Guildenstern are Dead*

Cheating occurs at all levels of education,[1–3] in medicine and elsewhere,[4–8] and postgraduate examinations are unlikely to be exempt. Cheating threatens examination validity and thereby health care. However, conventional invigilation is only partially effective in preventing cheating.[1] This paper describes Acinonyx, a computer program which adapts Angoff's validated method for identifying unduly similar answers from pairs of candidates.[9] Reasons for excessive similarity include copying and spontaneous or premediated collusion between candidates, perhaps supported by communications technology. Acinonyx cannot distinguish these processes, or determine which candidate has copied from which.

Method

Software—Acinonyx is written in *C++* and also uses the REGRESSION program of SPSS to implement a version of Angoff's *A* index.[9] It is applicable to any objectively marked examination (multiple true-false with or without negative marking; best of five; extended matching; etc), requiring only a knowledge of the questions answered correctly by each candidate.

Statistical method—Let candidate *I* answer R_i questions correctly in an exam, candidate *J* answer R_j questions correctly, and R_{ij} be the number of correct answers shared by the two candidates. R_{ij} is not a good measure of similarity because the number of similar answers increases with examinee knowledge. Acinonyx follows Angoff[9] in examining R_{ij} in relation to R_i and R_j, but assesses the unusualness of R_{ij} by calculating the residual of R_{ij} after regression on $\sqrt{(R_iR_j)}$ and R_iR_j. Residuals are distributed normally and expressed as probabilities.

Significance testing—With N candidates there are $N\times(N-1)/2$ pairs of candidates (that is, 1 999 000 pairs when N = 2000), making necessary a correction for alpha inflation (multiple significance testing). Acinonyx calculates a raw, uncorrected probability, P_{raw}, which is adjusted for multiple testing by a Bonferroni correction, giving a corrected probability, $P_{corrected}$:

$$P_{corrected} = P_{raw} \times N.(N-1)/2$$

An extended version of the paper and information on power calculations and Monte Carlo simulation are on bmj.com

Alpha is set conservatively at $P<0.001$, and a pair regarded as anomalous if $P_{corrected}$ is <0.001. Therefore, for 2000 candidates a pair is anomalous if P_{raw} is $<0.001/1\,999\,000 = 5.0\times10^{-10}$.

Centres and seating plans—Acinonyx does not know the seating of candidates (and many are sitting in different centres). Seating plans are used to check the validity of apparently anomalous pairings.

Examination—Until December 2003, Part 1 of the Membership Examination of the Royal College of Paediatrics and Child Health (the MRCPCH) consisted of a single paper. Since 2004 there are two papers, Paper One A (basic child health) and Paper One B (extended paediatrics). All examinations are marked by computer and do not use negative marking.

Results

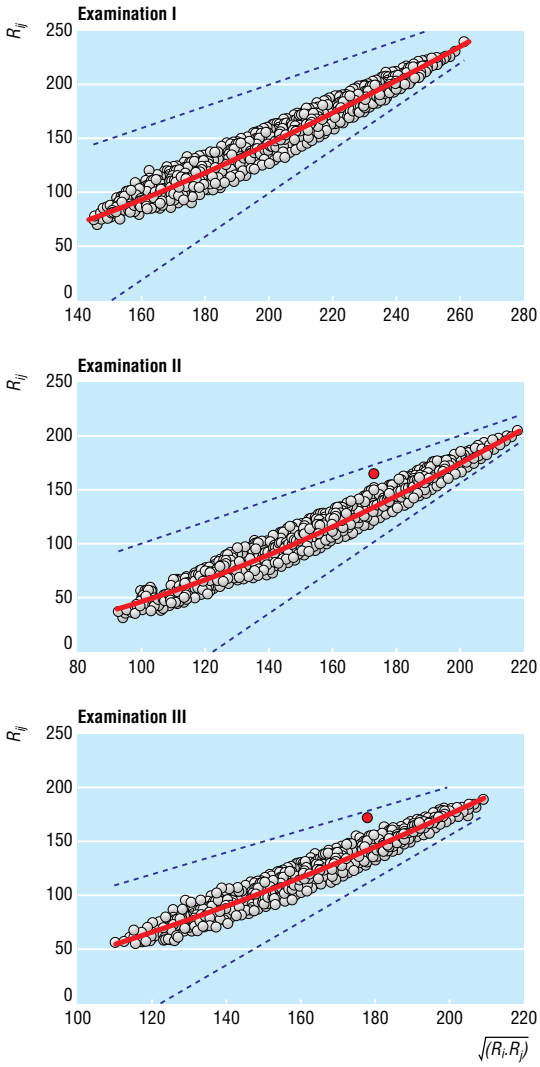
Here we describe three examinations; further examples can be found on bmj.com.

Examination I—MRCPCH Part 1, 2003/2, had 300 questions and was taken by 1099 candidates. The 63 351 pairs of candidates are plotted in the figure, with R_{ij} plotted vertically and $\sqrt{(R_i.R_j)}$ horizontally. The regression explains 95% of the variance ($R=0.975$), and residuals are normally distributed (see statistical appendix on bmj.com), with a range from -5.400 to 5.528, corresponding to raw, one tailed probabilities of 3.3×10^{-8} and 1.6×10^{-8} and corrected probabilities of 0.02 and 0.01, which do not reach the criterion of $P_{corrected}<0.001$. This examination showed no anomalies, and shows that residuals are normally distributed.

Examination II—MRCPCH 2004/2 Paper One A had 244 questions and was taken by 1298 candidates. The figure shows the 841 753 pairs of candidates; one pair, shown in red, has a standardised residual of 8.6, a raw probability of 1.1×10^{-17}, and a corrected probability of 9.0×10^{-12}, and hence $P_{corrected}<0.001$. These two candidates, who answered 170 and 178 items correctly, with 164 shared answers, were found on the seating plan to have been seated side by side; one passed and one failed. The latter subsequently took Parts 1A and 1B of the 2004/3 diet (941 and 1084 candidates), and was in the only anomalous pair in each of these examinations ($P_{corrected}=4.1\times10^{-24}$ and 3.7×10^{-21}; see bmj.com).

Examination III—Paper One B of MRCPCH 2004/2 had 244 questions and 1251 candidates. One of the 781 875 candidate pairs (figure) had a standardised residual of 7.8 ($P_{corrected}=7.1\times10^{-9}$). The computer file showed 177 and 180 correct items, with 172 shared answers. However, the candidates sat the examination in different cities. Questions are answered on a single response sheet with the 200 multiple true-false questions and 44 other questions scanned separately and the data sets then merged. The first 200 answers were identical. The actual answer sheets showed a scanning error had resulted in one answer sheet inadvertently being entered twice.

Overall results—We analysed 11 consecutive MRCPCH Part 1 examinations (2002/2 to 2004/3 A and B), which were taken by 11 518 candidates, comprising 6 178 628 candidate pairs. Seating plans were available only for the year 2004. One anomalous pair resulted from an administrative error, whereas 13 anomalous pairs were compatible with cheating (one pair for every 886 candidates), although two anomalous pairs consisted of the same two candidates. $P_{corrected}$ values for anomalies were in the range 2.4×10^{-5} to 4.1×10^{-24} ($N=13$; median $=1.6\times10^{-11}$). In the six exams where seating plans were available the candidates in each anomalous pair had been seated side by side. Of the 12 independent pairs, both candidates failed in seven cases, one passed in three cases, and both passed in two cases.



Plot of R_{ij} against $\sqrt{(R_i.R_j)}$ for examinations I (top), II (centre), and III (bottom). The red line shows the fitted quadratic regression. The diagonal straight lines are the maximum and minimum amounts of agreement that are theoretically possible. The points reaching a Bonferroni corrected criterion of P<0.001 are shown in red

Discussion

Acinonyx identifies anomalous pairs of candidates which require investigating (and meet standard forensic requirements for scientific evidence[10]). Action requires other evidence. Seating plans, notes in question booklets, changed answers, information from invigilators, other surveillance, and interviews with candidates may show culpability.

Examiners raise a number of questions and objections about Acinonyx that are worth considering (see also bmj.com).

Statistical issues—Although "the evidence is only statistical," statistics are facts and are widely used to guide actions throughout medicine. Although rare events do occur by chance, particularly with large numbers of candidates, Examination I shows that the method effectively eliminates type I errors. The extreme unlikelihood of some of the probabilities is sometimes difficult to interpret, and is better expressed in terms of games of chance: 10^{-20}, for instance, is the likelihood of tossing 64 successive heads, or of winning the UK National Lottery in three successive weeks. Additional statistical support also comes from seating plans: for examination II, with 1298 candidates, the probability that the second member of an anomalous pair was one of the eight seated adjacent to the first is only 1 in (1297/8) = 1 in 162, P = 0.006.

Candidates may give similar answers because they have studied together—If so then anomalous pairs would be found in candidates sitting in different centres, but they are not, here or elsewhere.[3]

The evidence is only circumstantial—"The rule of probability"[11] means that circumstantial evidence can be highly probative, particularly when corroborated by seating plans, coincidences in wrong answers in best of five and extended matching questions, answers erased in favour of another answer, annotation of question booklets, performance in previous examinations, and evidence from invigilators and other candidates.

The sensitivity, specificity, and validity of the technique are not known—Angoff demonstrated that his indices were substantially raised in 50 "known and admitted copiers."[9] Monte Carlo analysis confirms the sensitivity and specificity of Acinonyx (see bmj.com).

Postgraduate examinations should take other steps to prevent cheating—Measures to minimise cheating by close investigation, avoiding tiered lecture theatres or closely placed desks, and other methods should be taken. Acinonyx can itself be used to monitor the effectiveness of prevention.

It is a "victimless crime"—The victims are patients treated inappropriately by improperly qualified doctors. In the United Kingdom, cheating violates the guidelines in *Good Medical Practice* ("as a doctor you must be honest and trustworthy"),[12] and the General Medical Council has already disciplined a doctor for cheating, partly on the basis of statistical evidence.[13]

Conclusions

Acinonyx identifies anomalous pairs of candidates who are probably cheating, and also acts as a quality control, shown by detecting a scanning error which is so far unique to this and other examinations. Acinonyx does not provide direct evidence about which candidate did the copying, and further investigation is required. Although examining bodies dislike investigating innocent candidates, their obligations to other candidates, the profession, and patients require them to protect the integrity of examinations. Examining bodies should remind candidates of the importance of keeping answers concealed, and bodies that adopt such statistical methods should inform candidates about their use and should have appropriate investigative and disciplinary procedures in place.[14]

We are grateful to a number of colleagues who have discussed the ideas and analyses presented in this paper, and particularly to Mr Bertie Leigh for his detailed comments on the manuscript. The software is available free of charge to non-commercial and educational organisations from i.mcmanus@ucl.ac.uk.

Contributors: ICM developed the statistical analysis. Applying it to this examination was initiated by TL and SEW, with SEW assisting with statistical analysis. ICM wrote the first draft of the paper, and all authors contributed to the final draft. ICM is the guarantor for the paper.

Funding: None.

Competing interests: ICM has given unpaid educational advice to several postgraduate medical examination boards. TL is the honorary officer for examinations for the Royal College of Paediatrics and Child Health. SEW is a paid employee of the RCPCH examinations department.

Ethical approval: Not required.

What is already known on this topic

Cheating is common in examinations

Angoff's method is a validated technique for detecting copying in multiple choice examinations

What this study adds

Acinonyx, a computer program incorporating a modified Angoff's method, finds undue similarity ("anomalies") between pairs of candidates taking a postgraduate medical examination

Anomalous pairs of candidates are seated adjacent to one another, and the similarity probably results from copying

About one anomalous pair is found for every 1000 candidates taking postgraduate examinations

1 Cizek GJ. Cheating on tests: How to do it, detect it, and prevent it. Mahwah, NJ: Erlbaum, 1999.
2 Bay L. Factors associated with cheating among college students: a review. *Res Higher Educ* 1998;39:235-74.
3 Wesolowsky GO. Detecting excessive similarity in answers on multiple choice exams. *J Applied Statistics* 2000;27:909-21.
4 Sierles F, Hendrickx I, Circle S. Cheating in medical school. *J Med Educ* 1980;55:124-5.
5 Stimmel B,.Yens D. Cheating by medical students on examinations. *Am J Med* 1982;73:160-4.
6 Carlova J. How many doctors are cheating their way into practice? *Med Econ* 1984;61:84-92.
7 Rozance CP. Cheating in medical schools: implications for students and patients. *JAMA* 1991;255:2453-6.
8 Baldwin D-C Jr, Daugherty SR, Rowley BD, Schwarz MD. Cheating in medical school: a survey of second-year students at 31 schools. *Acad Med* 1996;71:267-73.
9 Angoff WH. The development of statistical indices for detecting cheaters. *J Am Stat Assoc* 1974;69:44-9.
10 Merwin SE, Moeller DW, Kenedy WE, Moeller MP. Application of the Supreme Court's *Daubert* criteria in radiation litigation. *Health Physics* 2001;81:670-7.
11 Nizer L. *My life in court*. New York: Doubleday, 1961.
12 General Medical Council. *Good medical practice*. London: GMC, 2001.
13 General Medical Council Professional Conduct Committee. Professional conduct hearing: Purmessur, Shyam Nundun Singh, 6-9 October 2003. www.gmc-uk.org/probdocs/decisions/pcc/2003/PURMESSUR_20031009.htm (accessed 8 april 2005).
14 Stearns SA. Administrative ramifications of student cheating. *J Assoc Communication Admin* 1997;(2):133-9.

(Accepted 22 March 2005)

Endpiece

Anticipation

Somewhere something incredible is waiting to be known.

Carl Sagan, American astronomer and author (b 1934)

Fred Charatan, retired geriatric physician, Florida

Supplementary information

With any diagnostic procedure for detecting anomalies there is a need to assess the sensitivity and specificity of the method, or, to put it in more statistical terms, to assess the power and the false positive rate. Although that is relatively straightforward for most diagnostic tests, since 'gold-standard' cases are relatively easy to define and identify, the process is not so straightforward when detecting cheating since it is difficult to identify individuals with certainty. Instead therefore one can use a Monte Carlo procedure to validate the statistical method, and to assess its specificity and sensitivity.

Specificity / False positive rate.

A potential problem of the statistical analysis is that it relies on the residuals being normally distributed, so that the probability of extreme values can be calculated. However whereas an assumption of multiple regression is that the data points are statistically independent, in the present analyses that is not strictly the case, since although there are N.(N-1)/2 data points in the regression, these are based on only N strictly independent cases. It is therefore possible, albeit unlikely, that the data points in the regression analysis are not independent, and that as a result some of the anomalous pairs occur due to statistical artefacts. In this section I use a Monte Carlo simulation of the process to show that that is not the case, that the residuals are well behaved and in accordance with normal distribution theory, and that the false positive rate is as would be expected.

Monte Carlo analysis. Data from multiple-choice examinations such as those analysed with *Acinonyx* are generally fitted well by means of a three-parameter item-response theory model, in which candidates differ in terms of their true ability (theta), and items differ in terms of their discrimination (a), difficulty (b), and chance-guessing probability (c) [1]. The program

PARDSIM [2] was used to simulate a typical examination in where there were 1000 candidates, with ability theta normally distributed with mean zero and standard deviation (SD) of 1. The examination had 200 items, which were regarded as being 'best-of-five', so that the guessing probability was of the order of 0.2. In particular discrimination parameters were normally distributed, with mean of .31 and SD of .15, the difficulty parameters were uniformly distributed in the range -3 to +3, and the chance guessing probabilities were normally distributed with mean of 0.2 and SD of .025. These parameters give results which are typical of many post-graduate examinations, with mean score of about 120 correct (60%), with candidates' marks in the range 70 to 170 (35% to 85%), and a reliability of about 0.85 (coefficient alpha).

A total of 50 Monte Carlo replications was carried out[1], the candidate and item parameters being sampled afresh for each replication. Table S1 summarises the overall results of the simulations. The key result concerns the extreme values of the standardised residuals, which are expressed as z-scores. The largest positive residual had a mean value of 4.761 (SD .276), and the smallest negative residual had a mean value of -4.724 (SD .218); the absolute values of these residuals do not differ significantly, given the 50 replications, and show that the distribution of residuals is symmetric. The largest absolute residual has a mean value of 4.879 (SD .247). The two-tailed probability can be calculated for the largest absolute z-score found in each replication. The arithmetic mean probability is .000001842 (1.84×10^{-6}), with a range of .000014776 to .00000000721 (1.4×10^{-5} to 7.2×10^{-9}).

The 1000 candidates in the simulated examination produce 1000 x 999/2 = 499,500 pairs of candidates. On average therefore it will be expected that the most extreme pair will

[1] Only 50 replications were carried out because the procedure was slow and tedious. The *Acinonyx* program takes about 15 minutes to generate half-a-million pairs from the data, and the SPSS program takes another10 minutes or so to analyse the data. A complete replication therefore takes about half an hour of computer time, meaning that the 50 replications took about 25 hours. This was particularly fiddly since each replication involved running *Notepad*, PARDSIM, SPSS, *Acinonyx*, and SPSS once again in sequence, with each having to be controlled from the keyboard.

have a probability of 1 / 499,500 = 0.000002002 (2.0×10^{-6}), a value which is consistent with the mean probability of .000001842 (1.84×10^{-6}). Using an alpha level of P<.001 and a Bonferroni correction, as in the main paper, would require a significance level of .001/499,500 = 0.000000002002 (2.0×10^{-9}); it is noteworthy that in none of the fifty replications does the largest residual reach this value. The specificity using an alpha level of P<.001 is therefore 100%, for these 50 replications. The 95% confidence interval for the specificity is 95.3% to 100%. A somewhat more conventional (but in this context too liberal) alpha level of 0.05 would require a Bonferroni corrected significance level of .05/499,500 = 0.0000001001 (1.0×10^{-7}), and of the 50 replications, two (4%) reach that level, one for a negative residual and one for a positive residual, a proportion close to the expected proportion of 5%.

Taken overall it can be concluded from the Monte Carlo analysis that the residuals are distributed normally, and that the Bonferroni correction is appropriate. The specificity of the test is between 95% and 100% [2].

Sensitivity / Power analysis

The sensitivity of Acinonyx can be assessed by generating data in which all of the candidates are strictly independent apart from a single candidate who copies a proportion of answers from another candidate. That process can be replicated many times for different proportions of questions being copied.

Monte Carlo analysis. PARDSIM was used to generate a similar set of data to that analysed earlier in the specificity analysis, with 1000 candidates (theta normally distributed with mean zero and standard deviation (SD) of 1), and 200 best-of-five items (discrimination: normally distributed with mean of .31 and SD of .15; difficulty uniformly distributed in the range -3 to

[2] Calculated using the Bayesian calculator at http://www.causascientia.org/math_stat/ProportionCI.html

+3; guessing probabilities normally distributed with mean of 0.2 and SD of .025). These were used to generate the 1000x999/2 = 499,500 pairs of candidates. The item parameters were saved and used in the generation of the anomalous pairs, who were therefore answering identical questions to the other candidates. For a particular rate of copying, 1000 candidates 'stronger' candidates were generated, who were 1 SD above the population mean (i.e. a typical candidate performing above the mean). 1000 'weaker' candidates were also generated by a similar procedure, who were 1 SD below the population mean, and each was paired with one of the stronger candidates. Each of the 1000 weaker candidates then copied a percentage, C, of the stronger candidate's answers[3]. Separate replications, each involving 1000 pairs of candidates were carried out with C varying by five point steps from 0 to 100%.

The 1000 anomalous pairs were then added to the database containing the 499,500 pairs who were not anomalous. The SPSS regression program calculated the standard *Acinonyx* regression equation based only on the non-anomalous pairs, and then the standardised residuals calculated for all pairs[4]. For each of the 1000 anomalous pairs, the program then calculated the standardised residual, the raw significance of the residual, and whether or not the residual was significant after Bonferroni correction with $p<.001$ (i.e. the raw significance level was less than .001/499,500 = 0.000000002002 (2.0×10^{-9}).

Table S2 summarises the results of the specificity analysis in a single situation which is probably typical of an averagely weaker candidate copying from an averagely stronger candidate. It can be seen that there is an 80% power when a candidate is copying about 61% or more of questions, and a 95% power when a candidate is copying 65% or more of the

[3] Of course some of these answers would have been identical to what the weaker candidate would have responded anyway, and in other cases the weaker candidate will have replaced a correct answer with a wrong one.

[4] Note that in effect this is equivalent to introducing a single anomalous pair on 1000 separate occasions, but is computationally far more efficient. The omission of a single anomalous pair from the distribution of residuals has only an infinitesimal effect on the mean or standard deviation of the residuals.

questions. With levels of copying below about 45% then few anomalies are detected[5].

Taken overall it is clear that as, as expected, the sensitivity of *Acinonyx 1* increases with an increasing proportion of copied answers. Experience with Acinonyx in practice certainly finds cases where 90% of the answers of two candidates are identical, and the method can readily detect such similarities with 100% power.

Summary

These Monte Carlo analyses confirm that Acinonyx 1 with a $p<.001$ criterion has a specificity of at least 95%, and a typical sensitivity of 80% when a weaker candidate has copied 65% or more of their answers from a stronger candidate.

[5] It is perhaps worth pointing out that these results do not mean that candidates can copy less than 40% of questions with alacrity. The sensitivity of *Acinonyx* varies somewhat with the ability levels of the candidates, and the program is substantially more sensitive if other statistics are also included, such as similarity of incorrect best-of-five answers and runs of identical answers, and if the Bonferroni correction is adjusted to take into account geographical propinquity. Those aspects of *Acinonyx* will be reported further in a subsequent paper.

Table S1: Specificity analysis. Results of 50 simulations of data with no copying between candidates. For details of parameter specifications see text.

	Mean	SD	se	Min	Max
Reliability (Cronbach's alpha)	0.8544	0.0105	0.0015	0.8347	0.8766
Mean percentage score on exam	60.0	1.3	0.2	56.8	63.1
SD of percentage score on exam	8.49	0.29	0.04	7.92	9.08
Minimum score on exam	34.4	2.4	0.3	26.5	39.5
Maximum score on exam	84.8	2.3	0.3	80.0	88.5
Minimum standardised residual	-4.724	0.218	0.031	-5.481	-4.315
Maximum standardised residual	4.761	0.276	0.039	4.209	5.786
Proportion of replications significant with P<.05 after Bonferroni correction	4%	-	-	-	-
Proportion of replications significant with P<.001 after Bonferroni correction	0%	-	-	-	-
Uncorrected significance of most extreme residual	0.0000018425	0.0000022867	0.0000003234	0.0000000072	0.0000150000

Table S2: Sensitivity analysis. Mean residual for 1000 replications with different proportions of answers being copied by a weaker candidate from a stronger candidate. The final column shows the proportion of cases which reach the criterion of $P_{corrected} < .001 = .001/499{,}500 = 0.000000002002$ (2.0×10^{-9}).

Percentage of answers copied	Mean standardised residual	SD of Standardised residuals	Mean raw significance level of standardised residuals	Percentage of cases significant with P<.001 after Bonferroni correction
0	-0.169	0.941	5.20×10^{-1}	0
5	0.289	0.927	5.15×10^{-1}	0
10	0.867	0.952	4.03×10^{-1}	0
15	1.451	0.924	2.57×10^{-1}	0
20	2.033	0.934	1.31×10^{-1}	0
25	2.613	0.944	5.53×10^{-2}	0
30	3.178	0.943	1.99×10^{-2}	0.1
35	3.759	0.942	5.90×10^{-3}	0.6
40	4.357	0.937	1.43×10^{-3}	4.8
45	4.953	0.878	1.50×10^{-4}	11.1
50	5.521	0.916	4.10×10^{-5}	30.8
55	6.098	0.865	4.26×10^{-6}	56
60	6.655	0.81	1.14×10^{-7}	78.7
65	7.249	0.795	1.31×10^{-8}	94.3
70	7.882	0.807	2.00×10^{-10}	98.4
75	8.477	0.797	4.56×10^{-10}	99.6
80	9.051	0.756	3.11×10^{-13}	100
85	9.66	0.708	1.77×10^{-16}	100
90	10.22	0.689	1.39×10^{-18}	100
95	10.81	0.639	8.47×10^{-20}	100
100	11.35	0.587	1.98×10^{-23}	100

Reference List

1. Weiss DJ, Yoes ME. Item Response Theory. In Hambleton RK, Zaal JN, eds. *Advances in Educational and Psychological Testing: Theory and Applications*, pp 69-95. Boston,London,Dordrecht: Kluwer Academic Publishers, 1991.

2. Yoes ME. User's manual for the PARDSIM Parameter and Response Data Simulation Program. St Paul, MN: Assessment Systems Corporation, 1997.