Applied Measurement in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Dissecting Knowledge, Guessing, and Blunder in Multiple Choice Assessments

Rashid M. Abu-Ghazalah, David N. Dubins & Gregory M.K. Poon

To cite this article: Rashid M. Abu-Ghazalah, David N. Dubins & Gregory M.K. Poon (2023) Dissecting Knowledge, Guessing, and Blunder in Multiple Choice Assessments, Applied Measurement in Education, 36:1, 80-98, DOI: 10.1080/08957347.2023.2172017

To link to this article: https://doi.org/10.1080/08957347.2023.2172017

APPLIED MEASUREMENT IN EDUCATION
2023, VOL. 36, NO. 1, 80–98
https://doi.org/10.1080/08957347.2023.2172017

Dissecting Knowledge, Guessing, and Blunder in Multiple Choice Assessments

[redacted]

[a]W. Booth School of Engineering Practice and Technology, Faculty of Engineering, McMaster University; [b]Leslie Dan Faculty of Pharmacy, University of Toronto; [c]Departments of Chemistry and Nutrition, Georgia State University

ABSTRACT

[redacted]

1. Introduction

[redacted]

CONTACT Rashid M. Abu-Ghazalah abughar@mcmaster.ca; David N. Dubins d.dubins@utoronto.ca; Gregory M.K. Poon gpoon@gsu.edu Departments of Chemistry and Nutrition, Georgia State University, Atlanta, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/08957347.2023.2172017

Budescu, & Attali, 2005; Diamond & Evans, 1973; Frary, 1989) and have been abandoned by major examinations such as the GRE and SAT (Bennett & von Davier, 2017). Attention has also been paid to more qualitative issues of item design, aimed at reducing the potential for construct-irrelevant variance and improving the identification of knowledgeable responses from lucky guessing (Haladyna, Downing, & Rodriguez, 2002). To this end, many discipline-specific improvements have been proposed (Breakall, Randles, & Tasker, 2019; Moore, Nguyen, & Stamper, 2021; Towns, 2014), including "ordered MC items" which recast distractors in terms of a staged progression in subject mastery (Lazenby, Balabanoff, Becker, Moon, & Barbera, 2021).

1.1. Review of Theoretical Frameworks

The prevailing theoretical framework for controlling guessing in MC responses is item response theory, or IRT. IRT models performance (probability of a correct response, P) on a given MC item as a function of a student's "latent ability" θ using a generalized logistic equation of the form:

$$P(\theta; a, b, c) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}. \tag{1}$$

The parameters a, b, c in Equation (1), define item characteristic curves (ICC), express item difficulty, item discrimination, and guessing as represented by the curve's position along the abscissa, its steepness, and intercept on the ordinate. In the IRT paradigm, knowledge is inferred on a *per-item basis*: a test consists of a set of items each with their own chracteristics, *individually* probing student ability to generate the observed test score.

A different perspective of MC assessment is to parameterize knowledge and other psychometric states directly from a phenomenological analysis of the MC test score (Wang & Calhoun, 1997), rather than the inferred functional characteristics of individual test items. Psychometric models provide a quantitative formulation of the intuition that MC scores integrate the test performance of multiple psychometric states. In this complementary approach, test scores X represent the probability of passing a student as follows (Dubins, Poon, & Raman-Wilms, 2016):

$$P(X = x) = \sum_{i=\max(0,x-n+k)}^{\min(k,x)} \left[\binom{k}{i} (1 - \beta)^i \beta^{k-i} \cdot \binom{n-k}{x-i} p^{x-i} (1 - p)^{n-k-x+i} \right]. \tag{2}$$

In Equation (2), a test of n MC items returns x correct responses. Of the correct responses, at most k items are due to knowledge and the remaining items are guessed with an overall probability of success p. Blunder β is the probability of an incorrect response despite knowledge. For modeling purposes, blunder is an empirical parameter without reference to the underlying reason *e.g.*, misinformation or some construct-irrelevant factor. Equation (2) takes as input an ensemble of MC items, so it is directly useful when quantification of knowledge of *whole-test* structures is desired. As is the case with phenomenological models, p and β reflect the data as observed and may be colored by partial knowledge. These limitations and strategies for mitigation are addressed in the subsequent Discussion section.

1.2. Rationale of This Study

Currently, methodological gaps exist in knowledge assessments from MC tests. IRT's treatment of test items as the unit of analysis lends itself to MC item design and, given a suitably diverse inventory of items, test optimization. Without sizable test banks, assessors in specialized coursework are not typically in possession of the large inventories needed to make significant adjustments to tests, nor may abrupt changes be desirable from a continuity perspective. In most classroom situations where test structures are at least partially constrained, one is often interested in the level of knowledge and guessing by the examinees writing the tests as constructed, rather than focusing on characteristics of

[redacted]

1.2.1. Research Questions

[redacted]

1. [redacted]
2. [redacted]
3. [redacted]

2. Methods

[illegible]

[illegible]

[illegible]

[illegible]

Supplementary Dataset

[illegible]

[illegible]

[illegible]

Table 1. MC Test set for this Study.

Test	# Items	# Students	# Analyzed	# Non-responses*
A	42	40	1,668	12 (0.7%)
B	58	40	2,302	18 (0.8%)
C	42	35	1,415	55 (3.7%)
D	40	37	1,461	19 (1.3%)
E	29	20	569	11 (1.9%)
F	26	25	641	9 (1.4%)
G	32	20	638	2 (0.3%)
H	17	25**	412	13 (3.1%)
Total	286		9,106	139 (1.5%)

Note. # analyzed = # questions × # students - # non-responses.
*Number and % of instances of MC or confidence non-response, not the number of respondents.
**These students took Test F also.

3. Results

3.1. MC Test Scores Alone are Not Robust Measures of Explicit Knowledge

To address Question 1, we begin with the simplest psychometric model of test scores. If only guessing is considered, test scores are probabilistically modeled by a binomial distribution:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3}$$

In this base model, knowledge is indirectly inferred as score distributions that are right-shifted from those expected from random selection among equiprobable choices; $p = .2$ for a five-choice MC test. Easier questions cluster toward to the left of the distribution and difficult questions to the right. We illustrate this model with Test A, which consists of $n = 40$ items written by 42 students (Table 1). Equation (3) affords a fit to the data by conventional frequentist maximum likelihood methods [Figure 1A]. To assess prior beliefs in a Bayesian realization of the model, we treated the prior expectation of p by its conjugate prior, which is a beta distribution [Appendix 2A]. This and other choices of priors made no meaningful effect on the posterior estimate for $p = .71$ [Figure 1B]. The other tests in the data set exhibited similar statistics. A simple model of MC responses as a basket of Bernoulli trials is therefore robust to prior beliefs on the test score.

To capture knowledge in MC test scores explicitly, the simplest extension of Equation (3) is to treat each question as either probabilistically or definitively selected (Dubins, Poon, & Raman-Wilms, 2016). Knowledge is directly modeled as the removal of a subset of k items, or equivalently, a fraction k/n of the whole test of n items, from probabilistic consideration:

$$P(X = x) = \binom{n-k}{x-k} p^{x-k} (1-p)^{n-x}, x \geq k \tag{4}$$

As with p, a Bayesian approach assigns a prior distribution to the knowledge parameter k, in this case (for a non-negative integer) a beta-binomial distribution [Appendix 2B]. In sharp contrast with the base model, posterior estimates of knowledge k and success rate p are strongly influenced by the choice of the prior distribution for k. Prior expectations of knowledge (k/n) at 30%, 50% and 70% result in altogether different posterior values of k. A bias toward higher levels of knowledge in the prior distribution inflates the posterior knowledge estimate while discounting the posterior estimate on the guessing efficiency [Figure 1C and Figure 1D]. Thus, Bayesian modeling of explicit knowledge based on test scores alone does not generate unbiased posterior estimates. Since Model (4) is the simplest formulation of knowledge as a probabilistic observation, this limitation would persist in more complex models that rely solely on test scores as input.

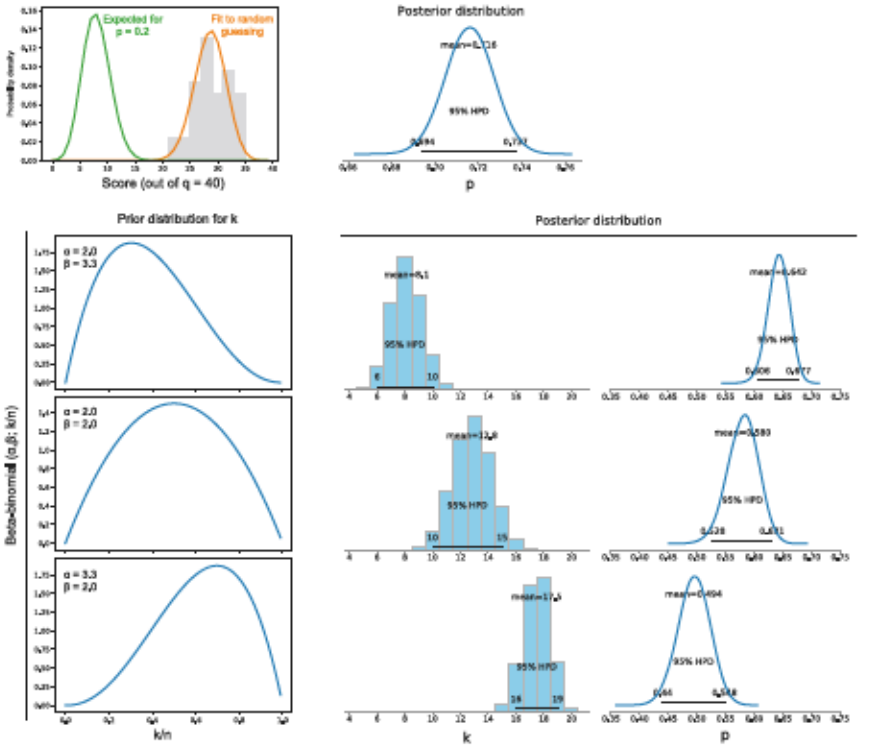


Figure 1. Limitation of Test Scores Alone in Empirical Modeling of Student Knowledge. [illegible] Histogram of scores from Test A (40 items × 42 students = 1,680 responses) are shown as an example. **A**, Binomial distributions for random selection [illegible] .2) and maximum likelihood [illegible]t to binned 5-option MC scores (orange). **B**, Bayesian [illegible]t of the basal model, Eq (3), to the data based on a prior expectation consistent with [illegible] .2. The posterior distribution for [illegible]s shown with the 95% credible interval. **C** and **D**, Bayesian inference according to Model (4), incorporating student knowledge [illegible] Shown are prior (beta-binomial) distributions for [illegible]peaking at 30%, 50%, and 70% knowledge in Panel C. Posterior distributions for [illegible]and [illegible]and their associated credible intervals are shown in Panel D.

[illegible]

Table 2. Psychometric Classification of Examinee Knowledge in MC Testing.

Confidence Level	Correct	Incorrect
1 (Confident)	Knowledge: "I was confident in my answer, and it was correct."	Blunder (misinformed or construct-irrelevant): "I was confident in my answer, and it was wrong."
2 (Partially confident)	Partial knowledge: "I was not sure of my answer, and it was correct."	Partial knowledge: "I was not sure of my answer, and it was wrong."
3 (Not confident)	Lucky guess (uninformed): "I was not confident in my answer, and it was correct."	Unlucky guess (uninformed): "I was not confident in my answer, and it was wrong."

Table 3. Summary Classification of Aggregate Test Data.

Confidence Level	Correct	Incorrect	Correct:Incorrect
1 (most)	Knowledge: 42.0% (n=3,887)	Blunder: 8.6% (n=730)	5.3
2	Partial knowledge: 20.4% (n=1,828)	Partial knowledge: 13.2% (n=1,262)	1.4
3 (least)	Lucky Guess/ Uninformed: 6.3% (n=614)	Unlucky Guess/Uninformed: 9.5% (n=787)	0.8

in improving scores in the direction of more confident responses was evident [Figure 2B]. To precisely resolve the relationships between confidence levels and knowledge, we stratified test performance by the self-reported confidence ratings. Specifically, we treated the confidence response as second random variable Q with three levels of denoted from 1 to 3. Since the three confidence levels span a complete sample space, Q is described by a trinomial distribution:

$$P(Q_1 = q_1, Q_2 = q_2) = \frac{n!}{q_1!q_2!(n - q_1 - q_2)!} p_1^{q_1} p_2^{q_2} (1 - p_1 - p_2)^{n-q_1-q_2} \quad (5)$$

For each level of confidence, test performance was modeled according to Model (3). Individual item scores and their paired confidence responses jointly entered the model. Bayesian inference yielded estimates of q_i ($i = 1$, 2, or 3) for each of the three confidence categories, and their corresponding performance p_i [Appendix 2C]. For Test A, the dispersion in the parametric estimates (95% credible interval) revealed non-overlapping confidence levels as well as ordered tiers of test performance [Figure 2C]. Specifically, items which students rated the most confident were correctly answered (p_1>80%) at a substantially higher probability than items on which students admitted to any level of uncertainty. An intermediate performance level associated with partially confident responses (p_2~60%) could be unambiguously distinguished from the least confident responses (p_3~40%). The least confident category performed with greater success than expected for random guessing for 5-option questions ($p = 20\%$). In addition, the partially confident category performed better than expected if the respondents eliminated options randomly: assuming an equal probability for students to eliminate 0 to n-2 options, $p = (0.2 + 0.25 + 0.333 + 0.5)/4 = 0.321$ for a 5-option MC item (Dubins, Poon, & Raman-Wilms, 2016). Thus, a self-assessed lack of confidence reflected a more considered approach than random guessing. We address potential reasons and mitigating solutions to these discrepancies in subsequent the Discussion section. Parenthetically, the well-resolved tiers of confidence and performance demonstrated that, at three categories, statistical precision was not at all limited by sample size.

From Table 2, blunder (which includes misinformation) may be directly taken as the rate of incorrect responses in the most confident category, *i.e.*, $\beta = 1 - p_1$. The posterior distributions give a range for β between 10% and 15% (95% HPD), which was within the range reported in the literature (Fayyaz Khan, Farooq Danish, Saeed Awan, & Anwar, 2013). In the data set, the higher success rate in educated guessing tended to offset the negative effect of blunder.

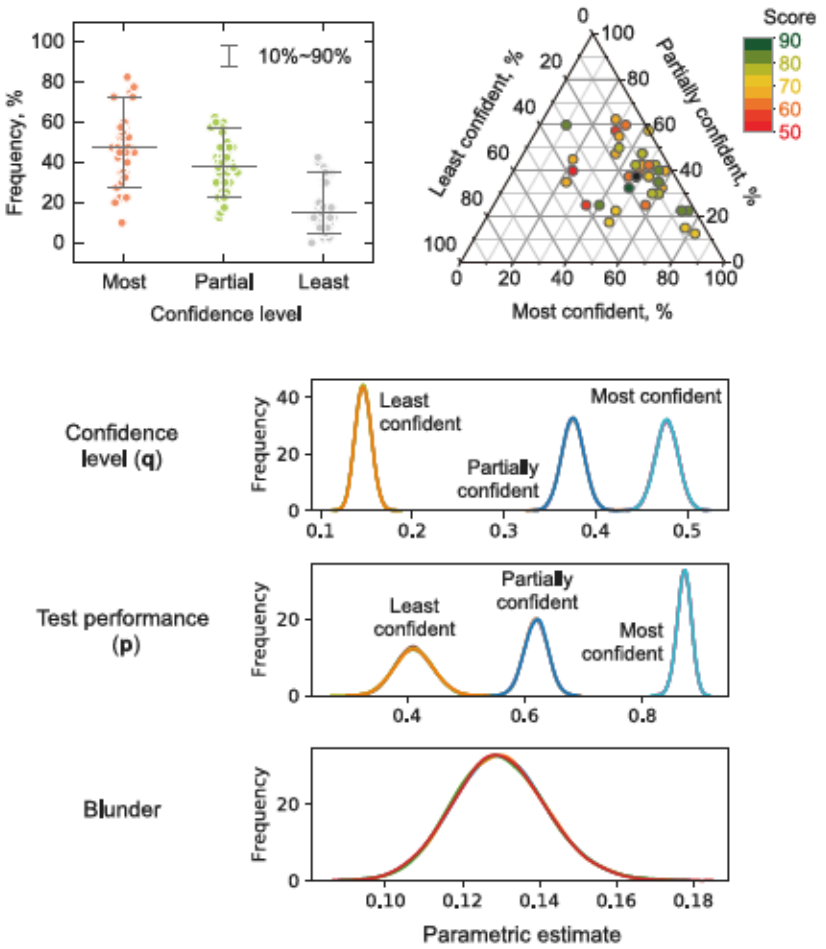


Figure 2. Self-Reported Levels of Confidence Distinguish MC Test Performance. Data from Test A is shown. **A**, Frequencies of each of the three confidence levels among the 42 MC items. Bars represent the mean $\pm 10^{th}/90^{th}$ percentile. **B**, Ternary plot showing the distribution of the three confidence levels in the context of individual respondents. To find the proportions of most, partial, and least confident responses for a respondent, track along the diagonal lines from a point back to the three respective axes. Overall % scores for the respondents are colored according to the scale as shown. **C**, Posterior plots of confidence levels and corresponding test performance following Bayesian analysis of Model (5). Posterior parametric estimates for all eight tests are given in Table 4.

Table 4. Summary of Bayesian Analysis of Test Performance According to Self-Reported Confidence Levels (%).

Test	# Q	Avg score	Most confident		Partially confident		Least confident		Blunder
			θ_1		θ_2		θ_3		$(1-\theta_1)$
A	40	71.5	47.8 (45.4, 50.1)	87.1 (84.7, 89.3)	37.6 (35.3, 39.9)	61.9 (58.1, 65.7)	14.7 (13.0, 16.4)	41.2 (35.2, 47.4)	12.9 (10.7, 15.3)
B	40	69.5	53.4 (51.3, 55.4)	83.7 (81.6, 85.7)	33.2 (31.3, 35.1)	54.7 (51.1, 58.1)	13.4 (12.0, 14.8)	37.2 (31.9, 42.6)	16.3 (14.3, 18.4)
C	35	66.2	45.1 (42.6, 47.7)	82.7 (79.8, 85.6)	33.3 (30.9, 35.7)	57.0 (52.5, 61.3)	21.6 (19.5, 23.7)	33.9 (28.6, 38.9)	17.3 (14.4, 20.2)
D	37	64.6	48.1 (45.6, 50.6)	79.2 (76.2, 82.1)	34.1 (31.7, 36.5)	55.8 (51.4, 0.6)	17.7 (15.7, 19.6)	36.8 (30.9, 42.6)	20.8 (17.9, 23.8)
E	20	74.0	46.7 (42.6, 50.7)	*86.8* (82.6, 90.6)	35.0 (31.1, 38.9)	*69.8* (63.4, 75.9)	18.4 (15.3, 21.6)	42.6 (33.4, 51.9)	13.2 (9.4, 17.4)
F	26	80.8	63.3 (59.7, 67.1)	88.4 (85.3, 91.4)	22.8 (19.6, 26.0)	52.2 (41.9, 62.2)	13.9 (11.3, 16.6)	52.2 (41.9, 62.2)	11.6 (8.6, 14.7)
G	20	72.2	48.0 (44.2, 51.9)	85.2 (81.3, 89.1)	30.2 (26.6, 33.7)	62.1 (55.3, 68.9)	21.8 (18.7, 25.0)	56.0 (47.8, 64.0)	14.8 (10.9, 18.7)
H	25	68.7	48.4 (43.7, 53.1)	82.2 (77.2, 87.5)	35.5 (31.1, 40.1)	59.5 (51.7, 67.2)	16.1 (12.6, 19.6)	37.2 (25.9, 48.1)	17.8 (12.5, 22.8)
Mean	30	70.9	50.1	*84.4*	32.7	59.1	17.2	42.1	15.6

[illegible] Test performance [illegible] was stratified by self-reported confidence data [illegible] for all eight tests according to Model (5). Posterior estimates of confidence levels [illegible] and their associated success rates [illegible] are given with 95% credible intervals (in gray).

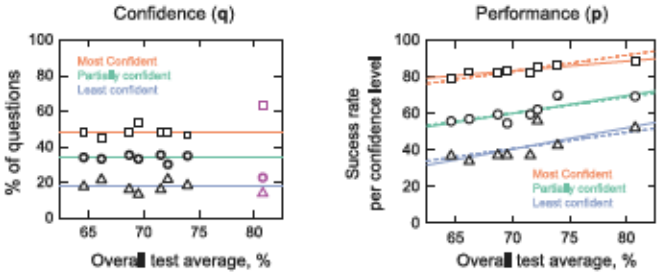


Figure 3. Correlation of Condence and Performance Levels with Overall Test Scores. [redacted] Estimates of condence and performance from Bayesian analysis of eight independent MC tests are plotted against the overall average test scores. Values are given in Table 4. Lines represent linear least-square ts to the data. **A**, Self-reported condence levels. The data in purple (for Test F) was excluded from the t. **B**, Test performance by condence level. Dashed lines represent a linear t to the three sets in which the slope is shared to guide the eye.

[redacted]

Table 5. Pass Marks (%) Required to Detect 50% and 60% Knowledge for a 5-Question MC, with ≤ 5% of a False Positive Rate.

# Items	Pass marks required for β=15.6%, probabilities and rates of guessing from the means of Table 5 (p=.5324[a])		Column C: Type II error:	Column D: Type I error:
	Column A: 50% Knowledge	Column B: 60% Knowledge	Probability that a student with 50% knowledge fails (pass mark from Column A, p=.5324)	Probability that a student with no knowledge passes (pass mark=50%, p=.421[b])
2	–	–	–	66.5%
4	–	–	–	56.1%
6	–	–	–	49.9%
8	100.0%	–	95.9%	45.5%
10	100.0%	100.0%	98.2%	42.4%
12	100.0%	100.0%	93.9%	39.2%
14	92.9%	100.0%	96.9%	36.7%
16	93.8%	93.8%	98.4%	34.6%
18	88.9%	94.4%	96.1%	32.7%
20	90.0%	90.0%	97.9%	31.0%
30	83.3%	86.7%	95.1%	24.3%
40	82.5%	85.0%	96.8%	19.7%
50	80.0%	84.0%	95.4%	16.2%
60	80.0%	81.7%	94.1%	13.4%
70	78.6%	81.4%	96.1%	11.2%
80	77.5%	81.3%	95.3%	9.4%
90	77.8%	80.0%	94.6%	8.0%
100	77.0%	80.0%	96.3%	6.8%
120	75.8%	79.2%	95.3%	4.9%
140	75.7%	78.6%	94.5%	3.6%
160	75.0%	78.1%	95.7%	2.7%
180	74.4%	77.8%	95.2%	2.0%
200	74.5%	77.5%	94.7%	1.5%
220	74.1%	76.8%	95.8%	1.1%
240	73.8%	76.7%	95.5%	0.8%
260	73.5%	76.5%	95.1%	0.6%
280	73.6%	76.4%	94.8%	0.5%

Note: An Excel template used to for model the predicted test outcomes is provided in **Supplementary Spreadsheet**. [a]Weighted average combining the partial and least confident categories; [b]Least confident category.

would pass the test (Column D). The model predicts that this hypothetical naïve student would pass more than 5% of the time on a test with fewer than 120 items, given a 50% pass mark. For an "average" test derived from our data set, consisting of 30 items, the model predicts a 24.3% chance of passing a naïve student, given a 50% pass mark. Thus, based on the average characteristics of the tests and students in the data set, a 5-option MC test would "fail to fail" a least-confident student about a quarter of the time.

4. Discussion

4.1. Implications for Knowledge Measurement

In this study, we used a series of empirical models of knowledge, guessing, and blunder to resolve their contributions in MC test scores. Compared with item response theory (IRT), which is geared at item analysis, the two approaches thus represent different perspectives and offer complementary benefits in modeling test scores. Where IRT provides detailed analysis on individual test items, our models resolve whole-test performance and enable comparison across tests. Applying these models to a set of real-world assessments, we showed that scores alone are insufficient to disentangle explicit knowledge (number of questions known) among probabilistic outcomes (Question 1). More precisely, the Bayesian analysis showed that prior belief of the knowledge level strongly pre-disposed the inferred knowledge level when only test scores were considered. This is not a trivial result. Without explicitly

sensitivity is more important in formative assessments that emphasize growth and change (Ding, Davison, & Petersen, 2005). The techniques presented here furnish the tools for guiding these decisions and provide useful information in curricular assessment.

As a practical matter, we are not advocating that confidence surveys be administered with MC test as a routine scoring aid, which would likely bias examinee response. Instead, confidence surveys are more useful in test development – such as the first few iterations of the test – to establish pass marks for future sittings of the test based on the trial results. Subsequently, surveys can be re-administered periodically (using the existing parameters for the priors) as part of existing curricular self-studies or assessing the validity and reliability of test structures or items. Informed consent by the examinees would include an explicit statement of their participation in continuous test improvement. Since the results would update future versions of the test, and not influence how the present instance would be scored, we do not expect this knowledge to significantly bias examinee behavior.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the National Science Foundation, Division of Molecular and Cellular Biosciences [2028902]; National Institutes of Health, the National Heart, Lung, and Blood Institute (NHLBI) [HL155178.]

ORCID

Gregory M.K. Poon http://orcid.org/0000-0001-5107-9458

References

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, *4*(1), 2005/06/1, 3–12. doi:10.1007/s11299-005-0001-z

Bennett, R. E., & von Davier, M. (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Springer International Publishing. https://books.google.com/books?id=ZtpCDwAAQBAJ

Borracci, R. A., & Arribalzaga, E. B. (2018, Sep-Oct). The incidence of overconfidence and underconfidence effects in medical student examinations. *Journal of Surgical Education*, *75*(5), 1223–1229. doi:10.1016/j.jsurg.2018.01.015

Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry [10.1039/C8RP00262B]. *Chemistry Education Research and Practice*, *20*(2), 369–382. doi:10.1039/c8rp00262b

Burke, N. J., Bird, J. A., Clark, M. A., Rakowski, W., Guerra, C., Barker, J. C., & Pasick, R. J. (2009, Oct). Social and cultural meanings of self-efficacy. *Health Education & Behavior*, *36*(5 Suppl), 111S–128S. doi:10.1177/1090198109338916

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 2001/01/1, 41–50. doi:10.1080/02602930020022273.

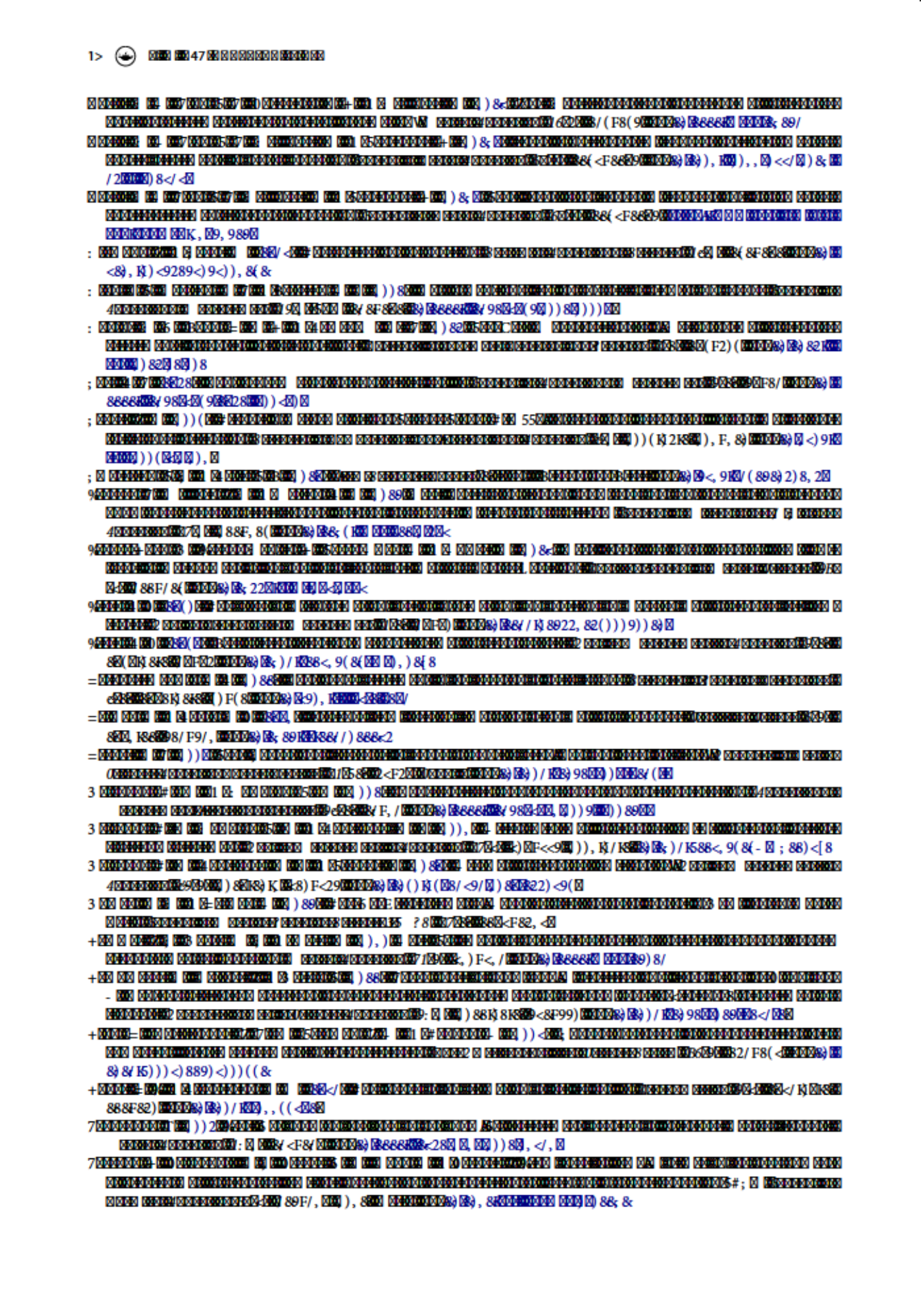
Burton, R. F., & Miller, D. J. (2006). Statistical modelling of multiple□choice and true/false tests: Ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, *24*(4), 399–411. 1999/12/1. 10.1080/0260293990240404

Bush, M. Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*. *40*(2), 218–231. 2014, Feb 17. doi: 10.1080/02602938.2014.902192

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, *25*(2), 2001/06/1, 157–163. doi:10.1080/03098770120050828.

Chaloner, K., & Duncan, G. T. (1987). Some properties of the dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics - Theory and Methods*, *16*, 511–523. doi:10.1080/03610928708829384

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 1951/09/1, 297–334. doi:10.1007/bf02310555.



APPENDICES

Appendix 1. Technical Analysis of Test Data Set

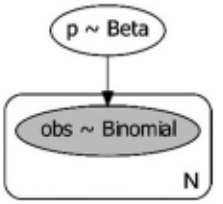
Test	A	B	C	D	E	F	G	H
Mean (SD)	71.5 (8.7)	69.5 (15.6)	66.2 (12.6)	64.6 (15.1)	74.0 (13.2)	80.8 (13.4)	72.2 (16.7)	68.7 (13.9)
Difficulty Indices*								
Discrimination Indices**								
KR20	0.433	0.818	0.671	0.798	0.551	0.720	0.751	0.591

*Distribution of the difficulty indices: < 0.25 (red); 0.25 to 0.75 (yellow); > 0.75 (green).
**Distribution of the discrimination indices: < 0.1 (red); 0.1 to 0.3 (yellow); > 0.3 (green).

Appendix 2. Bayesian Modeling of MC Scores

[redacted]

[redacted]

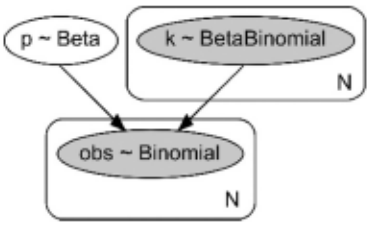


[redacted]

Appendix 2B. Explicit Modeling of Knowledge

To model knowledge directly in a probabilistic context, knowledge is treated as the functional removal of a subset of $k \leq q$ items from the test, Equation (1). For Bayesian analysis, this extended model is:

$$\begin{aligned} obs &\sim \text{Binomial}(p-k, q-k) \\ p &\sim \text{Beta}(\alpha_1, \beta_1) \\ k &\sim \text{Beta}-\text{binomial}(\alpha_2, \beta_2, q) \end{aligned} \tag{S2}$$



The new discrete parameter representing student knowledge, k, is modeled as a beta-binomial prior distribution which spans the interval [0, q]. Beta-binomial distributions are well suited for priors for k as they model the fraction of known items (k/q, out of the test of q items) as a random variable within the beta framework.

As shown in Figures 1C and D, posterior estimates of p and k are highly sensitive to the prior distribution for k. As **Figure S1** shows, the dispersion of these estimates *i.e.*, widths of the posterior distributions also depend strongly on the prior for k.

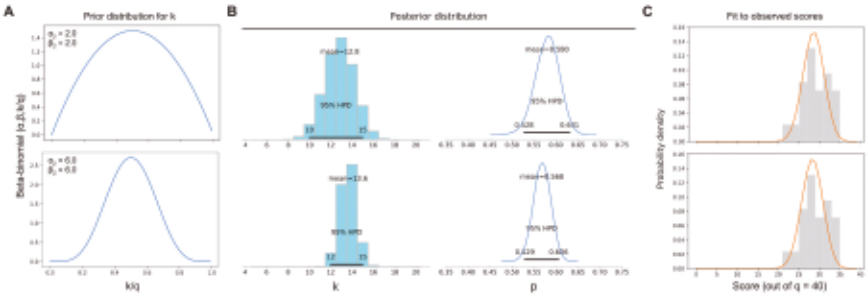
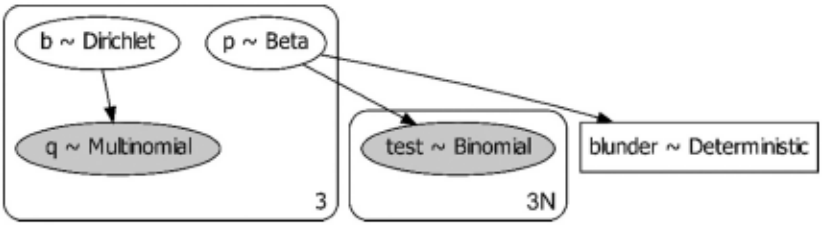


Figure S1. Influence of the Dispersion in the Prior Distributions on Posterior Estimates by a Simple Model Incorporating Student Knowledge. Test A scores (42 items) are shown as an illustrative example. **A**, Two prior distributions for k, both concentrated at k/q = 50% (q = 40) but differing in dispersion. **B**, Posterior distributions of k and p following 105 steps of MCMC simulation, discarding 104 steps of burn-in (not shown). **C**, Fits of the resultant binomial distributions to the observed scores.

Appendix 2C. Incorporation of Confidence in MC Score Analysis

To overcome the sensitivity of modeled knowledge level to its prior distributions from test scores alone, knowledge enters independently from a paired survey of confidence levels from the respondents. Since the three confidence levels span a complete sample space, they were subject to Bayesian analysis in a Dirichlet-multinomial framework (Chaloner & Duncan, 1987) in which the confidence levels were represented as a trinomial distribution with a Dirichlet prior, Equation (3). To express the differential performance among the three confidence levels, test scores corresponding to each confidence level were modeled separately as binomial distributions with beta-distributed priors. Since the number of confidence responses q from the three levels sum to the total number of questions in the text, the success rates (probability of correct answers) over the three confidence levels ($i = 1$ to 3) link the confidence responses with performance data for the N students:

$$q = \sum_{j=1}^{N} \sum_{i=1}^{3} q_{i,j}$$
$$p_i = \frac{x_i}{q_i} \tag{S3}$$



As before, the analysis was performed numerically by MCMC simulations, here initialized with uniform a Dirichlet distribution for **q** (confidence levels) and beta priors for **p** (correct answers). (Bold symbols denote vectors spanning the three confidence levels.) The simulations resolved estimates of **q** and **p** from the posterior distributions (Figure 2D). Note that the success rates p associated with each level of confidence relate to the *subset* of items on which the students offer that opinion.

As originally formulated, blunder (including misinformation) is the probability of an incorrect answer despite knowledge, whatever the proximal cause (Dubins, Poon, & Raman-Wilms, 2016). Since knowledgeable respondents are mapped as most confident, blunder was realized as:

$$\text{blunder} = 1 - p(\text{most confident}) \tag{S4}$$

Aggregate analysis of the fitted data as a function of mean test scores (Figure 3B) shows a differential contribution to test performance among the three confidence levels. Figure 3B included a confidence outlier as shown in Figure 3A. For completeness, exclusion of this out-lying test in **Figure S2** below did not alter the conclusion of the analysis.

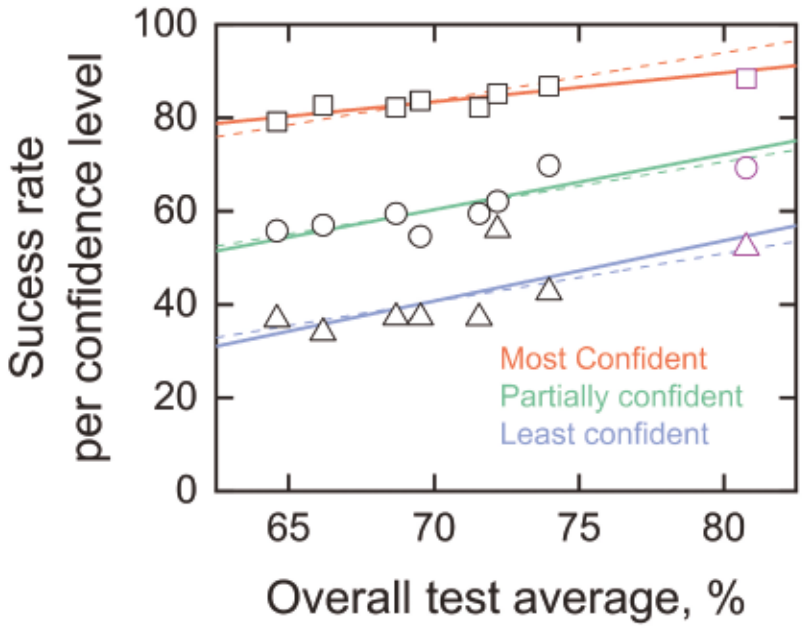


Figure S2. Correlation of Success Rates Per Confidence Level Minus Outlier.
Note. The same fits as described for Figure 3B in the main text were applied to the data set in Table 2 excluding the data for Test F (purple symbols). The least confident category was still more efficient in performance than the other two categories in tests over ~70% average.