Enriching Assessment of Physics Quantitative Literacy: IRT with Multiple-Response Items

Trevor I. Smith
Rowan University

Philip Eaton
Stockton University

Alexis Olsho
United States Air Force Academy

Suzanne White Brahmia
University of Washington

Keywords: Quantitative Literacy, Item Response Theory, Physics

Students in introductory physics courses are often asked to use familiar mathematics in novel ways. Improving students' quantitative literacy within physics contexts (i.e. physics quantitative literacy, or PQL) is a goal for many physics instructors. A challenge in assessing PQL is the blended relationship between students' general mathematical reasoning and conceptual understanding of physics. A recently developed multiple-choice instrument for measuring PQL includes several multiple-choice-multiple-response (MCMR) items for which students may select as many responses as they think are correct (White Brahmia, et al., 2021). The MCMR item format provides an opportunity to see the interaction between students' quantitative literacy and conceptual physics understanding.

Previous approaches to analyzing data from MCMR items fall into three categories: scoring response combinations as either completely correct or incorrect (Kubinger & Gottschall, 2007; Smith, et al., 2020; White Brahmia, et al., 2021), defining an *a priori* partial credit score for each response combination based on expert opinions (Wilcox & Pollock, 2014), or avoiding item scores by treating responses as categorical variables (Smith, et al., 2019; Vlach & Plašil, 2006).

In this poster we present a novel approach to defining a data-driven partial credit score for each combination of responses to MCMR items based on item response theory (IRT) analyses. We seek to address three research questions: 1) What score is assigned to each combination of responses based on IRT parameters? 2) How do correct and incorrect responses impact these scores? 3) How do scores based on IRT parameters relate to traditional dichotomous scores?

Data come from 3424 introductory physics students at a large research university. We apply IRT methods to MCMR data by treating each item as a combination of individual dichotomous response-items and grouped nominal response-items (Smith, et al., 2022). IRT parameters are used to define a score for each response combination (Smith & Bendjilali, 2022).

Our results show that the correct response combination for each item always has the highest IRT-defined score, even though the correct responses are not specified in the model. Correct individual responses have positive scores, and incorrect responses have negative scores. Total IRT-defined test scores strongly correlate with dichotomous test scores ($r = 0.8$). This supports the interpretation of IRT parameters as indicating levels of correctness and understanding.

Some analyses of grouped responses yielded counterintuitive results: we identified one item for which certain incorrect responses corresponded to positive parameters, and another item for which selecting two (out of three) correct responses was scored lower than selecting only one.

Our overall results indicate a strong potential for using IRT analyses to inform partial credit scores for MCMR items used to measure students' PQL; however, more consideration is needed to define the transformation from IRT parameters to partial credit scores.

Acknowledgments

Supported by NSF grants DUE-1832836, DUE-1832880, DUE-2214283, and DUE-2214765.

References

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item format – An experiment in fundamental research on psychological assessment. *Psychology Science*, *49*(4), 361.

Smith, T. I., White Brahmia, S., Olsho, A., & Boudreaux, A. (2019). Developing a reasoning inventory for measuring physics quantitative literacy. In A. Weinberg, D. Moore-Russo, H. Soto, & M. Wawro (Eds.), *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education.* (pp. 1181–1182).

Smith, T. I., Eaton, P., White Brahmia, S., Olsho, A., Boudreaux, A., & Zimmerman, C. (2020). Toward a valid instrument for measuring physics quantitative literacy. In S. Wolf, M. Bennett, & B. Frank (Eds.), *Physics Education Research Conference 2020* (pp. 496–502). https://doi.org/10.1119/perc.2020.pr.Smith_T

Smith, T. I., & Bendjilali, N. (2022). Motivations for using the item response theory nominal response model to rank responses to multiple-choice items. *Phys. Rev. Phys. Educ. Res.*, *18*(1), 10133. https://doi.org/10.1103/PhysRevPhysEducRes.18.010133

Smith, T. I., Eaton, P., White Brahmia, S., Olsho, A., Boudreaux, A., & Zimmerman, C. (2022). Toward a valid instrument for measuring physics quantitative literacy. In B. Frank, D. L. Jones, & Q. X. Ryan (Eds.), *Physics Education Research Conference 2022* (pp. 496–502). https://doi.org/10.1119/perc.2022.pr.Smith

Vlach, P., & Plašil, M. (2006). Analysis of Multiple-Response Data. *Applications of Mathematics and Statistics in Economics 2006, Conference Proceedings*. https://statistika.vse.cz/konference/amse/PDF/Plasil+Vlach.pdf

White Brahmia, S., Olsho, A., Smith, T. I., Boudreaux, A., Eaton, P., & Zimmerman, C. (2021). Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. *Phys. Rev. Phys. Educ. Res.*, *17*(2), 020129. https://doi.org/10.1103/PhysRevPhysEducRes.17.020129

Wilcox, B. R., & Pollock, S. J. (2014). Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Phys. Rev. ST Phys. Educ. Res.*, *10*(2), 020124. https://doi.org/10.1103/PhysRevSTPER.10.020124

[1] Rowan University

[2] Stockton University

Enriching Assessment of Physics Quantitative Literacy: IRT with Multiple-Response Items

Trevor I. Smith[1], Philip Eaton[2], Alexis Olsho[3], Suzanne White Brahmia[4]

[3] United States Air Force Academy

[4] University of Washington

DUE-1832836
DUE-1832880
DUE-2214765
DUE-2214283

Goal: Define partial credit for response combinations using both dichotomous and nominal IRT models to analyze data from MCMR items

Introduction

- Improving students' physics quantitative literacy (PQL) is a goal for many physics instructors.
- PQL is the blended relationship between students' general mathematical reasoning and conceptual understanding of physics.
- The **P**hysics **I**nventory of **Q**uantitative **L**iteracy (PIQL), includes multiple-choice-multiple-response (MCMR) items; students may select as many responses as they think are correct (White Brahmia et al., 2021).
- The MCMR items can show the interaction between students' quantitative literacy and conceptual physics understanding.
- Previous approaches to analyzing data from MCMR items fall into three categories:
  - scoring response combinations as either completely correct or incorrect (Kubinger and Gottschall, 2007; Smith et al., 2020; White Brahmia et al., 2021),
  - defining partial credit score based on expert opinions (Wilcox and Pollock, 2014),
  - treating responses as categorical variables without scores (Smith et al., 2019; Vlach and Plašil, 2006).

Methods

- PIQL response data analyzed using item response theory (IRT).
- Data come from 3424 introductory physics students.
- Each item is a combination of individual dichotomous response-items and grouped nominal response-items (Smith et al., 2022). See Table 1
- IRT parameters are used to define a score for each response combination (Smith and Bendjilali, 2022).

Item Response Theory

The probability of choosing a particular response k, given the student's overall physics quantitative literacy θ.

Dichotomous Two-parameter Logistic Model (2PL): A response option is either chosen or not chosen, e.g., response "A" to item 17.

$$P_{2pl}(k|\theta) = \frac{e^{a_k\theta + d_k}}{1 + e^{a_k\theta + d_k}}$$

Nominal Response Model (NRM): A response option is a set of related responses or combinations of responses, e.g., responses "D", "E", "both D& E", or "neither D nor E" to item 18

$$P_{nrm}(k|\theta) = \frac{e^{a_k\theta + d_k}}{\sum_i e^{a_i\theta + d_i}}$$

Joint probability of selecting a particular combination of responses is the product of individual functions.

$$P_{16}(A, C, F|\theta) = P_{2pl}(C|\theta)P_{2pl}(D'|\theta)P_{nrm}(A \cap F|\theta)P_{nrm}(B' \cap E'|\theta)$$

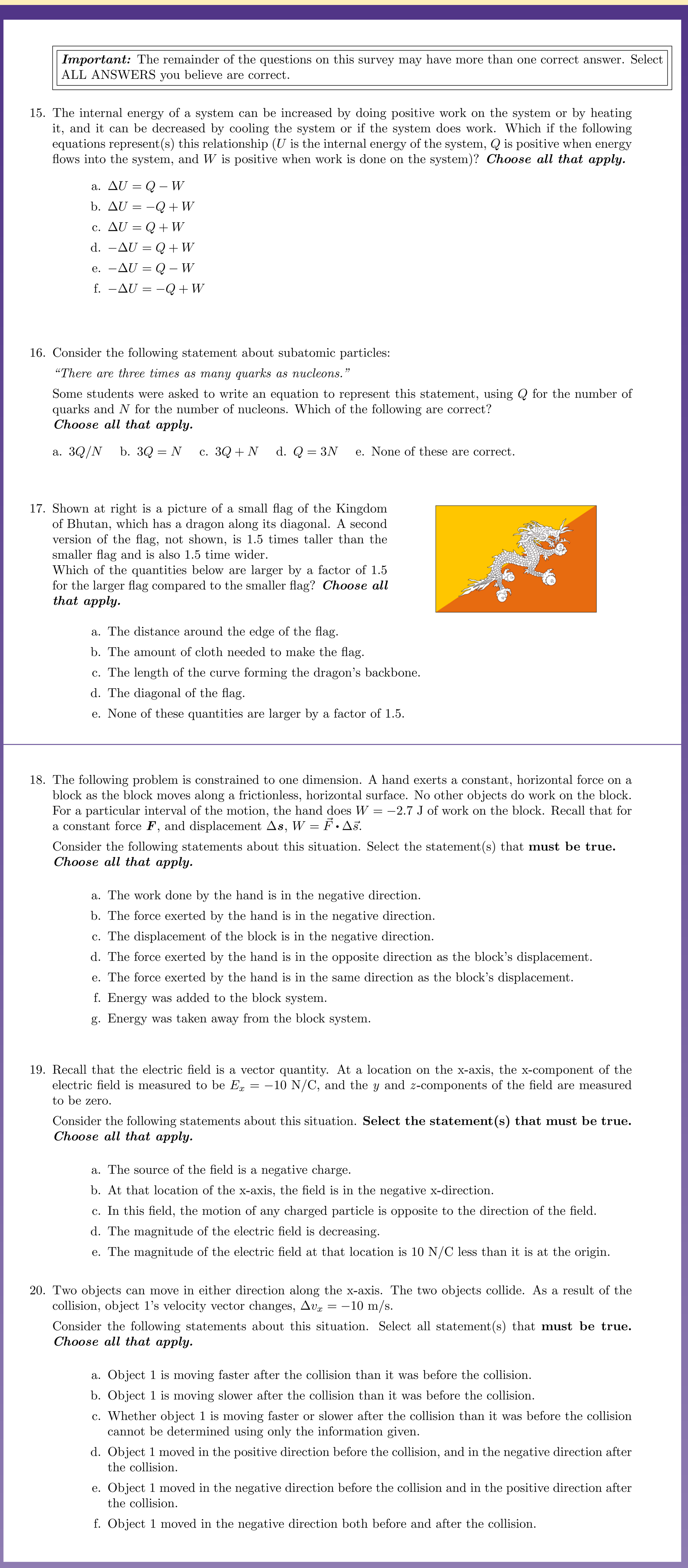
$$= \left(\frac{e^{a_C\theta + d_C}}{1 + e^{a_C\theta + d_C}}\right)\left(\frac{1}{1 + e^{a_D\theta + d_D}}\right)\left(\frac{e^{a_{AF}\theta + d_{AF}}}{\sum_i e^{a_i\theta + d_i}}\right)\left(\frac{1}{\sum_j e^{a_j\theta + d_j}}\right)$$

$$= \frac{\exp\left[(a_C + a_{AF})\,\theta + d_C + d_{AF}\right]}{Z_{16}(\theta)}$$

where $i = \{0, A, F, AF\}$ and $j = \{0, B, E, BE\}$. $Z_{16}(\theta)$ is a summation including one term for every possible combination of responses.

The "a_k" parameters indicate level of PQL (Smith and Bendjilali, 2022), and $a_0 = 0$. We use these to define partial credit scores.

For the dichotomous 2PL model, a negative a_k value indicates that the probability of selecting response k is a monotonically decreasing function of θ. But for the nominal model, positive or negative a_k values indicate higher or lower PQL than not selecting any of the responses, respectively.



Important: The remainder of the questions on this survey may have more than one correct answer. Select ALL ANSWERS you believe are correct.

15. The internal energy of a system can be increased by doing positive work on the system or by heating it, and it can be decreased by cooling the system or if the system does work. Which if the following equations represent(s) this relationship (U is the internal energy of the system, Q is positive when energy flows into the system, and W is positive when work is done on the system)? ***Choose all that apply.***
    a. $\Delta U = Q - W$
    b. $\Delta U = -Q + W$
    c. $\Delta U = Q + W$
    d. $-\Delta U = Q + W$
    e. $-\Delta U = Q - W$
    f. $-\Delta U = -Q + W$

16. Consider the following statement about subatomic particles:
    "There are three times as many quarks as nucleons."
    Some students were asked to write an equation to represent this statement, using Q for the number of quarks and N for the number of nucleons. Which of the following are correct?
    Choose all that apply.
    a. $3Q/N$ b. $3Q = N$ c. $3Q + N$ d. $Q = 3N$ e. None of these are correct.

17. Shown at right is a picture of a small flag of the Kingdom of Bhutan, which has a dragon along its diagonal. A second version of the flag, not shown, is 1.5 times taller than the smaller flag and is also 1.5 time wider.
    Which of the quantities below are larger by a factor of 1.5 for the larger flag compared to the smaller flag? ***Choose all that apply.***
    a. The distance around the edge of the flag.
    b. The amount of cloth needed to make the flag.
    c. The length of the curve forming the dragon's backbone.
    d. The diagonal of the flag.
    e. None of these quantities are larger by a factor of 1.5.

18. The following problem is constrained to one dimension. A hand exerts a constant, horizontal force on a block as the block moves along a frictionless, horizontal surface. No other objects do work on the block. For a particular interval of the motion, the hand does $W = -2.7$ J of work on the block. Recall that for a constant force $\boldsymbol{F}$, and displacement $\Delta\boldsymbol{s}$, $W = \vec{F} \cdot \Delta\vec{s}$.
    Consider the following statements about this situation. Select the statement(s) that **must be true.** ***Choose all that apply.***
    a. The work done by the hand is in the negative direction.
    b. The force exerted by the hand is in the negative direction.
    c. The displacement of the block is in the negative direction.
    d. The force exerted by the hand is in the opposite direction as the block's displacement.
    e. The force exerted by the hand is in the same direction as the block's displacement.
    f. Energy was added to the block system.
    g. Energy was taken away from the block system.

19. Recall that the electric field is a vector quantity. At a location on the x-axis, the x-component of the electric field is measured to be $E_x = -10$ N/C, and the y and z-components of the field are measured to be zero.
    Consider the following statements about this situation. **Select the statement(s) that must be true.** ***Choose all that apply.***
    a. The source of the field is a negative charge.
    b. At that location of the x-axis, the field is in the negative x-direction.
    c. In this field, the motion of any charged particle is opposite to the direction of the field.
    d. The magnitude of the electric field is decreasing.
    e. The magnitude of the electric field at that location is 10 N/C less than it is at the origin.

20. Two objects can move in either direction along the x-axis. The two objects collide. As a result of the collision, object 1's velocity vector changes, $\Delta v_x = -10$ m/s.
    Consider the following statements about this situation. Select all statement(s) that **must be true.** ***Choose all that apply.***
    a. Object 1 is moving faster after the collision than it was before the collision.
    b. Object 1 is moving slower after the collision than it was before the collision.
    c. Whether object 1 is moving faster or slower after the collision than it was before the collision cannot be determined using only the information given.
    d. Object 1 moved in the positive direction before the collision, and in the negative direction after the collision.
    e. Object 1 moved in the negative direction before the collision and in the positive direction after the collision.
    f. Object 1 moved in the negative direction both before and after the collision.

Figure 1. MCMR items on the PIQL

Dichotomous and Nominal Response-items

Item	Independent Dichotomous Responses	Grouped Nominal Responses	Item	Independent Dichotomous Responses	Grouped Nominal Responses
15	C, D	A+F, B+E	18	A, B, C	D+E, F+D
16	A, C	B+D	19	A, B, C, D, E	–
17	A, B, C, D	–	20	D, E, F	A+B+C

Table 1. Strongly correlated response options are grouped and analyzed using the nominal response model (Smith et al., 2022).

IRT Parameter Results

	Expected Ranking	IRT Score		Expected Ranking	IRT Score
Item	C*	+0.67	Item	DG*	+1.39
15	AF	-0.20	18	D†	+0.04
	BE	-0.53		G†	+1.35
	D	-0.74		EF	+0.07
	A	-1.08		E; F	-0.46; +0.53
	F	-0.50		A; B; C	-0.20; -0.36; -0.57
	AC‡	-0.40		DE‡; FG‡	-0.48; -0.20
Item	D*	+2.27	Item	B*	+0.49
16	B	+1.54	19	A	0.00
	A	-0.41		C	-0.51
	AB	+1.13		D	-0.84
	AD‡	+1.86		E	-0.38
Item	ACD*	+1.48	Item	C	+1.76
17	A†D†	+1.11	20	C†D	+0.90
	A†C†	+1.17		D	-0.86
	A†	+0.77		A	+0.23
	C†; D†	+0.37; +0.34		B	+0.99
	B	-0.89		AB‡	-0.23
				ABC‡	-0.10
				E	-1.22

Table 2. Expected rankings of responses to MCMR items (higher on top), and associated partial credit "scores" defined from IRT parameters. Asterisks (*) indicate completely correct responses. Daggers (†) indicate partially correct responses. Double daggers (‡) indicate contradictory response combinations.

Takeaways

- Choosing mathematically equivalent responses indicates higher understanding than just choosing one of a pair. Example from item 15: $a_{AF} > a_A + a_F$ and $a_{BE} > a_B + a_E$.
- Incorrect independent (dichotomous) responses always have negative parameter values. Incorrect responses that are part of a nominal group may have positive or negative IRT parameter values. This suggests that any response from some groups indicates higher physics quantitative literacy (PQL) than not choosing any from the group.
- The order of the IRT parameter "scores" matches the expected ranking for most items, but there are some discrepancies. For example: choosing the two incorrect responses "E" and "F" on item 18 has a higher score than choosing the single correct response "D". More work is needed to reconcile these results.

Research Questions and Results

RQ1: What score is assigned to each combination of responses based on IRT parameters?

- See Table 2

RQ2: How do correct and incorrect responses impact these scores?

- All correct response scores are positive.
- All incorrect dichotomous response scores are negative.
- Incorrect nominal responses may be positive or negative, but all are lower than correct.

RQ3: How do scores based on IRT parameters relate to traditional dichotomous scores?

- IRT parameters used as scores give a range of total test scores from about 52 to 82 (compared with [0,20] for dichotomous scoring)
- When partial credit scores are scaled to have a range of [0,20], most (but not all) students have higher total scores when including partial credit.

Partial Credit vs. Dichotomous Scores

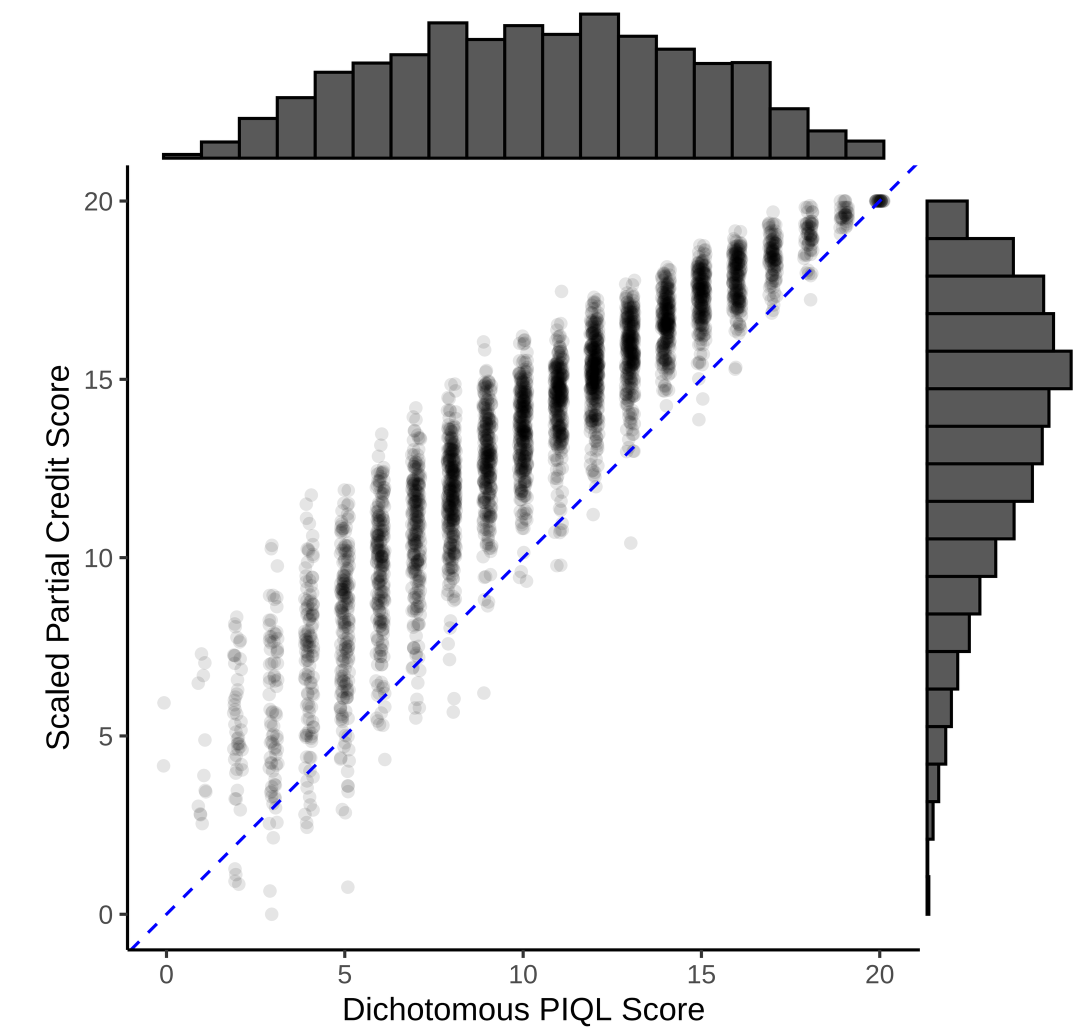


Figure 2. Partial credit scores have been scaled to have the same range as dichotomous scores: [0,20]. Points above the dashed blue line indicate higher partial credit scores, and vice versa. Histograms show the overall distribution for each score.

Conclusion

- The correct response combination for each item always has the highest IRT-defined score, even though the correct responses are not specified in the model.
- Correct individual responses have positive scores, and incorrect responses have negative scores.
- Total IRT-defined test scores strongly correlate with dichotomous test scores ($r \approx 0.8$). This supports the interpretation of IRT parameters as indicating levels of correctness and understanding.
- Some analyses of grouped responses yielded counterintuitive results: we identified one item for which certain incorrect responses corresponded to positive parameters.
- Our overall results indicate a strong potential for using IRT analyses to inform partial credit scores for MCMR items used to measure students' PQL; however, more consideration is needed to define the transformation from IRT parameters to partial credit scores.
- IRT parameters indicate how much knowledge or understanding is indicated by each response. Translating this into a partial credit score is not entirely straightforward.

References

Kubinger, K. D. and Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item format — An experiment in fundamental research on psychological assessment. *Psychology Science*, 49(4):361.

Smith, T. I. and Bendjilali, N. (2022). Motivations for using the item response theory nominal response model to rank responses to multiple-choice items. *Phys. Rev. Phys. Educ. Res.*, 18(1):10133.

Smith, T. I., Eaton, P., White Brahmia, S., Olsho, A., Boudreaux, A., and Zimmerman, C. (2020). Toward a valid instrument for measuring physics quantitative literacy. In Wolf, S., Bennett, M., and Frank, B., editors, *Physics Education Research Conference 2020*, PER Conference, pages 496–502, Virtual Conference.

Smith, T. I., Eaton, P., White Brahmia, S., Olsho, A., Zimmerman, C., and Boudreaux, A. (2022). Analyzing Multiple-Choice-Multiple-Response Items Using Item Response Theory. In *Physics Education Research Conference Proceedings*, pages 432–437. American Association of Physics Teachers.

Smith, T. I., White Brahmia, S., Olsho, A., and Boudreaux, A. (2019). Developing a reasoning inventory for measuring physics quantitative literacy. In Weinberg, A., Moore-Russo, D., Soto, H., and Wawro, M., editors, *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education.*, pages 1181–1182, Oklahoma City, OK.

Vlach, P. and Plašil, M. (2006). Analysis of Multiple-Response Data. In *Applications of Mathematics and Statistics in Economics 2006, Conference Proceedings*.

White Brahmia, S., Olsho, A., Smith, T. I., Boudreaux, A., Eaton, P., and Zimmerman, C. (2021). Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. *Phys. Rev. Phys. Educ. Res.*, 17(2):020129.

Wilcox, B. R. and Pollock, S. J. (2014). Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Phys. Rev. ST Phys. Educ. Res.*, 10(2):020124.