A Tool to Assess Fine-grained Knowledge from Correct and Incorrect Answers in Online Multiple-choice Tests: an Application to Student Modeling

Patricia Albacete

Learning Research and Development Center, University of Pittsburgh, USA

Scott Silliman

Learning Research and Development Center, University of Pittsburgh, USA

Pamela Jordan

Learning Research and Development Center, University of Pittsburgh, USA

Paper presented at the 25th World Conference on Educational Media & Technology

Washington, DC, June 2017

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150155 to the University of Pittsburgh. The opinions are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

A Tool to Assess Fine-grained Knowledge from Correct and Incorrect Answers in Online Multiple-choice Tests: an Application to Student Modeling

Patricia Albacete
Learning Research and Development Center
University of Pittsburgh
United States
palbacet@pitt.edu

Scott Silliman
Learning Research and Development Center
University of Pittsburgh
United States
ssilliman@pitt.edu

Pamela Jordan
Learning Research and Development Center
University of Pittsburgh
United States
pjordan@pitt.edu

Abstract: Intelligent tutoring systems (ITS), like human tutors, try to adapt to student's knowledge level so that the instruction is tailored to their needs. One aspect of this adaptation relies on the ability to have an understanding of the student's initial knowledge so as to build on it, avoiding teaching what the student already knows and focusing on the knowledge the student lacks or understands poorly. One way of acquiring this initial student knowledge state is by having the student take a multiple-choice test. However, the overall results commonly provided by multiple-choice tests may not be at the level of granularity needed by the ITS. This paper presents a tool that allows the extraction of fine-grained knowledge from correct and incorrect answers given in multiple-choice tests. Although the tool was developed to be used by ITSs, we argue that it could become a useful instrument for teachers in classroom evaluations.

Introduction and Motivation

We are currently working on a project to enhance Rimac, an online dialogue-based intelligent tutoring system. Rimac engages high school students in post-problem reflective dialogues that address the conceptual knowledge embedded in physics problems (Albacete, Jordan, & Katz, 2015; Jordan, Albacete, & Katz, 2016). The goal of the current project is to provide the system with the ability to adapt the support it gives to students depending on their knowledge level as they work through their homework problems. To provide this adaptation, we are augmenting Rimac with a student model. This model will provide an estimated probability of a student knowing each piece of knowledge needed to solve the conceptual questions the system asks. This information is then used to adapt the support that is provided to the student when asking questions, responding to students' answers—in particular incorrect ones—and in deciding what to discuss next. The student model uses machine learning techniques to make the predictions. One important aspect of tailoring the student model to each user is the ability to provide a fairly detailed account of the initial knowledge state of the students when they start interacting with the system. This initial "picture" of the student's knowledge will be provided by the results of the pretest that the student takes before they start interacting with Rimac. In our case, we plan to initialize every piece of knowledge that is tracked by the system with the corresponding performance level shown in the pretest. Moreover, because the initialization of the student model needs to be done immediately after the student takes the pretest, the test needs to be given online and graded automatically. This is why we would need to use an online multiple-choice test rather than a constructed-response test, which may better show students' conceptual knowledge (Dufresne, Leonard, & Gerace, 2002; Kuechler & Simkin, 2010), since the latter needs to be graded by hand.

The use of multiple-choice tests in the US is ubiquitous at all levels of education (e.g., Hamblem, 2006). They are used for everyday in-class testing as well as for extensive assessments, such as the PSSA—Pennsylvania System School Assessment—which is given to students in 3rd through 8th grade in public schools, the SAT—Scholastic Achievement Test—for students applying to college, or the TOEFL—Test of English as a Foreign Language—. Their widespread use is due to the fact that they can be graded easily and accurately, they can be machine graded, they are perceived as being more objectively graded, and they can be returned to students in a timely manner. Their widespread use indicates that they are considered a reliable way of assessing student knowledge.

Usually the resulting score of a multiple-choice test represents the percent of correct answers. In some occasions, partial credit is used to give students partial scores for incorrect choices that show some understanding, albeit incomplete or not entirely correct, of the knowledge tested by the question (Baranchik & Cherkas, 2000; Frary, 1989). In either case, the goal of the test is to provide an overall assessment of the student knowledge of the topic under study. However, when using a multiple-choice test to initialize a student model, it is desirable to be able to assess the student's competence on individual pieces of knowledge. One proposal would be to have a multiple-choice question for every knowledge component[1] (KC) that is to be tracked in the student model. In our case, there are more than one hundred KCs that are represented; hence that would mean having a pretest of over a hundred test items. Even at one minute per test item, this would require 100 minutes of class time and high school classes are typically ~40 minutes long. So to keep the test to no more than 35 items required the ability to test several KCs per test item. We investigated the systems available to implement online multiple-choice tests and found that none provided the kind of support that allows the extraction of detailed knowledge from a combination of selections—or lack of selection—to model the student level of knowledge of each KC embedded in each test item. So we developed a tool, McKnowAT, to do so.

A Closer Look at Multiple-choice Questions and What They Usually Do Not—but Could—Reveal About a Student's Knowledge State

There are two types of commonly used multiple-choice (MC) test items: those that allow students to choose only a single answer, an example of which can be seen in Figure 1, and those that require the student to select several choices for the answer to be considered completely correct. An example of this second type of MC test item can be found in Figure 2.

Which of the following is a statement of the definition of acceleration?

A. Acceleration is average velocity divided by time.
B. Acceleration is change in velocity divided by time.
C. Acceleration is displacement divided by time.
D. Acceleration is change in velocity divided by the square of the time.
E. None of the above.

Figure 1: MC question to test for the definition of acceleration (italics show correct answer)

Single selection test items, sometimes also called 'radio button items', such as the one shown in Figure 1, are usually graded as correct or incorrect—correct if the right selection was made and incorrect otherwise. For example, a student selecting choice B in Figure 1 would be considered to know the definition of acceleration and would get full credit for it. A student selecting any of the other choices would be considered to lack knowledge of the definition of acceleration and would get no credit. Occasionally, partial credit may be given to an incorrect selection if it reflects a partially correct answer.

Multiple selection test items, also called 'checkbox items', such as the one in Figure 2, can either be graded as correct/incorrect (i.e., a test item is graded as correct when all necessary choices are selected and as incorrect

[1] Knowledge components refer to the units into which knowledge can be decomposed. A knowledge component is defined as a piece of knowledge that can be learned and applied independent of other knowledge components (Hausmann & VanLehn, 2010).

otherwise) or each selection can be given partial credit so that the final grade is a sum of the correct and incorrect selections. For example, in Figure 2, one possible partial credit assignment rubric is to give each correct selection a score of 0.5 and each incorrect selection a score of – 0.5. With this partial credit assignment, if a student selects choice i alone (a correct choice), his final score is 0.5; if he selects i (a correct one) and ii (an incorrect one) his final score is 0.5 – 0.5 = 0, and likewise with all possible combinations of selections (if a final score falls below 0, it is considered 0).

Which of the following statements are true about the relationship between net force on an object and that object's acceleration?

i. *The net force and acceleration vector always point in the same direction.*
ii. If the net force is nonzero, then the acceleration may be either zero or nonzero.
iii. *If the acceleration is decreased, that means the net force has been decreased.*
iv. None of the above.

Figure 2: MC question to test knowledge of Newton's second law, Fnet=m*a (italics show correct answers)

In both types of test items presented, the score the student receives reflects a general assessment of his competence regarding the overall target knowledge tested in the item. However, it does not show the student's understanding of the fine-grained knowledge that each individual choice in the item presents. For example in the test item shown in Figure 2, the overall target knowledge being tested is Newton's Second Law (Fnet=m*a). This law shows the relationship between the applied net force on an object and the object's acceleration. Each choice in this test item reflects a correct or incorrect application of this law. But the final scoring of the test item (whether it is graded using correct/incorrect or partial credit) does not reveal which specific application of the law—and the background knowledge to support it—a student knows or lacks. It just gives information about the overall knowledge of the law. For example, the student would get a score of 0 whether he selects only choice ii (an incorrect one) or if he chooses both choice ii (an incorrect one) and choice iii (a correct one). Moreover, standard grading does not reveal anything about the implications of *not* selecting a correct or incorrect answer. For example, in the item in Figure 2, if a student *does not* select answer i it may be due to lack of understanding of the vectorial nature of the net force and acceleration. Furthermore, standard scoring does not show the possible implications of an incorrect selection. For example, in the test item of Figure 1, if a student selects C—which is the definition of velocity—this would suggest that the student not only does not know the definition of acceleration but also that he does not know the definition of velocity either. Finally, it may be the case that the selection—or non-selection—of a *group* of choices reveals a lack of understanding of a particular aspect of a concept. For example, in the test item presented in Figure 3, if the student chooses not to check choices A and B this may indicate that he knows the following piece of knowledge: "in projectile motion, there are no forces applied in the horizontal direction." Having information about these fine-grained pieces of knowledge would help inform the decisions the ITS makes with respect to what knowledge to teach and what to skip, which could enhance the overall effectiveness and acceptability of the tutor system.

A thorough review of the tools available to build and grade multiple-choice tests, such as wQuiz [Open Source (GPL) PHP based web quiz engine, http://www.penguintutor.com/wquiz.php], revealed that none provided the kind of capability that would allow extraction of this kind of information from students' selections. McKnowAT was developed for this purpose and is described in the following section.

The Tool: Multiple-choice Knowledge Assessment Tool (McKnowAT)

The basic idea of the tool is that it allows the creator of a test item to specify what the selection of a choice(s) (or lack thereof) means in terms of the knowledge that a student may have or lack. This specification can be done at any desirable level of specificity.

A language was developed that allows the author of a test item to state the KCs that are affected by a student selection (or non-selection) and to specify an estimate of the student knowledge level of each of those KCs given the selection (or lack thereof). This is done by writing rules that describe how to make the estimate assignments. For each individual choice in a test item or for a group of choices in a test item, the author specifies a rule of the form:

Operator (x,y,…) Action (KC, Value)

Where:
Operator is one of the operators described in Table 1.
x,y,.. are the options students may choose to answer the test item.
Action is one of the actions described in Table 2 which involve assigning the assessment value for a KC.
KC is the knowledge component associated with the option and on which an Action will take place.
Value is the value assigned to the KC by the Action. Table 3 describes all possible values.

Operator	Allowable question types	Meaning
if(x)	Checkbox, radio	Option x is chosen
not(x)	Checkbox, radio	Option x is not chosen
and(x,y,…)	Checkbox	All options listed are chosen
nand(x,y,…)	Checkbox	Not all of the options listed are chosen (could be none)
or(x,y,…)	checkbox, radio	One or more of the options listed are chosen
nor(x,y,…)	checkbox, radio	None of the options listed are chosen
xor(x,y,…)	checkbox	Only one of the options listed are chosen
only(x)	checkbox	The option listed is the only option chosen

Table 1. Description of Operators (Checkbox = 1 or more options can be chosen; Radio = only 1 option can be chosen)

Action	Meaning
set(KC,value)	Provide evidence (value) that the student knows the KC (the total evidence that a student knows the KC is the *average* of *all* sets collected by any other rules).
override(KC,value)	Instructs the tool to ignore any *sets* for this KC; provides a value to associate with the KC instead.

Table 2. Definition of Actions

The operator and its operands define the rule preconditions. When they are satisfied, the rule fires and its actions are performed.

Once a set of rules is created for a test item, the tool applies the rules and calculates final estimates for each KC (this is done for each possible combination of selections). The final estimate for each KC, corresponding to each possible combination of selection, for a test item is calculated in the following way. First, if no rule applies to a KC, the KC is assigned a value of nil, meaning nothing is known from the selections made by the student in this test item about how much the student knows about this KC. Second, if any rule has an Override Action on a KC, the Value assigned by that Override Action is the final one regardless of any other value that was assigned to this KC. This allows the author to decide that when a certain choice is made by the student that means that he knows this KC at a certain level regardless of anything else that the student chooses in that test item. Finally, if a KC has one or more Set actions applied to it and no Override Action, the final value is calculated as the average of those Values. If the final number is negative, it is set to 0, meaning the student does not know this KC. An output with the final estimates for each KC for each possible combination of choices is produced for the author to evaluate and make corresponding changes if necessary. An example of such an output table can be seen in Table 4. When all rules have been adjusted to the author's satisfaction, a final set of rules is submitted to be used when students take the online test.

Value	Valid Action types	Meaning
0 or 0.0	set, override	student does not know the KC
1 or 1.0	set, override	student knows the KC
0.0 > value < 1.0	set, override	student knows the KC with that probability or knows that percent of the KC
< 0.0	set	Used to provide *negative* evidence that a student knows a KC (will reduce the average of all other set values for this KC)
Nil	override	Nothing is known about the student's knowledge of this KC

Table 3. Definition of Value

Given that the test consists of many test items, there is a need to also specify how to combine the estimates assigned to KCs in individual test items when all test items are considered together. The final estimate of the knowledge that a student has of a particular KC after answering *all* the test items in a test can be calculated in different ways, depending on how a test item relates to a KC and the level of specificity of the KC. There are three basic ways in which this calculation can be made:

1. Consider that each test item, for which the KC is relevant, is a different context in which the KC can be applied. In this case the total KC score would be calculated as the sum of the KC estimates in the relevant test items divided by the total number of test items for which the KC is relevant. In the particular case when the KC score is 1 (knows) or 0 (does not know) in all relevant test items, the final KC score would represent the proportion of contexts in which the student knows how to apply the KC and would be an indication of how much the student knows of that KC. For example, if KC1 is relevant in Test items T1, T2, and T3 and has an estimate of KC1=1 in T1 and T2 and KC1=0 in T3 then the final score for KC1 = (1+1+0)/3=2/3=0.66. Hence the student knows KC1 in two thirds of the contexts or with a 67% probability. A concrete example would be: KC1=direction of acceleration, T1=test item testing "direction of acceleration" in the context of "speeding up", T2=test item testing "direction of acceleration" in the context of "slowing down", and T3=test item testing "direction of acceleration" in the context of "going in a circle at constant speed".
2. Consider that each test item is evaluating part of the KC. In this case, each test item would have a "weight" associated with it representing how large of a piece of the KC it is evaluating. For each KC, these weights would add up to 1. The total score of the KC would be calculated by adding the weight of the test item multiplied by the KC score in that item over all relevant test items. In this case, each test item is independent of the other but they test a piece of a whole. For example, if KC1 is relevant in test items T1, T2, and T3 and T1 tests 30% of KC1, T2 tests 50% of KC1, and T3 tests 20% of KC1, then if a student gets KC1=1 in T1 and T2, and KC1=0 in T3, final-KC1=.3*1+.5*1+.2*0=.8. In other words, we believe that the student knows 80% of the total of KC1.
3. Consider that the estimate of a KC obtained in a test item represents the probability of the student knowing that KC. For example, if in T1 KC1=.9, this would be interpreted as the student knowing KC1 with a 90% probability. The idea is that the test items would be independent of each other and, unlike in the previous two propositions (in the first approach each test item is independent of the others, but each is a context in a space of contexts), they would test the KC with minor variations. In this case, the final KC score could be obtained by: a) taking the largest of the estimates (i.e., take an optimistic view); b) taking the smallest of the estimates (i.e., take a pessimistic view), and c) taking the average of the KC values in all test items for which it is relevant. For example, if KC1 is relevant in Test items T1, T2, and T3 with associated values .9, .3, and .6, in the optimistic view, the probability of the student knowing KC1 would be 90%; in the pessimistic view it would be 30%; and in the average it would be 60%.

The author would specify how to combine the results of all test items for a final—all test—estimate of each KC.

Below is an example that shows the rules created by an author for a test item and how they are used by the tool to assign values to the associated KCs for that test item. The example also presents the detailed knowledge that can be extracted from the selection made by a particular student and contrasts the results with what would be obtained from a standard scoring of the same test item.

The Tool in Action: An Example

Imagine a baseball in the air after being hit hard horizontally by a baseball bat. What force or forces are acting on the baseball while it is in the air (ignore air resistance)? Check all that apply.

__A. Applied force from the bat
__B. Horizontal friction force
__C. Normal force
__D. *Weight (aka gravitational force)*

Figure 3. MC question to test knowledge about applied forces on a flying object (italics show correct answer)

Suppose an author has generated the test item in Figure 3. To reflect what she thinks the student's selection of a certain choice or choices means in terms of the knowledge the student has (or lacks), she writes the following rules:

question type: checkbox
answers: A,B,C,D
KCs: 1,2,3,4,5,6,7
1: For an object in projectile motion, the only force acting on it is the gravitational force.
2: For an object in projectile motion, there are no horizontal forces.
3: Concept of friction force
4: Concept of normal force
5: Concept of contact force
6: The force of gravity is applied on a flying object
7: Applied force stays with the moving object after contact ceases (misconception)

Rules:
Rule1 "If student chooses A, student has misconception 7, does not know KC1 no matter what else is chosen, and does not know about KC2 no matter what else is chosen. Additionally he may have an erroneous understanding of the concept of contact force (KC5)"
If(A) Set(7,1) Override (1,0) Override (2,0) Set(5,-.3)

Rule2 "If student does not choose A, student does not have misconception 7, knows something about contact forces (KC5)"
Not(A) set(7,0) set(5,.5)

Rule3 "If student chooses B, student does not know about KC1 and KC2 no matter what other answers are selected and seems to have an erroneous understanding of the concept of friction force (KC3)"
If(B) Override(1,0) Override(2,0) Set(3,-.3)

Rule4 "If student chooses C, student has an erroneous perception of normal force (KC4) and of the concept of contact force (KC5). Additionally student does not know about KC1 no matter what other answers are selected"
If(C) Set(4,-.5) Set(5,-.5) Override(1,0)

Rule5 "If C is not selected, student has some understanding of normal force (KC4) and contact force (KC5)"
Not(C) Set(4,.7) Set(5,.5)

Rule6 "If student chooses D, student knows that the force of gravity is applied on a flying object"
If(D) Set(6,1) Set(1,1)

Rule7 "If student does not choose D, student does not know the force of gravity is applied on a flying object"
Not(D) Set(6,0) Set(1,0)

Rule8 "If student chooses neither A nor B, student knows that in a projectile motion there are no horizontal forces applied"
Nor(A,B) Set(2,1)

The author then submits these rules to McknowAT. The tool applies the rules to all possible combinations of selections and returns the corresponding estimates of individual KCs, as shown in table 4. The author can then look at these results and decide if they reflect what they intended to show with regards to the student knowledge. For example, the author can look at row 8 in table 4 (choices B, C—both incorrect choices) and check the values assigned to the KCs. In this case, she will see that all KCs have a value of zero meaning that the student does not know any of the KCs and that he also does not harbor the misconception captured by KC7. These KC values agree with what the author would expect for this selection of choices. The author would proceed in the same way with each row in the table refining the rules if the KC values are not the desired ones. This would be done iteratively until she is satisfied with the final KCs' estimates for all possible combination of choices in the test item. Once the author settles on a final set of rules they are submitted to be used when students take the online test.

Choice	**KC final values for test item of Figure 3**
A	KC1=0.0, KC2=0.0, KC3=nil, KC4=0.7, KC5=0.1, KC6=0.0, KC7=1.0
B	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.7, KC5=0.5, KC6=0.0, KC7=0.0
C	KC1=0.0, KC2=1.0, KC3=nil, KC4=0.0, KC5=0.0, KC6=0.0, KC7=0.0
D	KC1=1.0, KC2=1.0, KC3=nil, KC4=0.7, KC5=0.5, KC6=1.0, KC7=0.0
A,B	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.7, KC5=0.1, KC6=0.0, KC7=1.0
A,C	KC1=0.0, KC2=0.0, KC3=nil, KC4=0.0, KC5=0.0, KC6=0.0, KC7=1.0
A,D	KC1=0.0, KC2=0.0, KC3=nil, KC4=0.7, KC5=0.1, KC6=1.0, KC7=1.0
B,C	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.0, KC5=0.0, KC6=0.0, KC7=0.0
B,D	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.4, KC5=0.5, KC6=1.0, KC7=0.0
C,D	KC1=0.0, KC2=1.0, KC3=nil, KC4=0.0, KC5=0.0, KC6=1.0, KC7=0.0
A,B,C	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.0, KC5=0.0, KC6=0.0, KC7=1.0
A,B,D	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.7, KC5=0.1, KC6=1.0, KC7=1.0
A,C,D	KC1=0.0, KC2=0.0, KC3=nil, KC4=0.0, KC5=0.0, KC6=1.0, KC7=1.0
B,C,D	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.0, KC5=0.0, KC6=1.0, KC7=0.0
A,B,C,D	KC1=0.0, KC2=0.0, KC3=0.0, KC4=0.0, KC5=0.0, KC6=1.0, KC7=1.0

Table 4. Final estimates of individual KCs for all possible combination of choices in the test item of Figure 3.

Suppose, for example, that a student taking the online test selects choice A and D for the test item shown in Figure 3. Then the following rules will fire and assign a value to the corresponding KCs:

Rule 1 KC7=1, KC1=0, KC2=0 KC5=-.3
Rule 5 KC4=.7 KC5=.5
Rule 6 KC6=1 KC1=1

Those values are then combined, depending on the operators that assigned them, to give the final KC values for this test item:

KC1=0 (Override action in rule 1 overrides the Set action in rule 6)
KC2=0 (Override action in rule 1 and no other actions elsewhere).
KC3=nil (no rule assigned it a value so nothing is known of the student knowledge about KC3)
KC4=.7 (Set action in rule5 and no other actions elsewhere)
KC5=.1 (average of set actions in rule1 and rule5: (-.3+.5)/2=.1)
KC6=1 (Set action in rule 6 and no actions elsewhere)
KC7=1 (Set action in rule 1 and no actions elsewhere)

In contrast to the fine-grained information obtained using McKnowAT, a standard scoring of the test item of Figure 3, would assign full credit if choice D is selected and no credit otherwise (there is no possible assignment of partial credit in this test item since all other selections are incorrect). This would only reveal whether the student knows that the force of gravity (choice D) is the only force applied to a flying object, but it would not be able to detect, for example, that the student has the common misconception, “a force applied on an object remains with the object after contact ceases” if choice A was selected, or that his understanding of normal force (KC4) is poor if choice C was checked, or whether he knows that in a projectile motion there are not horizontal forces applied (KC2) if choice A or B were selected. To detect each of these pieces of knowledge a dedicated test item would be necessary. In the example presented, this would mean that at least 5 test items would be necessary to measure the fine-grained knowledge assessed by McknowAT from 1 test item.

Summary and Conclusion

This paper presented, McKnowAT, a tool that allows developers of multiple-choice tests to create test items that can evaluate fine-grained knowledge using a reduced number of test questions. The task is primarily accomplished by specifying what piece(s) of knowledge a student may be assumed to know (or not know) when he makes a selection, a group of selections, or decides not to check an option. Additionally, we discussed how the use of this tool allowed us to initialize the student model which will be used to guide the level of support provided by a dialogue-based intelligent tutoring system. Moreover, we believe that this tool could be used by teachers in classrooms to obtain statistics about the specific pieces of knowledge that have been mastered by her students and which are not being grasped in the desired manner. This knowledge would allow teachers to focus their teaching efforts more effectively. Furthermore, the additional time required to create the tests may be saved in teaching that is appropriately tailored to material that students have not yet acquired and that is revealed by the test. We conducted a pilot testing of McknowAT with a high school physics teacher and with an experienced physics tutor. It took both of them half an hour to learn how to use the tool and develop a couple of test items. They both saw much promise in the use of this tool.

References

Albacete, P., Jordan, P., & Katz, S. (2015). Is a Dialogue-Based Tutoring System that Emulates Helpful Co-constructed Relations during Human Tutoring Effective? In *Proceedings of the 17th International Conference on Artificial Intelligence in Education,* pp. 3-12, Madrid, Spain. Springer International Publishing.

Baranchik A. & Cherkas, B. (2000). Correcting grade deflation caused by multiple-choice scoring. *International Journal of Mathematical Education in Science and Technology, 30*(3), 371-380.

Dufresne, R. J., Leonard, W.J., & Gerace, W.J. (2002). Making Sense of Students’ Answers to Multiple-Choice Questions*. The Physics Teacher, 40*(3), 174-180.

Frary, R. B. (1989). The “none-of-the-above” option: An empirical study. *Applied Measurements in Education*, *4*(2), 115-124.

Hamblem, M. (2006). Users struggle on road to Cisco certification. *Computerworld, 40*(28), 20.

Hausmann, G. M. & VanLehn K. (2010). The Effect of Self-Explaining on Robust Learning. *International Journal of Artificial Intelligence in Education, 10*(4), 303-332.

Jordan, P., Albacete, P., & Katz, S. (2016). Exploring Contingent Step Decomposition in a Tutorial Dialogue system. In *Proceedings of the 24th Conference on User Modeling, Adaptation and Personalization*, Halifax, Canada. Association for computing machinery (ACM).

Kuechler, W. L. & Simkin, M. G. (2010). Why is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and Empirical Test. *Decision Sciences Journal of Innovative Education, 8*(1), 55-73.

Acknowledgements

The authors would like to thank Sandra Katz, Dennis Lusetich, and Sarah Birmingham for their contributions to this work. This research was supported by the Institute of Education Sciences, U.S. Dept. of Education, through Grant R305A150155 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Dept of Education.