AAPT Studies in Pedagogy 5 (2019): From Research to Learning
DOI: 10.5840/aaptstudies2019121144

Beyond Information Recall: Sophisticated Multiple-Choice Questions in Philosophy

J. Robert Loftis
Lorain County Community College

Abstract: Multiple-choice questions have an undeserved reputation for only being able to test student recall of basic facts. In fact, well-crafted mechanically gradable questions can measure very sophisticated cognitive skills, including those engaged at the highest level of Benjamin Bloom's taxonomy of outcomes. In this article, I argue that multiple-choice questions should be a part of the diversified assessment portfolio for most philosophy courses. I present three arguments broadly related to fairness. First, multiple-choice questions allow one to consolidate subjective decision making in a way that makes it easier to manage. Second, multiple-choice questions contribute to the diversity of an evaluation portfolio by balancing out problems with writing-based assessments. Third, by increasing the diversity of evaluations, multiple-choice questions increase the inclusiveness of the course. In the course of this argument, I provide examples of multiple-choice questions that measure sophisticated learning and advice for how to write good multiple-choice questions.

Introduction

For many people, the multiple-choice question symbolizes an industrial, mechanistic approach to education that is utterly inimical to authentic human learning. In practice they think, "We may have to resort to using multiple-choice questions because of the number of students we have to teach or because of pressure from administrators to demonstrate measurable outcomes. But this is something we should only do reluctantly, as a concession to the realities of our fallen state." Kenneth Howe, for instance, writes that fixed-response tests are optional. "They are useful primarily for summative evaluation, and should be restricted to measuring *knowledge*."[1] I argue that this statement is wrong on every front. Multiple-choice tests should generally be used whenever essay evaluations are used, and multiple-choice tests are very good for formative evaluation. But the last front is the most important

ISSN 2380-4076

one. The bad image that people have of multiple-choice questions is a result of the belief that multiple-choice questions are only capable of measuring the shallowest sort of content knowledge, but this is simply not the case. Multiple-choice questions are also the main weapon in the arsenal of large-scale standardized testing, and critics of standardized testing often do not distinguish the kind of question used on a test from the overall design of a test or the agenda of people promoting that style of testing. Finally, many people are probably put off from multiple-choice questions by their first-hand experience with frustrating, badly written questions.

The purpose of this article is to challenge this view. Multiple-choice questions can do much more than simply measure shallow content knowledge and, as a result, should be part of the assessment portfolio of most philosophy courses, including upper level courses and courses aimed at majors. Multiple-choice questions should be a part of a diversified portfolio because 1) they consolidate problematic subjectivity in a way that makes it easier to manage fairly, 2) they increase the diversity of evaluation portfolios in a way that balances out the virtues and vices of writing assignments, and 3) by increasing the diversity of the evaluation portfolio they increase the inclusiveness of the course. On the way to showing that philosophers can and should often include multiple-choice questions as part of their assessment portfolios, I provide examples of multiple-choice questions that measure higher order learning and advice for how to construct good multiple-choice questions.

Multiple-Choice Questions Can Measure Sophisticated Learning

I am using the term "multiple-choice question" as a synecdoche for any kind of mechanically gradable or selected response question, including things like matching and true–false questions. The best way to describe this family of questions is as "selected response," as opposed to "constructed response," because the student is selecting from a predefined list of answers. They can also be described as "mechanically gradable," although as we shall see, I do not actually recommend using machines to do the grading. Sometimes these questions are called "objective."[2] Following Michael Scriven, I think that is a really horrible way to put it.[3] As I argue below, selected response questions do not attempt to reduce or eliminate subjectivity. They merely move it to the design phase of the evaluation, where it is easier to manage.

The other term that needs to be defined is "sophisticated learning." I use Benjamin Bloom's taxonomy as a framework for measuring sophistication. In 1956 a committee of teachers led by Bloom set out to improve communication among people working in education.[4] Their lasting achievement was to establish a six-level hierarchy of cognitive skills. A revised version of the taxonomy was released in 2001, which is what I use here.[5] Bloom's taxonomy is far from perfect, but it is very

Table 1: Revised Bloom's Taxonomy.

Outcome		Multiple Choice Evaluation?
Remember	1.1 Recognize	Yes
	1.2 Recall	
Understand	2.1 Interpret	Yes
	2.2 Exemplify	Yes
	2.3 Classify	Yes
	2.4 Summarize	Yes
	2.5 Infer [extrapolate a pattern]	
	2.6 Compare	Yes
	2.7 Explain	?
Apply	3.1 Execute	Yes
	3.2 Implement	?
Analyze	4.1 Differentiate	Yes
	4.2 Organize	Yes
	4.3 Attribute	Yes
Evaluate	5.1 Check	?
	5.2 Critique	
Create	6.1 Generate	
	6.2 Plan	
	6.3 Produce	

widely used and thus accomplishes the primary goal with which the committee was tasked. Also, the developers of the taxonomy are explicit about the role multiple-choice questions can play at many of their levels of outcomes.[6]

The critics of multiple-choice questions assume that multiple-choice questions can only be used to evaluate outcomes at the bottom of Bloom's six-level hierarchy, which the revised version of the taxonomy labels "remembering": "When the objective of instruction is to promote retention of the presented material in much the same form it was taught, the relevant category is *Remember*."[7] In the updated version, "remember" is divided into level 1.1, recognizing (one knows it when one sees it), and level 1.2, "recall" (one can reconstruct the information from traces in one's long-term memory). Selected response questions, it seems, can only function on level 1.1.

This appearance, however, is merely an illusion. It is directly contradicted by the authors of the revised taxonomy. Table 1 shows the levels of the revised taxonomy and whether the authors assert that selected response questions are possible at that level. Of the nineteen subcategories, ten have evaluation methods that are explicitly labeled selected response, and another three subcategories have

evaluations that seem like they could be implemented as selected response questions but are not explicitly labeled that way. The appendix to this article shows how these can be implemented in philosophy. It contains twenty-three questions from my personal question bank, organized by their level in Bloom's taxonomy to serve as exemplars for generating further questions. For example, at the lower levels, there are five different kinds of questions about the concept of moral luck. Ideally, a student who has grasped the concept of moral luck should be able to answer any of these, no matter how they approach the concept. In fact, if one is unfamiliar with moral luck, one will probably be able to pick up a fair amount about it just from these questions.

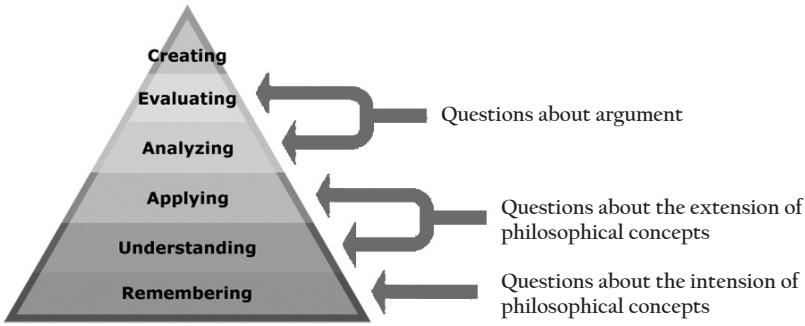

Figure 1: Mapping Kinds of Questions onto Bloom's Taxonomy[8]

Figure 1 summarizes how different kinds of questions in philosophy courses map onto different levels of Bloom's Taxonomy. At the lower levels of Bloom's taxonomy, one is simply testing students' grasp of philosophical concepts. As it turns out, there is a difference here depending on whether one is investigating the concept intensionally, through its lexical definition, or extensionally, through examples that fall under the concept. Questions about the intension of philosophical concepts evaluate at the lowest level of Bloom's taxonomy, remembering. Questions about extension introduce an element of novelty such that levels 2 and 3 of Bloom's taxonomy, understanding and applying, are being evaluated. To move to the fourth and fifth levels of Bloom's taxonomy, one needs to move from asking about individual concepts to asking about arguments. Let us look at these levels in more detail.

Levels 4 and 5 of the Bloom hierarchy are analyzing and evaluating. In philosophy this typically means analyzing the structure of arguments and evaluating their strength, validity, and soundness. Consider the following questions, which can be modified to ask either about simplified passages or passages in primary sources. Correct answers are in italics. For questions with a single correct answer, the penalty for incorrect answers is given in parentheses on a ten point scale.

For questions one and two, consider the following passage:

A longtime member of the ACLU is talking local politics with her friends. She argues that automated surveillance cameras should not be used to catch people speeding or running red lights. Once we allow the government to begin to use cameras like that, they will be used for other purposes. For instance, if someone is caught robbing someone on a camera that was supposed to just be used for traffic violations, surely people will want that evidence used in court. Pretty soon, there will be surveillance cameras everywhere, and it will be just like living in a police state.

1) Which of the following is the best statement of the conclusion of the argument given by the ACLU member above? (Select one)
   a. Automated surveillance cameras should not be used to catch people speeding or running red lights.
   b. Living in a police state. (-*10*)
   c. Cameras should not be used to identify robbers. (-5)
   d. If we use cameras to catch people running red lights, we will soon use them to catch robbers. (-5)

   Explanation (optional):

2) Which of the following are premises in the argument given by the ACLU member? Focus on the meaning of the premises, not how closely the wording in the answers matches the wording in the passage. (Select all that apply.)
   a. Automated surveillance cameras should not be used to catch people speeding or running red lights.
   b. If we use cameras to catch people running red lights, we will soon use them to catch robbers.
   c. Accepted uses for automated cameras will soon be everywhere, and we wouldn't want to live like that.
   d. Automated surveillance cameras are creepy.

   Explanation (optional):

I believe these questions require students to differentiate, which is Bloom's level 4.1 skill. Arguments could be made to place it at a different sublevel, however as long as the question is filed somewhere on level 4, it illustrates my point. Questions fifteen through eighteen in the appendix ask about the premises or conclusions of an author the student has been reading. These are not simply recollection problems, because they require students to distinguish what an author said from what is actually a premise in the relevant argument. In general, the use of true statements as distractors marks out more sophisticated questions.

In a very important way, questions that require students to perform higher level thinking go against the training most students have had with multiple-choice

questions in secondary school. Many students come to college having only seen multiple-choice testing used for factual recall. They may be thrown off by options that are true statements but are not correct answers. One might think the confusion students will face here is a reason not to use true statements as distractors, but I do not think we should give up so easily. We need to educate students about the sorts of things they will be tested on, not only because it helps prepare them for our tests, but also because it demonstrates what our priorities in education are. We are teaching them to reason, not just recognize true statements. Students should know this and know that they will be tested on it.

Why Philosophy Teachers Should Use Multiple-Choice Questions

The examples above and in the appendix show that multiple-choice questions can evaluate student performance at more sophisticated levels of thinking than many people think possible. In this section, I argue that we furthermore should use multiple-choice questions. Considerations of fairness and the imperfect nature of any one form of evaluation considered on its own imply that the portfolio of evaluations used in a course should be diversified. Further, multiple-choice questions do a very good job of balancing out the flaws of other modes of evaluation, particularly essay writing.

Let me be very clear about what I mean here. I am not saying that multiple-choice questions are more fair than other kinds of questions. Multiple-choice questions, just like any form of evaluation, have biases. My claim is that an assessment repertoire that includes multiple-choice questions is more likely to be fair than one that does not. The specific portfolio of evaluations one should use obviously depends on the learning objectives for the course. However, one should not just assume that multiple-choice questions should only be used in lower-level courses or courses aimed at non-majors. In most courses, if one uses essay-based evaluations, one should also use some multiple-choice evaluations. Consider three arguments in defense of this claim.

First, multiple-choice questions allow one to consolidate subjectivity in a way that makes it more manageable. They do not eliminate subjectivity or impose a false objectivity on evaluations that fundamentally require teacher judgment, but they do allow evaluators to move subjectivity around in a way that helps them manage it more fairly.

Since "subjective" and "objective" are such laden philosophical terms, I should be very clear about what I mean here. I call a method of evaluation "subjective" if it 1) involves a significant amount of disagreement between the judgments of different evaluators or the same evaluator at different times and 2) this disagreement can be characterized as a reasonable difference of opinion. This is meant to be a commonplace definition that sidesteps any number of philosophical quagmires

regarding objectivity. For instance, I'm not going to worry about the relative importance of knowing subject and known object in the production of knowledge. This definition is phenomenological: it is just about the presence of disagreement and how it appears to us.

I follow Scriven in saying that machine-gradable tests do not remove all subjectivity from grading. In the end, with any evaluation, one has to make a judgment call. When grading papers, one makes a fresh judgment with each paper. With mechanically gradable questions, one makes a single judgment when one sets up the question. Subjectivity is consolidated in this very straightforward sense. This consolidation has two benefits. First, it allows one to apply the same judgment consistently and to improve that judgment over multiple iterations of the test. Second, it gives one more flexibility in how to manage grading time, so that one is likely to be using one's best judgment. These are two separate factors, and I want to emphasize each of them.

To see the advantage of applying the same judgment consistently, consider question six from the appendix:

> Which of the following are examples of people who have internalized moral norms? (Select all that apply)
>
> a. Steve grows up in a drug- and crime-ridden community. He decides not to join a gang for the sole reason that he fears going to jail or being killed by a member of another gang.
>
> *b. Barbara's employer trusts her to deposit the cash that came into the store in the bank at the end of each day. Barbara could easily pocket thousands of dollars and then alter the receipts so that no one would ever know the difference. She doesn't do this, though, because she believes that stealing is wrong and doesn't want to violate the trust of her employer.*
>
> *c. Bill has problems with explosive anger. He is aware of this and has taken anger management classes. After a particularly hard day at work, he gets into an argument with his wife Jenny and punches her in the face. He instantly feels ashamed of what he did. Unable to look his wife in the eye, he locks himself in the bedroom to sulk. As his shame mixes with guilt, he realizes that he must make amends. Summoning his courage, he goes downstairs, looks Jenny in her now blackened eye, and begs forgiveness. They agree to go into couples' therapy and that he should stay with his brother for a while.*
>
> d. Sid is a stone-cold killer. He has to be if he wants to stay on top of his drug ring. His enemies need to know he is willing to have them killed. His subordinates need to know he is willing to personally kill them. Preachers and policemen have told him many times that gang life is immoral, but their lectures are all just words to him.
>
> Explain your answer (optional):

In my courses, I define the internalization of moral norms as the ability to self-regulate according to some set of moral norms without the influence of outside rewards and punishments and to identify and correct one's moral mistakes. This definition is supposed to be neutral about what moral rules get internalized. As a result, some students have trouble with item (d), imagining that Sid has internalized some kind of gang-loyalty ethic. I have done my best to word this question to minimize the impression that there is honor among thieves in this case. Still, someone might reasonably disagree. I cannot rule out the possibility of this disagreement. But by applying the same grading system, I am applying my own standard consistently.

Another important fact about this question is that its wording has evolved each time I have used it in a test. Writing good mechanically gradable questions is hard. Howe calls it "sheer drudgery."[9] The hardest part of writing a multiple-choice question is being sure it has a clear correct answer. Even long-established, widely used standardized critical thinking tests—for example, the California Critical Thinking Skills Test or the Ennis-Weir—have problems with ambiguous questions. Actually, for any widely used test, there will be a literature challenging the questions. Sometimes the challenges are cavils, but often real ambiguities are identified.[10] Ambiguity comes up because philosophical ideas are ill defined or contested, because there can be reasonable disagreement about the strength of inductive arguments, and because natural language in general is shot through with ambiguity and vagueness. At rock bottom, though, the main reason ambiguous questions crop up is instructor error. No matter how hard one tries, ambiguity will remain in sophisticated multiple-choice questions, especially the first time they are used. It is important for teachers to see how students cope with the question for the same reason researchers need an outsider to copyedit their writing. One simply will not see many of the ambiguities in one's writing because one already knows what one means.

Because multiple-choice questions are so hard to write, it is important to rework the same question many times. In all my multiple-choice questions, I explicitly give the students the option to explain their answers. This is a technique that Bob Ennis has been recommending since the early 1990s,[11] but which I first learned about in his talk at the biennial workshop-conference of the American Association of Philosophy Teachers (AAPT).[12] This option is a small thing, but it is very useful. As Ennis notes, "[i]t is fairly quick, can be comprehensive, provides forgiveness for unrefined multiple-choice questions, and allows for differences in student backgrounds and interpretation of items."[13] In my experience, adding the prompt for explanations does not significantly increase grading time over hand-graded multiple-choice tests without the prompt, because most students do not use it, and those who do generally just restate the prompt or the answer. Including this option, however, does mean that one cannot actually use a machine

to grade every part of these "machine-gradable" questions. One must look at each test directly and check for explanations of answers. But I am not advocating for multiple-choice questions on the basis of convenience. I am advocating for them on the basis of fairness, and the occasional feedback one gets that does identify real ambiguities in a question is vital for fairness. In upper level courses or courses aimed at majors, one can expect a lot more of this sort of feedback.

Just as in psychological surveys or political polling, small changes to the way a prompt is presented can dramatically influence the quantity and quality of the feedback. In my experience, providing a single space at the end of the test for all comments about the questions dramatically reduces the quantity and quality of the feedback. Not providing adequate space beneath the prompt can also influence feedback. I generally explain the use of the feedback space explicitly to students during a practice test. Providing examples of good and bad feedback improves its quality.

So, multiple-choice questions allow one to make the same judgment call consistently and improve that judgment over multiple iterations of the course. They also make it more likely that graders will be using their best judgment in the first place. Most teachers are all too familiar with the way fatigue can impact judgment. The farther graders get into a stack of papers, the more tired they can feel and the more the responses all seem to blur together. But people often do not take this decision fatigue seriously enough to structure their work to take it into account. When grading a large number of constructed response evaluations, decision fatigue is inevitable. Yet, prompt and constructive feedback is crucial for student success. So, after the test is given, there is only a short amount of time to return graded tests to students with detailed feedback, and each constructed response presents a new challenge. One might promise to only grade when rested and well-fed, but that is not realistic. Making the difficult judgments when one is initially writing the question reduces the likelihood of judging similar work differently. The questions could be written even before the course starts, and one can reuse or repurpose questions from other courses.

The case for taking decision fatigue seriously can be strengthened by looking at the psychological literature on the subject, but it is important to be careful here, because the consensus among psychologists on this matter is in flux. For a long time, the dominant model of decision making under fatigue was an idea called "ego depletion" developed by Roy Baumeister and Dianne Tice.[14] The ego-depletion literature features a number of eye-catching experiments, including a famous study of Israeli probation judges. A team of psychologists from Columbia and Ben Gurion universities tracked the decisions of eight probation judges and discovered that the biggest factor in whether they granted a favorable decision to the prisoner was how recently the judge had eaten.[15] Their favorable rulings start at around 65 percent, go down to near zero as the judges make more and more decisions, and jump up

to 65 percent after food. Importantly, the significant factor here is the number of decisions made, not the time passed. This study has become the emblematic for the ego-depletion model of decision fatigue, because it highlights several key things. First, it shows that decision making itself takes effort. Second, it illustrates how later decisions in a sequence tend to be more instinctive and cautious than earlier decisions. Third, it indicates that decision making draws on some kind of finite but replenishable resource linked to eating, specifically blood glucose. All of these points have been backed by a large body of research.[16]

If the ego-depletion model fits the way decision making under fatigue works, it amplifies the common-sense case for using multiple-choice questions to take fatigue into account when structuring your grading workflow. First, ego depletion goes some distance to quantifying the effect of fatigue and showing that the size of the effect has a significant impact on fairness. If one is up for probation, one really wants see the judges right after they have eaten their lunch. Second, since it is the act of decision making that causes the fatigue, one cannot eliminate the grading fatigue by fussing with external factors, such as grading in a calm environment. Third, the model says that decision making under fatigue will be more cautious than decision making in other circumstances. In higher education, this means decision fatigue is linked to grade inflation. In the study of the probation judges, fatigue led the judges to deny prisoner requests. Generally, these were requests for parole, but sometimes they were for things like transfer to a different prison. The authors of the study note that the safe thing to do in all of these cases is to deny the request.[17] This preserves the status quo and avoids the risk that a paroled convict will commit another crime. When the judges were tired—when their egos were depleted—they took the safe route and denied prisoner requests. For grading, the equivalent effect is going to be grade inflation. A low grade is something that a teacher might be called on to justify. The decision fatigue model bolsters the argument here because it says that decision fatigue is quantifiable, caused by the act of making a decision itself, and linked to grade inflation.

Unfortunately, the ego-depletion model of decision making under fatigue has come under a lot of criticism as a result of the replication crisis in social psychology. The criticism has been focused on part of the model that I am not directly drawing on, that decision making draws on a finite but replenishable resource linked to blood sugar, but it has implications for the whole model. Much of the research testing this model was done using the *sequential task paradigm*.[18] Subjects are asked to perform one task that requires will power, and then their ability to perform a second, distinct task requiring will power is measured. Their performance can then be compared with a variety of other experimental groups, such as a control group who did not receive the first task, or another, luckier experimental group that received a sugary treat like a cookie. A great deal of research was published using

this paradigm, until the arrival of the "replication crisis" in social psychology led to a widespread reevaluation of established experimental techniques.[19]

In response to this, an institutional framework was developed to encourage large-scale replications that register their objectives in advance to guard against data fishing.[20] Subsequently, a large-scale, multi-lab study using this framework failed to replicate the results of the sequential task paradigm.[21] The challenge affects the claim that diverse sorts of decisions all draw on the same pool of willpower, which is what experiments in this paradigm purported to show. However, the replication failure could be seen as undermining the basic idea that decision making is a measurable cause of fatigue. If that is the case, then my argument here may have to fall back on the common sense premise that one ought to get some rest before making decisions that affect student grades.

My second argument concerns the value of diversity in methods of evaluation in general and the usefulness of multiple-choice questions in particular for balancing out essay-based assignments. The desired outcomes of philosophy classes generally require one to use a very diverse portfolio of methods of evaluation. This already sets up a *prima facie* case for including any method of evaluation that increases diversity in one's assessment portfolio. Additionally, multiple-choice questions have a set of virtues and vices that is specifically useful in balancing out the virtues and vices of essay-based evaluations.

Philosophy education outcomes are generally hard to define and measured by proxy indicators. Blood pressure can be measured with a blood pressure cuff. However, the judgment that someone is likely to develop heart disease is arrived at by evaluating multiple indicators such as blood pressure and cholesterol levels. Philosophy outcomes such as "think critically" and "live philosophically" are much more like predicting heart disease than measuring blood pressure or cholesterol levels.

Because philosophy employs proxy indicators, it is best if the total set of indicators are balanced and diverse. If we only use one thing as a proxy for our larger outcome, then we may come to think that our indicator is the outcome itself. There's an old maxim in social sciences: When an indicator becomes a target, it ceases to be a good indicator. For instance, if the government decides to use single indicator as a measure of economic health then it ceases to be a good indicator, because speculators can use it to predict government action.[22] This effect has been specifically observed in education when the government sets targets for certain education outcomes, and people learn to game the system to get those outcomes.[23] Just as avoiding a heart attack is the desired outcome, and lower bad cholesterol numbers are an indicator of that outcome, so too is improved critical thinking the desired outcome, and earning an "A" in a critical thinking class an indicator of that outcome.

The need for balance and diversity is something that applies in any area with large amounts of uncertainty. When people talk about a healthy diet or a well-managed stock portfolio, "balanced" and "diversified" are words that often come

up. The same applies to a repertoire of evaluations. The fact that we have to rely on proxy indicators in a sea of uncertainty means that anything that can reasonably be used as an indicator has a *prima facie* case for inclusion in an evaluation repertoire.

In addition to having balance and diversity, it is also important that evaluations be frequent and low stakes, and this further supports the use of multiple-choice questions. A significant problem with standardized testing in primary and secondary education in the United States right now is that the fate of both teachers and students rests on a single data point: performance on a yearly standardized test. In higher education, teachers are usually fortunate to have control over how we evaluate students, so we can avoid this mistake. If teachers are going to use frequent, low-stakes tests, however, they need to use a testing format that can be scaled up. Once instructors have established a bank of well-designed multiple-choice questions, it is fairly simple to write large numbers of unique tests, making frequent, low-stakes testing possible.

It may seem odd to advocate for multiple-choice questions in frequent, low-stakes evaluations, because most people associate multiple-choice questions with the one-time, high-stakes standardized tests used in secondary education. But the use of multiple-choice questions must be separated from the myriad problems with standardized testing. Multiple-choice questions are the go-to resource for large-scale standardized testing because, once the tests are designed, they can be cheaply implemented across an entire state, province, or nation. But this same power can just as easily be harnessed for frequent, low-stakes testing. The net effect will actually be the reverse of what it is for standardized tests: the amount of stress for the student will go down, and the amount of information used in deciding overall evaluations will go up, because one is sampling information about performance at multiple times and in multiple circumstances.

Multiple-choice questions are especially useful for balancing out the drawbacks of essay-based assignments. As Peter Collins points out, central among the drawbacks is that grading essays is incredibly subjective and scattershot, and using essays as a primary form of student evaluation favors a certain sort of student.[24] Essay writing is the most cognitively sophisticated student work that we typically evaluate in philosophy. In the revised Bloom terminology, it asks students to *create*. It is also one of the places where we ask students to actually *do* philosophy, and we evaluate them on how well they do it. But evaluating essays is also an incredibly subjective act. There is a large range of reasonable disagreement in how essays should be graded, and the assessment itself measures many different things at the same time without sufficiently distinguishing them. In any essay, one measures the clarity and persuasiveness of the writing, grasp of philosophical concepts, quality of argumentation, originality, and an incredible host of other factors. This ambiguity can be reduced through the use of a rubric that differentiates the criteria used to evaluate the essay. But on a deeper level, they are all intertwined, because

the student has to be working on all of them at once. Difficulty in writing clearly can interfere with the ability to formulate a good argument. Time spent trying to get the concepts down is time lost on coming up with something original. In the end, the fact is that essay evaluation is holistic and subjective. Student performance on essays can provide evidence of learning regarding things that more structured evaluations cannot, but the cost is precision. Essay assignments need to be balanced by modes of evaluation that have the reverse virtues and vices. Multiple-choice questions do not evaluate creativity. Nor are they especially good at measuring how the student deals with the interaction of multiple factors. But they do allow us to measure individual skills more precisely and, as I have just argued, to isolate subjectivity in a way that makes it more manageable.

A final way in which essay writing and multiple-choice questions are complementary is in terms of coverage of the material and outcomes of the course. If teachers put too much emphasis on, for instance, a large essay project, students will work on the material relevant to their essay topic and ignore everything else. Because multiple-choice questions can target specific topics or outcomes, teachers can carefully deploy them to direct student attention to many or all aspects of the course. For instance, if there is a specific fine distinction that one wants to emphasize in a course, then using a variety of well-designed multiple-choice questions is a good way to emphasize it. This is a point made by Howe, but it actually contradicts his claim that multiple-choice questions should only be used in summative evaluations.[25] Multiple-choice questions can be used to give a student feedback on their understanding of a topic before they decide to write a paper on it. If a student wants to write about moral luck, for instance, it might help if earlier on they practiced identifying examples of moral luck in extension-based multiple-choice questions, such as question five in the appendix.

My first argument claimed that multiple choice questions allow one to move the subjectivity of grading to a place in one's workflow where it is more manageable. The second argument discussed several ways multiple choice questions balance out the problems with essay-based evaluations. A third argument in favor of an assessment portfolio that contains multiple-choice questions is that by increasing the diversity of the evaluations, one increases the inclusiveness of the class. Classes that rely exclusively on one form of evaluation wind up rewarding one particular kind of student, often a student whose education has groomed them to succeed in that sort of task. A diversity of evaluations means a diversity of students have the opportunity to display a diversity of talents while at the same time working on the areas where they are weaker.

Using only writing-based evaluations winds up favoring a certain sort of student: the student who writes well. This is again furthering a point made by Collins.[26] A student who has mastered the art of college essay writing is going to come off better in these sorts of evaluations than a similar student who is equally

philosophically engaged in the course but has not learned how the essay game is played. Sometimes, genuinely superficial signs of good writing, like a broad vocabulary, can influence essay grading. Little failures in understanding writing mechanics—spelling mistakes, lack of familiarity with word processors—can have an outsized influence. Of course, philosophy courses typically have improved writing as a desired outcome, which means that teaching aspects of professional writing, both major and trivial, are a part of our job. But writing outcomes are not the only outcomes philosophy professors have, and often they are not the philosophically distinctive ones. If we use evaluations that are general writing ability loaded, our evaluations are distorted.

Favoring students who are good at a particular form of evaluation is not, in itself, a way of favoring students from already dominant or overrepresented groups, but it can indirectly have that effect. The college writing game is something that students from elite schools are trained to win at. Of course, students can also be groomed to answer multiple-choice questions. A course that only uses multiple-choice evaluations would also be a mistake. One can also write multiple-choice questions in a way that avoids some of the cruder tricks that have historically been used to game standardized tests. In general, however, a diversity of evaluations means that evaluations will actually track the outcomes one is trying to produce but which can only be measured using proxies and rough indicators. As a result, teachers will not favor privileged students who have been trained to perform superficial tricks to game the system.

Multiple-choice questions can further increase inclusiveness because students perceive them as being fairer than other types of assessments.[27] The questions have right and wrong answers that the students have either given or not given, so they feel that they are not being subjected to the prejudice or whims of the grader. This also has a positive effect on students' perception of philosophy. Wendy Turgeon, Katheryn Doran, and Michelle Saint point out that right and wrong answers send the message that philosophy is not a free for all:

> In a discipline such as philosophy, where we hear all too often, "There is no right or wrong answer, right?" it can be important to convey to students that yes, there are right answers when it comes to whether you understand a philosopher's claims and arguments, and a mastery of the basic ideas must precede reflective analysis.[28]

When students feel that they are being graded fairly on the basis of mastery of real content, they are more likely to buy into the course. This is very helpful for students who may rightfully have the sense that the larger system is rigged against them.

So, multiple-choice questions allow us to consolidate subjectivity, increase the diversity of our evaluations, and in so doing increase the inclusiveness of our courses. I believe these three fairness-based arguments give a better picture of the issue than many of the arguments surrounding multiple-choice questions. Usually, debates about the appropriateness of multiple-choice questions are framed in terms

of the need for time savers versus the depth of authentic education. However, multiple-choice questions are not really much of a time saver. As I have said, good multiple-choice questions are hard to write. They take a lot of time. What is really happening is that some of the test-grading work is being moved into the test-writing phase. Over the long run time is saved by building up a large bank of reusable questions of proven quality when one teaches the same courses across semesters. But this depends largely on having stable course preparations and well-organized question banks. This saves some time but not as much as one would think.

Reply to a Criticism

The arguments against multiple-choice questions are often hard to parse, because they are so bound up with arguments against the rise of high-stakes standardized testing in the United States and Canada. The Canadian Center for Policy Alternatives' objection to standardized testing is representative. They say that when standardized tests are used,

> [c]ognitive processes of analysis, comparison, inference and evaluation are replaced by isolated skills that do not have transferability. The tyranny of the single right answer does not engage students in the tasks which require sustained reasoning or an explanation concerning their thinking processes. Teaching techniques that are effective in raising test scores conflict with the kind of instruction that develops critical thinking, problem solving, and creativity.[29]

Even in context, it is hard to tell what aspect of high-stakes standardized testing the authors are objecting to. However, the remark about "the tyranny of the single right answer" indicates that at least part of what they have in mind is the multiple-choice format. The entanglement of high-stakes testing with the use of multiple-choice tests is very unfortunate, because they are really separate issues. I am not interested in large-scale standardized testing, whether it is multiple-choice or essay testing. I just want people to know that multiple-choice questions are a legitimate and good option to use in the myriad of tests instructors use that are individualized for each course.

When people do directly criticize multiple-choice questions, the criticisms take numerous forms: they make it too easy for students to guess; they are too easy; they are too hard; they inevitably have multiple answers one could make a *prima facie* case for; they are unable to capture the complexity of real world situations; they are just fundamentally unphilosophical.[30] But, most of all, multiple-choice questions have an undeserved reputation for only being able to test student recall of basic facts. I hope I have put this myth to rest. Similarly, the above discussion of consolidating subjectivity addresses the worry about the inevitability of multiple *prima facie* correct answers. I want to end this section by responding to the objection that multiple-choice questions are unable to capture the complexity of real world situations.

The complexity critique comes up most often in the context of arguments against standardized tests specifically for critical thinking.[31] Trudy Govier's argument makes this most clear. She notes that

> [m]echanically gradable tests have to deal with articulated thought about small, easily described issues where answers do not diverge due to differences in political or ethical perspective or on the basis of varying background knowledge. . . . [I]nteresting figures of speech, irony and sarcasm, and suggestive ambiguities will have to be avoided.[32]

Similarly, Leo Groarke argues that standardized multiple-choice tests are not good tools for measuring critical thinking, partially because they must rely on made up examples that "are removed from the real-life contexts where critical thinking must take place."[33]

I am sympathetic to this kind of objection, but at some point we run into the problem of the map that is actual size. The map that is actual size is a concept that has been explored in literature and stand-up comedy.[34] A map of a nation that is actual size could obviously never be made. Such a map would defeat the whole purpose of having a map. Maps by their nature simplify. They are not just smaller than the things they represent, but they also ignore certain features and highlight others. This is also what we do when we create learning environments. We do not just give people the real world. We give them a structured, simplified version of the real world in order to make certain parts of it perceptually salient to the students. The process of scaffolding is basically a matter of providing students with a series of simplified experiences that build on each other until eventually they get something that begins to approximate the complexity of the real world. A final summative evaluation might have this level of complexity, but even for summative evaluations, it is important to use simplified representations of the world to check the development of component skills individually.

Writing Good Multiple-Choice Questions

The entire argument for multiple-choice questions depends on the instructor actually using good multiple-choice questions, which, as I have said, are very difficult to write. I end with a few pieces of advice for writing such questions. I have already discussed the most crucial suggestion, which originates with Robert Ennis: provide a prominent space where students have the option of explaining their answer. Ennis and Stephen Norris offer further sound advice on writing multiple choice questions in chapter four of their book *Evaluating Critical Thinking*. Their itemized advice for avoiding ambiguous or badly worded questions includes these helpful items:

1) Either use direct questions or incomplete statements as the question stem.
2) Write items in clear and simple language.

3) State the central problem of the item clearly and completely in the stem.
4) Include most of the reading in the stem.
5) Base each item on a single central problem.
6) Construct options homogeneous in grammatical form.
7) Include in the stem any words that would otherwise need repeating in each option.
8) Emphasize negative words or words of inclusion (e.g., "not," "except") and avoid such words when possible.
9) Place options at the end of the stem, not in the middle of it.
10) Arrange the options in a logical order, if one exists.[35]

In addition to creating unambiguous answers, good questions do not provide hints about the correct answer that are irrelevant to the thing you are trying to measure. Test writers may not think they are doing this, but unless they are making a conscious effort not to, they probably are. This can be seen most easily going back to one of the original pieces of advice given to test takers by the father of test prep schools, Stanley "The Cram King" Kaplan: "If guessing, a good rule of thumb is: the longest choice is often the correct one."[36] Kaplan's big insight was that there were ways to improve your score on tests that have nothing to do with the content of the test. They were just ways to game the system. The longest-answer trick is a great example. The correct answer to a question must be stated very precisely, so it will generally be a long answer. But test writers are often too busy or too lazy to work on good distractors, and so distractors wind up being quite short. A textbook I use comes with an extended question package, where every question can be gamed this way. Test prep services teach a wide variety of "Kaplanning" techniques: 1) If two answers are really similar, one of them is probably the correct answer, 2) absolute statements tend to misrepresent information, and 3) avoid answers with a radically different tone than the prompt.

I have two pieces of advice regarding questions that are "Kaplannable." First, do not write test questions that can be "Kaplanned." Look over the test. Is the longest answer consistently the right one? Change that. If a test writer is in a hurry, using correct answers from other questions is a quick way to generate long distractors. Stanley Kaplan and his company were eventually embraced by the testing establishment when they realized that the challenges he was posing just forced them to write better tests.[37] Other test writers should have the same response.

Second, give students Kaplan-style advice for how to answer multiple-choice questions. Some Kaplanning techniques are like the answer-length trick: they allow the student to guess the correct answer using information that is entirely irrelevant to what the question is trying to measure. More often, though, the techniques are drawing on skills we are trying to teach and measure. Test prep services advise

test takers to rule out answers with a different tone and point out that absolute expressions often misrepresent. We want students to be able to recognize the tone of a passage. We want students to be wary of over generalizations. These may not be the exact thing the question is trying to measure, but they are not something completely arbitrary like answer length either. If you give Kaplan-style lessons in how to answer multiple questions you will simultaneously give a critical thinking lesson and a college survival lesson.

The usefulness in teaching critical thinking skills is especially important to emphasize here. Earlier, I noted that students might be thrown off by distractors that were true statements, but not premises in the argument. My response to those who said this might be unfair was to suggest that one continue to use such questions but teach students how to study for them. Such a lesson is again both a critical thinking and a college survival lesson. Michelle Saint similarly points out that if a distractor is of the form "A and B," and A is true, but B is false, students might still be drawn to it.[38] This is an instance of what Amos Tversky and Daniel Kahneman call the "conjunction fallacy."[39] We have a cognitive bias to average things, even when we should be combining them in some other way. The same fallacy explains why adding one low-value baseball card to a collection of high-value cards will lower the market value of the whole collection, even though all the high-value cards are still there.[40] Potential buyers average values when they should be summing them. Teachers can help students by using test preparation as an opportunity to warn people against this mistake.

Conclusion

Evaluating students is one of the most complicated and politically fraught parts of education, which is in turn one of the most complicated and politically fraught aspects of human society. Administrators and politicians have a standing interest in reducing that complexity to a single number, which they can then boast that they have improved. Evaluation techniques get caught up in this struggle to industrialize education. Any technique that seems like it can give the managers the single number they are looking for is praised by the managers, and this in turn creates a backlash against that technique from the people who want to resist industrialization. But there is nothing in the techniques themselves that the managers are promoting that requires us to use them reductively. In fact, a nonreductive approach to education is going to be by nature inclined to include them, precisely because it is nonreductive. In philosophy, our evaluation tools are generally proxies and rough indicators. This means that our evaluations need to be frequent, low stakes, and diverse. Mechanically gradable questions have an important role to play in such a program.

Appendix

Examples of Questions at Various Levels of Bloom's Taxonomy

The headings in this appendix follow Bloom's revised taxonomy. The numbers are not sequential, because not all categories from Bloom's taxonomy are implemented. The subheadings classify questions that evaluate grasp of concepts or grasp of arguments, and within those categories, further headings classify which aspects of concepts or arguments are tested. As before, correct answers are in italics, and the penalty for incorrect answers is given in parentheses on a ten point scale.

1.1 Remembering: Recalling

If a multiple-choice question asks about the intension of a philosophical concept, it is asking about the lexical definition of the concept, which is at the lowest level of Bloom's taxonomy— remembering. These questions rely on verbal memory, and one can study for them using flashcards. Questions can be made more challenging for students by substituting a constructed response question that is equivalent to the selected response question without significantly adding to grading time or really affecting the fairness. It is also really easy to group several level 1.1 questions into matching questions to increase variety on a test. I give two kinds of multiple-choice questions at this level. Questions one and two ask students to select the single best answer, while questions three and four ask students to identify several correct answers among a larger set.

Concept Questions: Intension

Term to Intension

1) Which of the following is the best definition of good moral luck? (Select one)
   a. Moral luck occurs when one chooses the right thing by accident. (-*10*)
   b. Moral luck consists of factors that are not in a person's control that nevertheless lead fully informed people to say that they are morally good.
   c. Moral luck consists of being given the opportunity to choose the right thing. (-*7*)
   d. Moral luck consists of positive character traits that are the result of genetics and upbringing, and not personal choice or deliberate effort. (-*4*)

   Explain your answer (optional):

Intension to Term

2) Which of the following terms is used to describe factors that are not in a person's control that nevertheless lead fully informed people to say that they are morally good? (Select one)
   a. Free will (-7)
   b. Determinism (-4)
   c. Moral luck
   d. Moral competence (-10)

Explain your answer (optional):

Important Properties

3) Which of the following statements are true of moral luck? (Select all that apply.)
   a. One counts as morally lucky if one commits a crime and does not get caught.
   b. One's moral luck is made up of factors beyond one's control.
   c. Moral luck is about the way fully informed people would evaluate you morally.
   d. Moral luck poses a philosophical problem, even if one accepts that free will exists.

Explain your answer (optional):

Subcategories

4) Which of the following are kinds of moral luck? (Select all that apply)
   a. The good fortune to be honored for virtues we do not possess
   b. The good fortune of having no morally bad unanticipated consequences result from our choices and actions
   c. The good fortune of not getting caught for the crimes we have committed
   d. The good fortune of having a good temperament and an optimistic personality
   e. The good fortune of not being thrown into morally difficult circumstances

Explain your answer (optional):

Having both sorts of questions is useful for getting at the same concept in multiple ways. Also, note that for the "choose one" questions, some distractors are more wrong than others and can cost the student more points. The amount of deduction, indicated in parentheses, reflects the degree to which a wrong answer is wrong. Sometimes I add an additional penalty for selecting answers that directly contradict each other. This allows teachers to rate students more precisely and can empower teachers to ask about more subtle distinctions. The least wrong answer for question one is actually the definition of a subtype of moral luck. Distinguishing type from subtype might seem too subtle a thing to ask students to do, but if teachers are only taking four points off, it is not a big deal. Note also that I tend to use four or five options for multiple-choice questions. There is an

extensive literature on how many options should be used in a multiple-choice test, with many studies pointing to three as the optimal number.[41] I prefer four or five options, because I want to maximize the sensitivity of my scale for partial credit.

Scoring on the "select all that apply" questions also allows for more precise evaluation. I have a somewhat unusual grading method for "select all that apply" questions. I grade each item individually and take off points either if it was marked, but should not have been marked, or was left unmarked when it should have been marked. Effectively, each item is a true/false question. The standard method of grading such questions, which is built into most if not all learning management systems, looks at the percentage of correct answers that were checked by the student and then multiplies that by the percentage of incorrect options that were left blank. This means that the student does not get any credit when they leave an option blank when it should in fact have been left blank. It also over penalizes false-positive responses. This asymmetry is meant to penalize "guessing," but, as I argued above, "educated guesses" are continuous with critical thinking skills that philosophers actually want to teach and evaluate. So, I reject the asymmetry between selecting and not selecting that is built into many learning management systems.

2.2 Understanding: Exemplifying

Levels two and three of Bloom's taxonomy are understanding and applying. At these levels, one does not just repeat a verbal formulation, one understands its content and can apply it in novel situations. Lorin W. Anderson et al. note that if "assessment tasks are to tap higher-order cognitive processes, they must require that students cannot answer them by relying on memory alone."[42] Moving to the extension of a term allows us to introduce novel examples and ask students how to classify them. In level 2.2 (exemplifying) students are given a concept in the prompt and have to decide which novel examples fall under that prompt. Questions five and six do this for the ideas of moral luck and internalization.

Concept Questions: Extension

Term to Extension

5) Which of the following people have experienced good moral luck? (Select all that apply)
    a. George robs a bank and, because the security system malfunctions, escapes with a million dollars to Mexico, where he lives out his life on a beach resort under an assumed name.
    b. Ahmed was born with a strong genetic predisposition to alcoholism and violent outbursts. Fortunately, he lives his whole life in Muslim countries where alcohol is illegal and almost impossible to find. Without the toxic effect of alcohol, he finds it easy to control his temper.

c. Susan was married for fifty years to a cheating husband, Bob, but never found out about his activities. She goes to her grave thinking only good thoughts about him. (Remember that the question is whether Susan is morally lucky, not whether her husband is.)

d. A combination of good genes, loving parents, and a strong religious background leads Hans to grow up with courage and strong moral vision. When the Nazis take power, he sees the threat and risks his own life to get hundreds of Jews safely to what is now Israel. He is remembered for generations as a hero.

e. For five years Jenny was in the habit of horsing around with an unloaded handgun at parties, pretending to shoot her friends. Fortunately, the time she accidentally left a bullet in the chamber, the gun never went off.

f. Bob wins the lottery, quits his job, invests his money wisely, and lives lavishly off the interest for the rest of his life. He travels the world, sends his children to the finest schools, and so forth.

Explain your answer (optional):

6) Which of the following are examples of people who have internalized moral norms? (Select all that apply)

a. Steve grows up in a drug- and crime-ridden community. He decides not to join a gang for the sole reason that he fears going to jail or being killed by a member of another gang.

b. Barbara's employer trusts her to deposit the cash that came into the store in the bank at the end of each day. Barbara could easily pocket thousands of dollars and then alter the receipts so that no one would ever know the difference. She doesn't do this, though, because she believes that stealing is wrong and doesn't want to violate the trust of her employer.

c. Bill has problems with explosive anger. He is aware of this and has taken anger management classes. After a particularly hard day at work, he gets into an argument with his wife Jenny and punches her in the face. He instantly feels ashamed of what he did. Unable to look his wife in the eye, he locks himself in the bedroom to sulk. As his shame mixes with guilt, he realizes that he must make amends. Summoning his courage, he goes downstairs, looks Jenny in her now blackened eye, and begs forgiveness. They agree to go into couples' therapy and that he should stay with his brother for a while.

d. Sid is a stone-cold killer. He has to be if he wants to stay on top of his drug ring. His enemies need to know he is willing to have them killed. His subordinates need to know he is willing to personally kill them. Preachers and policemen have told him many times that gang life is immoral, but their lectures are all just words to him.

Explain your answer (optional):

2.3 Understanding: Classifying

In level 2.3 (classifying), students are given a novel example in the prompt and are asked to find the concept under which it falls. Question seven does this for the concept of validity. Questions eight and nine combine asking about both extension and intension in a way that requires students to justify their answers to an earlier question. Question eight asks students to correctly classify a novel example, so it is about extension. Question nine asks students to justify the answers they gave to question seven, so it is about intension. Often the grading scale I use for questions of the second type vary depending on how the previous question was answered. This kind of justification is exactly the sort of thing multiple-choice questions are sometimes thought to not allow for.

Concept Questions: Extension

Extension to Term

7) You are talking with your friend Cindy, and she puts forward an argument with premises that you completely disagree with. You would never in a million years take those premises to be true. However, you have to admit, that if those premises were true, Cindy's conclusion would absolutely have to be true. Which of the following terms best describes Cindy's argument from your perspective? (Select one)
   a. strong (*-4*)
   b. valid
   c. persuasive (*-10*)
   d. sound (*-4*)

   Explain your answer (optional):

Combine Term, Intension, and Extension While Asking for Reasons

8) Fritz is adamantly pro-life, and when asked why, he replies that the Bible says "Thou shall not kill." This rule, though, seems to imply that all war is morally wrong. He considers that maybe the rule has been misinterpreted and that really the rule should be "Thou shall not take an innocent life." After all, he thinks, what could be more innocent than a baby? But innocent people die in war, too. Even the enemy soldiers may have signed up for morally good reasons—to defend their country, for instance. Based on this reasoning, Fritz decides that in addition to being pro-life, he should be a pacifist. What is the name given in this course for the reasoning process Fritz went through? (Select one)
   a. Values clarification (*-4*)
   b. Reflective equilibrium
   c. Moral knowledge (*-10*)
   d. Cultural relativism (*-7*)

   Explain your answer (optional):

9) Which is the best reason for giving the answer you gave above? (Select one)
   a. because he is deciding what he believes about a moral issue (-*10*)
   b. because he is weighing his intuitions about individual cases, like the death penalty and killing in war, against the general rule "thou shall not take innocent life"
   c. because he is looking at a rule, "thou shall not take innocent life," and making its value clearer (-*7*)
   d. because he is torn between different ideas about what the right thing to do is, but then he reaches a stable conclusion (-*4*)

Explain your answer (optional):

2.4 Understanding: Summarizing

Level 2.4 is summarizing, which is a commonplace activity in the humanities. Below is a question about Hume where the student has to identify the speaker and interpret the passage. These questions might also fall under "2.1 Understanding: Interpreting" or even "4.3 Analyze: Attribute." For all three of these categories, Andersen et al. give examples that amount to "what is the main point of this passage," a staple in standardized testing at the primary and secondary level. One might argue that the example below does not evaluate a sophisticated cognitive process by the standards of the Bloom committee because the passage is from the students' reading and thus not new to the students. But given the difficulty students have reading Hume, the odds that students will simply recognize the passage and pull the speaker and meaning from their memory are low. Students will probably have to read the passage and construct its meaning, aided by a general familiarity with the issue. Memory in general is reconstructive, so the line between level 1 questions and higher level questions is unlikely to be very bright in any case.[43]

Concept Questions

For questions ten and eleven, consider the following passage:

> "It is my opinion, I own, replied ____, that each man feels, in a manner, the truth of religion within his own breast, and, from a consciousness of his imbecility and misery, rather than from any reasoning, is led to seek protection from that Being, on whom he and all nature is dependent. So anxious or so tedious are even the best scenes of life, that futurity is still the object of all our hopes and fears. We incessantly look forward, and endeavour, by prayers, adoration, and sacrifice, to appease those unknown powers, whom we find, by experience, so able to afflict and oppress us. Wretched creatures that we are! What resource for us amidst the innumerable ills of life, did not religion suggest some methods of atonement, and appease those terrors with which we are incessantly agitated and tormented?"

10) Who is speaking here? (Select one)
    a. Philo (-5)
    b. Cleanthes (-5)
    c. Demea
    d. Pamphilus (-*10*)

Explain your answer (optional):

11) Which of the following statements best summarizes the passage? (Select one)
    a. Everyone instinctively knows the truths of religion, because they are revealed by a rational examination of the world. (-5)
    b. No matter how hard life seems, no matter how much suffering you have to endure, you can always know that God doesn't throw anything at you that you can't handle. (-*10*)
    c. Everyone instinctively knows the truths of religion, because life is miserable and religion offers the only hope.
    d. Everyone instinctively feels the misery of life, and therefore doubts the existence of an all-good creator. (-5)

Explain your answer (optional):

3.1 Applying: Executing

Level 3 is similar to level 2, in that one has to apply something one learned to a novel situation. The difference is that in level 3 the student executes a process rather than applies a concept. The level is divided into "executing" (3.1) and "implementing" (3.2), based on how similar the novel situation is to the training situation and how much the procedure needs to be modified. For my example I use the process of reflective equilibrium. This question could be classified as either level 3.1 or 3.2, depending on how the prior learning was conducted.

Concept Questions

12) Cameron has always lived by the principle that the good of the many outweighs the good of the few. So, when they first heard about the classic version of the trolley problem, they thought it was obvious that one should throw the switch. But when they heard the fat man version of the trolley problem, they were perplexed. Cameron recognized the process they were going through as a case of reflective equilibrium and decided the thing to do was to stick by the rule, rejecting their intuition about the case. Which of the following best describes the outcome of their decision? (Select one)
    a. Cameron will say you should both pull the switch and push the fat man.
    b. Cameron will say you should pull the switch, but not push the fat man. (-5)

c. Cameron will say you should neither pull the switch nor push the fat man. (*-5*)
d. Cameron will say you should push the fat man but not pull the switch. (*-10*)

Explain your answer (optional):

4.1 Analyze: Differentiate

Examples of argument questions targeting levels 4 and 5 are discussed in the body of this article. Below are additional questions one could ask that rely on argument to raise the sophistication level.

For questions like these, students can use their understanding of how argumentation works to compensate for imperfect recollection of the reading. The different kinds of distractors used here offer different ways to do that. In questions thirteen and fourteen, for instance, the distractors are items that cannot serve as reasons in an argument, or even go against the conclusion of the argument. Question fifteen includes a distractor that would make the argument circular. Question sixteen uses a distractor that is a reason but not one that the author would invoke. Distractors such as these will throw off some students. If they are being asked to find reasons that support a statement, and one of the options is a reason to oppose the statement, they may not understand why it is even an option in the first place. However, a sophisticated student will recognize that this simply means the options must be incorrect answers. One might think this is letting the student get away with something, because it means that someone who has not even done the reading can figure out the correct answer, but really this is substituting a higher level skill for a lower level one. It helps if prior to an exam containing questions such as this, teachers explain that if they are trying to find reasons a statement might be true, they should not be confused by reasons that would actually show the statement is false. This is again a lesson both in test taking and critical thinking. Question seventeen asks the student to fill in missing premises in an argument given in a reading. Questions like seventeen can also be done as constructed response questions, which are just as quick to evaluate as multiple-choice questions.

Questions that Just Ask What Premises or Conclusions an Author Used

Distractors That Are Reasons to Reject the Conclusion

13) Which of the following are reasons Kant gives for saying that reason is the only thing that can serve as the justification and motivation for moral behavior? (Select all that apply.)
    a. Emotions are not stable, so a morality motivated by emotion will not last.
    b. Emotions have no cognitive content, so they cannot be used to judge right and wrong.
    c. People who lack emotions are unable to find any meaning in life, so a morality without emotion would not motivate.

d. Emotions are closely linked to the right and wrong things to do, so a morality based in emotion has more than an accidental link to goodness.
e. People who act out of emotion are only satisfying their own needs, so a morality founded on emotion is not morally worthy.

Explain your answer (optional):

Distractors That Are True Statements, But Not Used in the Argument

14) Which of the following are *criticisms* that have been made of the use of proverbs in deliberation, according to the text and our classroom discussion? (Select all that apply.)
a. Proverbs can aid memory by serving as a repository of past judgment.
b. Proverbs are often vague and contradictory.
c. Proverbs are platitudes given in place of thinking.
d. A proverb is a short, commonly repeated saying that reflects the wisdom of a culture.

Explain your answer (optional):

Distractors That Would Make the Argument Circular

15) Which of the following are reasons Mill gives for the harm principle? (Select all that apply)
a. People have no right to interfere with someone's behavior if they are not harming others.
b. Outsiders are less likely than you to know what is good for you.
c. Sometimes people don't know what's best for them, so an outsider needs to make that choice.
d. Outsiders are less likely to care what is good for you.
e. People are better off making their own choices, even when they are not making the best choices.

Explain your answer (optional):

Distractors That Are Reasons, But Not Ones Invoked by the Author

16) Which of the following are reasons Radcliffe-Richards et al. give for allowing kidney sales? (Select all that apply)
a. A ban on kidney sales hurts both the potential sellers and the potential buyers.
b. Potential sellers can be given extensive counselling and information, so we can be sure their participation is based on informed consent.
c. If it is true that you own your body, you must also be allowed to sell it.
d. Purchasing of kidneys can be done by a central body to ensure just distribution.

Explain your answer (optional):

Fill in Missing Premises in an Argument Given in a Reading

17) Consider the following argument, often called the argument from non-paradigm humans or the argument from marginal cases:

P_1 Suppose we tried to deny animal rights by using sapience as a criterion for moral status

P_2 We would then also deny moral status to non-paradigm humans, such as the permanently mentally disabled.

P_3 _____

C: Therefore, we cannot deny animal rights by using sapience as a criterion for moral status.

What is the missing premise of the argument?

a. We do not want to deny moral status to the permanently mentally disabled.

b. The permanently mentally disabled are still human beings.

c. The permanently mentally disabled include people with Down syndrome.

d. We should not deny rights to animals.

Explain your answer (optional):

4.2 Analyze: Organize

The following questions ask students to pick apart an argument. The final question is obviously not a selected response question, but it was part of the set and illustrates how these questions can be integrated.

For questions eighteen through twenty, consider the following passage.

At a college meeting, faculty are considering changing student loan policy to prevent fraud.

Faculty Member 1: You know these students. When they get their loan checks, they go straight to the bookstore and buy an iPod and a bunch of fancy gear.

Faculty Member 2: Do you have any evidence of that?

Faculty Member 1: Just ask anyone who works at the bookstore!

18) Which of the following is the best way to represent this argument in canonical form? (Select one)

a. P_1: When students get their loan checks, they buy an iPod and a bunch of fancy gear

C: Just ask anyone who works at the bookstore!

b. P_1: Any employee at the bookstore will tell you students use loan money to buy iPods and fancy gear

C: Students just use loan money to buy iPods and fancy gear

c. P_1: Do you have any evidence of that?

C: Students just use loan money to buy iPods and fancy gear

d. P_1: You know these students.

C: Just ask anyone who works at the bookstore!

Explain your answer (optional):

19) Which of the following best describes the argument? (Select one)
   a. It is an argument from authority, specifically the authority of faculty members.
   b. It is an ad hominem attack against bookstore employees.
   c. It is an ad hominem attack against students.
   d. It is an argument from authority, specifically the authority of any bookstore employee.

Explain your answer (optional):

20) Evaluate the argument, including assessing the truth of the reasons, the strength of the reasoning, and any warrants that should be added to it.

5.1 Evaluate: Check

The highest Bloom level where one can use selected response questions is 5.1, "Evaluate: Check." Anderson et al. specifically use "does this premise support this conclusion" and "is there a fallacy in this argument" as examples here.[44] Anderson et al. do not specify that selected response questions can be used for this, but teachers of critical thinking should recognize this type of question.

21) Consider the following passage

Kimmy Schmidt, a put-upon but resilient sitcom character on Netflix's *The Unbreakable Kimmy Schmidt*, says, "You can get through anything. I learned a long time ago that a person can stand just about anything for ten seconds. Then you just start on a new ten seconds."

Assume this is meant as a rational argument for the conclusion: "You can get through anything," and not as general encouragement and advice to take everything one day at a time. Which of the following best describes the fallacy Kimmy is committing here?

   a. Red herring (*-7*)
   b. Straw man (*-10*)
   c. Slippery slope (*-4*)
   d. Continuum fallacy
   e. No fallacy is being committed here. This is a good way to reason. (*-7*)

Explain your answer (optional):

Notes

This article was inspired by workshops given by Bob Ennis and Michael Scriven that I attended at AAPT workshop-conferences years ago (Scriven, "Novel Approaches to Testing"; Ennis, "Appraising Critical Thinking Tests"; and Ennis, "Writing Critical Thinking Test Items"). Later, I held my own workshop on multiple-choice questions (Loftis, "Beyond Information Recall"), which became the basis for this article. Thanks to everyone involved with all these events.

1. Howe, "Evaluation Primer," 325.

2. See, for instance, Possin, "A Field Guide to Critical-Thinking Assessment," 212.

3. Scriven, "Novel Approaches to Testing in Philosophy."

4. Bloom, *Taxonomy of Educational Objectives*.

5. Anderson et al., *A Taxonomy for Learning, Teaching, and Assessing*.

6. When Anderson et al. give individualized descriptions of each level of their taxonomy, each description has a subsection entitled "Assessment Formats" (Anderson et al., 69–88). For the taxonomy levels marked with a "yes," Anderson et al. list in that subsection a form of assessment they label "selected response" or "multiple-choice." For the taxonomy levels marked with a "?," they describe assessment formats where one can imagine using selected response questions, even though they do not explicitly say so.

7. Anderson et al., 66.

8. The pyramid on the left-hand side of the figure is taken from A North Carolina History Online Resource (https://www.ncpedia.org/media/blooms-taxonomy-revised) and used under a Creative Commons BY-CA-SE license (https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode).

9. Howe, "Evaluation Primer," 323.

10. For good examples, see Govier, *Problems in Argument Analysis and Evaluation* and Possin, "A Field Guide to Critical Thining Assessment." For less compelling critiques, see Groarke, "What's Wrong with the California Test?"

11. Ennis, "Critical Thinking Assessment."

12. Ennis, "Writing Critical Thinking Test Items for Classroom Use."

13. Ennis, "Critical Thinking Assessment," 184.

14. For a scientific review of the ego depletion model, see Hagger, Wood, et al., "Ego Depletion and the Strength Model"; and Inzlicht and Schmeichel, "What Is Ego Depletion?" For a more popular discussion, see Baumeister and Tierney, *Willpower: Rediscovering the Greatest Human Strength*.

15. Danziger, Levav, and Avnaim-Pesso, "Extraneous Factors in Judicial Decisions."

16. Hagger, Wood, et al., "Ego Depletion and the Strength Model"; Inzlicht and Schmeichel, "What Is Ego Depletion?"; and Baumeister and Tierney, *Willpower: Rediscovering the Greatest Human Strength*.

17. Danziger, Levav, and Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," 6890.

18. Hagger, Chantzisarantis, et al., "A Multilab Preregistered Replication of the Ego-Depletion Effect." See also Baumeister, Vohs, and Tice, "The Strength Model of Self-Control."

19. Hartshorne and Schachner's "Tracking Replicability" was influential in establishing the scope of the problem for psychology. For a popular presentation of the current state of the replication crisis, see Yong, "Psychology's Replication Crisis Is Running Out of Excuses."

20. Open Science Collaboration, "Collaborative Effort to Estimate the Reproducibility of Psychological Science," and "Estimating the Reproducibility of Pscyhological Science."

21. Hagger, Chantzisarantis, et al., "A Multilab Preregistered Replication of the Ego-Depletion Effect."

22. The maxim has different names and different phrasings in different disciplines. In economics, it is known as Goodhart's law. See Goodhart, "Problems of Monetary Management." In sociology, it is known as Campbell's law, and can be traced back to Campbell, "Assessing the Impact of Planned Social Change."

23. Campbell, Assessing the Impact of Planned Social Change," 51.

24. Collins, "Examining Philosophy," 147.

25. Howe, "Evaluation Primer," 322.

26. Collins, "Examining Philosophy," 147.

27. Collins, 146.

28. Turgeon, Doran, and Saint, "Helping Students with Multiple Choice Exams."

29. Meaghan and Casas, "Bias in Standardized Testing," 37.

30. Govier, *Problems in Argument Analysis*, 265; Groarke, "What's Wrong with the California Test?," 50; and Collins, "Examining Philosophy," 145.

31. Govier, *Problems in Argument Analysis*, 248; Groarke, "What's Wrong with the California Test?," 47; and McPeck, *Critical Thinking and Education*, 145.

32. Govier, *Problems in Argument Analysis*, 248.

33. Groarke, "What's Wrong with the California Test?," 50.

34. Carroll, *Sylvie and Bruno Concluded*; Borges, "On Exactitude in Science"; and Wright, Saturday Night Live Appearance, January 14, 1984.

35. Norris and Ennis, *Evaluating Critical Thinking*, 108–109.

36. Caldwell, "The Opposite of Education."

37. Caldwell.

38. Turgeon, Doran, and Saint, "Helping Students with Multiple Choice Exams."

39. Tversky and Kahneman, "Extensional versus Intuitive Reasoning," 293–315; and Kahneman, *Thinking, Fast and Slow*, 158–65.

40. List, "Preference Reversal of a Different Kind," 1636–43.

41. For a review, Rodriguez, "Three Options Are Optimal for Multiple-Choice Items."

42. Anderson et al., *A Taxonomy for Learning, Teaching, and Assessing*, 71.

43. Schacter, Norman, and Koutstaal, "The Cognitive Neuroscience of Constructive Memory," 289–318.

44. Anderson et al., *A Taxonomy for Learning, Teaching, and Assessing*, 83.

References

Anderson, Lorin W., David R. Krathwohl, Peter Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.

Baumeister, Roy F., and John Tierney. *Willpower: Rediscovering the Greatest Human Strength*. New York: Penguin Publishing Group, 2011.

Baumeister, Roy F., K. D. Vohs, and D. M. Tice. "The Strength Model of Self-Control." *Current Directions in Psychological Science* 16(6) (2007): 351–55. https://doi.org/10.1111/j.1467-8721.2007.00534.x

Bloom, Benjamin S. (Ed.), Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain*. New York: Longmans, Green and Co., 1956.

Borges, Jorge Luis. "On Exactitude in Science." In *A Universal History of Infamy*. Translated by Norman Thomas de Giovanni. New York: Penguin, 1972.

Caldwell, Christopher. "The Opposite of Education." *The Financial Times*, August 28, 2009. https://www.ft.com/content/f6539326-9400-11de-9c57-00144feabdc0.

Campbell, Donald T. "Assessing the Impact of Planned Social Change." Public Affairs Center, Dartmouth College, Occasional Papers Series 8, 1976.

Carroll, Lewis. *Sylvie and Bruno Concluded*. London: MacMillan and Co., 1893.

Collins, Peter. "Examining Philosophy: 'Choose the Best Answer.'" *Teaching Philosophy* 16(2) (1993): 145–54. https://doi.org/10.5840/teachphil199316215

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences of the United States of America* 108(17) (2001): 6889–92. https://doi.org/10.1073/pnas.1018033108

Ennis, Robert H. "Appraising Critical Thinking Tests." Presented at the American Association of Philosophy Teachers 16th Beinnial Workshop/Conference, Washington, PA, August 2–6, 2006.

Ennis, Robert H. "Critical Thinking Assessment." *Theory into Practice* 32(3) (1993): 179–86. https://doi.org/10.1080/00405849309543594

Ennis, Robert H. "Writing Critical Thinking Test Items for Classroom Use." Presented at the American Association of Philosophy Teachers 16th Beinnial Workshop/Conference, Washington, PA, August 2–6, 2006.

Govier, Trudy. *Problems in Argument Analysis and Evaluation*. Dordrecht: Foris, 1987. https://doi.org/10.1515/9783110859249

Groarke, Leo. "What's Wrong with the California Critical Thinking Skills Test? CT Testing and Accountability." In *Critical Thinking Education and Assessment: Can Higher Order Thinking Be Tested?*, edited by Jan Sobocan and Leo Groarke, 35–54. London, ON: Althouse Press, 2009.

Hagger, Martin S., N. L. D. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, D. P. Calvillo, W. K. Campbell, P. R. Cannon, M. Carlucci, N. P. Carruth, T. Cheung, A. Crowell, D. T. D. De Ridder, S. Dewitte, M. Elson, J. R. Evans, B. A. Fay, B. M. Fennis, A. Finley, Z. Francis, E. Heise, H. Hoemann,

M. Inzlicht, S. L. Koole, L. Koppel, F. Kroese, F. Lange, K. Lau, B. P. Lynch, C. Martijn, H. Merckelbach, N. V. Mills, A. Michirev, A. Miyake, A. E. Mosser, M. Muise, D. Muller, M. Muzi, D. Nalis, R. Nurwanti, H. Otgaar, M. C. Philipp, P. Primoceri, K. Rentzsch, L. Ringos, C. Schlinkert, B. J. Schmeichel, S. F. Schoch, M. Schrama, A. Schütz, A. Stamos, G. Tinghög, J. Ullrich, M. vanDellen, S. Wimbarti, W. Wolff, C. Yusainy, O. Zerhouni, and M. Zwienenberg. "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11(4) (2016): 546–73.

Hagger, Martin S., Chantelle Wood, Chris Stiff, and Nikos L. D. Chatzisarantis. "Ego Depletion and the Strength Model of Self-Control: A Meta-Analysis." *Psychological Bulletin* 136(4) (2010): 495–525. https://doi.org/10.1037/a0019486

Hartshorne, J. K., and A. Schachner. "Tracking Replicability as a Method of Post-Publication Open Evaluation." *Frontiers in Computational Neuroscience* 6 (2012): 1–14. https://doi.org/10.3389/fncom.2012.00008

Howe, Kenneth. "An Evaluation Primer for Philosophy Teachers." *Teaching Philosophy* 11(4) (1988): 315–28. https://doi.org/10.5840/teachphil198811478

Inzlicht, Michael, and Brandon J. Schmeichel. "What Is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control." *Perspectives on Psychological Science* 7(5) (2012): 450–63. https://doi.org/10.1177/1745691612454134

Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

List, John A. "Preference Reversal of a Different Kind: The 'More Is Less' Phenomenon." *American Economic Review* 92(5) (2002): 1636–43. https://doi.org/10.1257/000282802762024692

Loftis, J. Robert. "Beyond Information Recall: A Workshop on Sophisticated Multiple Choice Questions in Philosophy." Presented at the American Association of Philosophy Teachers 21st Beinnial Workshop/Conference, Saginaw, Michigan, July 27–31, 2016.

McPeck, John. *Critical Thinking and Education*. New York: St. Martin's Press, 1981.

Meaghan, Diane E., and François R. Casas. "Bias in Standardized Testing and the Misuse of Test Scores." In *Passing the Test: The False Promise of Standardized Testing*, edited by Marita Moll, 35–50. Ottawa, ON: Canadian Center for Policy Alternatives, 2004.

Norris, Stephen P., and Robert H. Ennis. *Evaluating Critical Thinking*. Pacific Grove, CA: Midwest Publications Critical Thinking Press, 1989.

Open Science Collaboration. "Estimating the Reproducibility of Pscyhological Science." *Science* 349(6251) (2015): 943–45. https://doi.org/10.1126/science.aac4716

Open Science Collaboration. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Perspectives on Psychological Science* 7(6) (2012): 657–60. https://doi.org/10.1177/1745691612462588

Possin, Kevin. "A Field Guide to Critical-Thinking Assessment." *Teaching Philosophy* 31(3) (2008): 201–28. https://doi.org/10.5840/teachphil200831324

Rodriguez, Michael C. "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research." *Educational Measurement: Issues & Practice* 24(2) (2005): 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Schacter, Daniel L., Kenneth A. Norman, and Wilma Koutstaal. "The Cognitive Neuroscience of Constructive Memory." *Annual Review of Psychology* 49(1) (1998): 289–318. https://doi.org/10.1146/annurev.psych.49.1.289

Scriven, Michael. "Novel Approaches to Testing in Philosophy." Presented at the American Association of Philosophy Teachers 15th Beinnial Workshop/Conference, Toledo, Ohio, August 4–8, 2004.

Turgeon, Wendy, Katheryn Doran, and Michelle Saint. "Helping Students with Multiple Choice Exams." *APA Blog: The Teaching Workshop*. January 26, 2016. https://blog.apaonline.org/2016/01/21/the-teaching-workshop-helping-students-with-multiple-choice-exams/.

Tversky, Amos, and Daniel Kahneman. "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90(4) (1983): 293–315. https://doi.org/10.1037//0033-295X.90.4.293

Wright, Steven. *Saturday Night Live*, Season 9, episode 9, aired January 14, 1984, on NBC.

Yong, Ed. "Psychology's Replication Crisis Is Running Out of Excuses." *The Atlantic*, last modified November 19, 2018. https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.