Perspectives In Learning

Volume 20
Number 1 *TiLT Special Issue*

Article 8

2-27-2023

A Multiple-Choice Study: The Impact of Transparent Question Design on Student Performance

John LeJeune
Georgia Southwestern State University, john.lejeune@gsw.edu

A Multiple-Choice Study: The Impact of Transparent Question Design on Student Performance

John LeJeune
Georgia Southwestern State University

Abstract

This university classroom study seeks to better understand how, and to what extent, designing more transparent (or TILTed) multiple-choice questions would impact student performance. Ninety-two students in an introductory American Government class were randomly assigned "TILTed" and "unTILTed" versions of thirty-five test questions. Questions were "TILTed" and "unTILTed" in one of three ways—involving either (a) adding or eliminating unnecessarily difficult vocabulary from the stem; (b) adding or eliminating "all-of-the-above" and "none-of-the-above" answer options; or (c) adding or omitting additional cues or context. Statistical analysis showed that TILTing questions generally increased student scores, with twelve questions showing positive statistical significance at the $p \leq .10$ level. The most robust positive effects involved simplifying question vocabulary and avoiding all-of-the-above options (none-of-the-above was not examined in isolation). Adding additional cues and context produced mixed, and in some cases negative, results.

Transparent instruction, or TILT (Transparency in Learning and Teaching), has "demonstrably increased several important predictors of college student success: academic confidence, sense of belonging in college, metacognitive self-awareness of skill development, and persistence" (Winkelmes, 2019a, p. 1). These benefits are especially pronounced for "first-generation, low-income, and underrepresented college students" (Winkelmes, 2019a, p. 6).

Noting this, transparent instruction is especially important to consider at higher education institutions where many students arrive from disadvantaged or historically underserved backgrounds. For a variety of reasons, including job and family responsibilities, lack of confidence, less academic preparation, and lack of campus connectedness or familiarity with the college experience (Ma & Shea, 2021; Stebleton & Soria, 2012; Unverferth et al., 2012), these students face high obstacles to academic and career success, particularly in their first years of college. And in recent years, transparent instruction has emerged as one of the most credible, flexible, and easy-to-implement classroom strategies to address these challenges, providing benefits to all students, but especially to students of underprivileged backgrounds.

The essence of TILT is communicative, and it entails presenting classroom assignments broken down into

three component parts (Winkelmes, 2019b). First, the "Purpose" section relates the skills, knowledge, and practical relevance of the assignment. This helps students understand why the assignment exists, and why its tasks are structured as they are. Second, the "Task" section outlines the requirements of the assignment, usually as a set of sequential action steps. By minimizing ambiguity, presenting assignments in terms of discrete and manageable parts, and (ideally) offering successful models to emulate, students will approach their assignments as intended, with a concrete sense of efficacy. Finally, the "Criteria" section provides a rubric or grading framework that explains how submitted work will translate into grades. This section defines the qualities of successful work, and shows how different parts of the assignment will be weighted.

Transparent communication has many positive effects. Not only does it help students effectively tackle their assignments and succeed overall; it also helps overcome the alienation or "imposter syndrome" that first-generation students often feel on college campuses (Stebleton & Soria, 2012, p. 16). Notably, students facing academic or scholastic presentation styles for the first time may mistakenly process their discomfort with the material as a lack of belonging, when the real problem is a breakdown in communication that alternative presentation styles would solve. Transparent approaches acknowledge this problem, helping students embrace their college status by including them fully in the conversation.

To date, the bulk of pedagogical research into transparency has focused on assignment design (Winkelmes et al., 2016; Winkelmes et al., 2019); and a pedestrian google search shows many universities supporting a culture of transparent syllabus design as well. Considerably less research has focused on transparent testing. This may stem from the "panoptic" quality of tests, which makes them tactically distinct from assignments. For a variety of reasons—promoting practice, encouraging broad preparation, and preventing answer sharing—most in-class tests (and especially multiple-choice tests, which we examine below) are not entirely transparent, because most instructors do not distribute their test questions in advance. Teachers may put great effort into preparing effective study guides—and study guides may reference everything that *might* be on the test—but certain test specifics must remain hidden until taken. Thus, on its face, it is not surprising that tests would not be first on the minds of transparency advocates.

But transparency is no less relevant for tests. Whether one tests in short-answer, long essay, or multiple-choice formats, transparent communication (or a lack thereof) matters in equal measure, and in at least two respects. First, studies have shown that while college students frequently adopt suboptimal study strategies (Anthenien et al., 2018; Hartwig & Dunlosky, 2012), certain strategies like self-testing have been positively associated with long-term retention (Roediger & Butler, 2011) and GPA (Hartwig & Dunlosky, 2012). But this strategy, in turn, is adopted more consistently by learners with "growth" rather than "fixed" mindsets (Yan et al., 2014). For students to study more (and better), they must believe that doing so will actually help them. This suggests that more transparent tests, facilitated by more transparent study guides—i.e. study guides that offer students specific strategies for studying, specific content for self-testing, and clear explanation of *how* these strategies will promote learning growth and benefit their grade—may be especially effective at steering students towards more

frequent and effective studying. We do not pursue this here.

Second, test questions themselves can be more or less transparent in terms of what they ask students to do or to recall. Some questions are vague, some questions are clear. Some prompts are open and general, others are targeted and specific. Some grading criteria are explicit, others are flexible or even implied. Depending on the context, any of these approaches may be appropriate—but the tendencies towards greater or lesser transparency in each are easily identifiable, and all else equal, we would expect that more transparent test questions (clear; specific; explicit) will benefit more students more of the time.

In my own experience giving multiple-choice tests, a common refrain in post-exam debriefings is that certain questions "tricked" or "psyched out" a student, or led them to "second guess" and mistakenly "change their answer." In a recent learning community on multiple-choice tests at my university, other instructors noted the same. And while on many occasions such criticisms may seem unwarranted or even absurd, it is a mistake to dismiss them out of hand. For upon inspection, it may be the case that a student's difficulty stemmed less from a lack of requisite knowledge to answer the question, than from a lack of understanding of, or comfort with, the question itself, i.e. a lack of transparency. As we discuss below, a range of arbitrary barriers or distractors can prevent students from successfully relaying what they know. And thus, if test questions *could* be reworded to do things like clarify their purpose, eliminate ambiguity, and better distinguish between answers, one would expect this to eliminate those barriers, and to help students who do have the tested knowledge to achieve scores that reflect it.

Our study begins from this basic intuition, grounded in the philosophy of TILT, that in writing tests, no less than in writing assignments, it is incumbent on instructors to emphasize transparency—to write questions that are clear, complete, straightforward, and unambiguous. We theorize that more transparent questions will improve student performance, and will in the process lead to fairer and more accurate assessment of each student's knowledge. In this study, we are particularly interested in the transparency of multiple-choice questions, although the general principle may be applied to others, including short-answer and long essay.

Transparency and Multiple-Choice Tests

The literature on effective multiple-choice question writing is well developed (Haladyna & Downing, 1989a; Haladyna & Downing, 1989b; Burton et al., 1991; Marsh & Cantor, 2014) and might be divided into three broad schools. The first, or "critical thinking" school, responds to a stereotypical conception of multiple-choice questions as "only good for measuring simple recall of facts" (Burton et al., 1991, p. 8), and ambitiously considers whether it is "possible to construct multiple-choice questions that simultaneously test both factual knowledge and critical thinking" (Karras, 1978, p. 211; Karras, 1993; Morrison & Free, 2001; Scott, 1993). For this school, the "real value of multiple-choice items…is their applicability in measuring higher-level objectives, such as those based in comprehension, application, and analysis" (Burton et al., 1991, p. 8).

The second school, which one might call the "test effect" school, examines the impact (short and long-term) of answering multiple-choice test items on student learning and retention (Butler et al., 2006; Butler &

Roediger, 2008; Cantor et al., 2015; Fazio et al., 2010; Little & Bjork, 2012, 2015; Marsh & Cantor, 2014; Marsh et al., 2007; Roediger & Marsh, 2005). Most of these studies are conducted in lab settings; and among other things, they consistently advise against exposing students to incorrect alternatives, and they advocate strongly for combining multiple-choice tests with immediate feedback to test-takers.

Third and finally is the "effective measurement" school, which examines the extent to which multiple-choice tests fairly, accurately, and reliably assess student learning, and how to improve on that score. Burton et al. (1991, p. 1) summarize the concern: "Well-written multiple-choice test questions do not confuse students, and yield scores that are more appropriate to use in determining the extent to which students have achieved educational objectives." These studies consider whether particular multiple-choice strategies—like avoiding negative stems, omitting "all of the above" and "none of the above" options, using question vs. completion form, keeping option lengths similar, and using "plausible" distractors—allow students to more reliably convert their knowledge into correct answers, and allow instructors to more accurately gauge student learning. In two groundbreaking articles, Haladyna and Downing (1989a, 1989b) analyzed 96 theoretical and empirical studies to determine if support existed for no less than 43 multiple-choice item-writing rules in the literature. Each rule was judged on the basis of whether its use in practice affected "(a) item difficulty, (b) item discrimination, (c) test reliability, and (d) test validity" (Haladyna & Downing, 1989b, p. 52). Several rules found strong support.

Many of the rules or strategies identified by Haladyna and Downing (1989a, 1989b) and others (Burton et al., 1991) overlap considerably, if implicitly, with the concept of transparency. And in what follows, we explore the implications of a transparency-centered test-writing approach by focusing on three strategies in particular, each with some support in the literature, that bear on transparency in different ways.

The first is to *keep vocabulary appropriate* (Haladyna & Downing, 1989b, p. 70). Relatively little research exists in this area, but it rests on the idea that the only challenging vocabulary in a test question should be *discipline specific*, and other vocabulary which might hinder a student's understanding of the question should be avoided. In a study of some novelty, Cassels and Johnstone (1984, p. 613) found that on college chemistry exams, empirical results "indicated that the substitution of words (in key positions) by simpler words brought about an improvement in performance." One example was the substitution of "choking" for "pungent." More research is needed to understand how arbitrary vocabulary hurdles impact test performance, but we would expect this to be an especially important concern amongst first-generation students, students entering college with other academic challenges, and students attending regional universities with historically lower average SAT/ACT scores.

Our second strategy is to *avoid the alternatives "all of the above" (AOTA) and "none of the above" (NOTA)* (Burton et al., 1991, pp. 26-27), an area with substantial research coverage. While a "lack of consensus" has existed regarding the impact of using AOTA as an alternative (Haladyna & Downing, 1989b, p. 69), the evidence against using NOTA as an alternative appears robust (although compare with Frary, 1991). As Haladyna & Downing (1989b, p. 61) report, the NOTA option has consistently been

shown to make questions "more difficult" and "less discriminating," while making test scores "less reliable" (Haladyna & Downing, 1989b, p. 61). Recent studies add that using NOTA as a *correct* answer is especially problematic (DiBattista et al., 2014), not only because it erroneously rewards students with knowledge deficiencies (thus effectively reducing item discrimination; see Gross, 1994), but also because it weakens the positive testing effects that students might otherwise glean from the test itself (Odegard & Koen, 2007), even in contexts where immediate feedback is given (Blendermann et al., 2020).

Our concern is distinct from this. From a transparency standpoint, one problem with both AOTA and NOTA options is their potential to encourage second-guessing and internal mind-games. We speculate that students who *have* the knowledge to correctly answer a question may be prompted by AOTA or NOTA options (or both!) to question whether their knowledge is complete, or whether the question is asking for something above and beyond what they know or recall. Eliminating such options, we believe, and having students choose from only concrete alternatives, should eliminate this problem and reduce second-guessing. And we would not be surprised if eliminating AOTA and NOTA as options was especially helpful to first-generation students, for whom lack of confidence is a documented problem. Eliminating such options should allow students to more confidently apply their own recall, rather than question whether that recall is sufficient. At the same time, because the existing research regarding NOTA questions (and their negative learning effects) appears robust, our study will lean heavily towards exploring the unresolved question of AOTA, and to a lesser extent the combination of AOTA and NOTA, as response options.

Our third and final strategy, the least examined in the literature, is to intentionally *cue marginal knowledge*. Regarding human memory, it is widely recognized that "the amount recalled is often an underestimate of what is stored in memory," and that "estimates of remembering will change depending on the particular retrieval cues available" (Cantor et al., 2015, p. 193). Scholars have used "marginal knowledge" to describe knowledge that people have but are unable to spontaneously access, absent a retrieval cue. And while studies have examined the impact of multiple-choice *answers* in "activating" prior knowledge during tests (Schimmelfing & Persky, 2020), and of multiple-choice tests in "stabilizing access" to marginal knowledge after the test (Cantor et al., 2015), the consideration of test *questions* as critical purveyors of recall cues is far less explored.

But this is counterintuitive, for a multiple-choice test question (or any test question) *is* a recall cue. A multiple-choice test question does more than simply request an answer—it also reflects a judgment on the instructor's part about which cues are necessary and sufficient to elicit the desired answer, without also telegraphing what it is. The art of writing questions that provide enough information to elicit correct responses from those with at least marginal knowledge, but not so much information as to reward those with insufficient knowledge—can be delicate. But not all "cues" are "hints," and strategically offering the former, while carefully avoiding the latter, can make both the question itself, and a student's knowledge of their own knowledge of the answer, more transparent.

Purpose of the Study

Our study seeks to better understand how making multiple-choice questions more transparent in these ways would benefit students at a small regional university with a large number of first-generation students. Recognizing that test transparency may come in several forms, we also seek to better understand which kinds of transparent strategies most benefit students, beginning with the three above-mentioned, namely—

1. Keeping vocabulary appropriate—thereby removing vocabulary barriers;

2. Avoiding "all-of-the-above" and "none-of-the-above" options—thereby removing psychological barriers; and

3. Cuing marginal knowledge—thereby aiding recall, by adding additional cues or context;

Methods

Participants

Our study had ninety-two participants, all of whom were students enrolled at Georgia Southwestern State University (GSW) in Americus, Georgia. GSW is a small regional university with total fall 2021 enrollment (undergraduate and graduate) of 3,158 students (BOR, 2021, p. 1). Among all undergraduates who were awarded bachelor's degrees in FY21, 54% had received the Pell grant, and 51% were first-generation students (GSW, 2021).

Participants were undergraduate students enrolled in the course "POLS 1101: American Government" taught by Dr. John LeJeune during the fall 2021 and spring 2022 semesters. American Government is a required core course for GSW undergraduates, and the study included two sections each from each semester. All classes were lecture-based and delivered face-to-face. The only substantive variation in teaching methods occurred in Spring 2022, when during the final four weeks the course slides were significantly re-structured. This could only potentially impact, if at all, results in questions 31-35 below. Of the 92 participants, 57 (62%) were female and 35 (38%) were male. These numbers are consistent with overall enrollment trends at GSW. The class breakdown among participants was as follows: freshman, 48 (52%); sophomore, 29 (32%); junior, 10 (11%); and senior, 5 (5%).

Prior to beginning the study, IRB exemption was sought and granted by the GSW IRB Committee. All participants submitted a signed voluntary consent form at the time of participation.

Research Design and Data Collection

Our study used a "between-subjects" design to compare student performance on "transparent" ("TILTed") versus "non-transparent" ("unTILTed") versions of thirty-five multiple-choice test questions in American Government. The questions were delivered as an optional, in-class extra-credit final exam, and covered material from throughout the course. The original exam included 40 questions—five each from all eight course units. However, the five questions from the eighth unit were omitted from the study because several students did not view the back page to answer them.

Participants received at random either Version 1 or Version 2 of the study test, which

in turn contained different versions—either "transparent" or "non-transparent" versions—of each of the 35 questions. For "Version 1" of the exam, Questions 1-20 were TILTed, and questions 21-35 were unTILTed. For "Version 2" of the exam, questions 1-20 were unTILTed, and questions 21-35 were TILTed. Forty-seven participants received Version 1, and 45 participants received Version 2.

Notably, each question was "transparent" or "non-transparent" in one of three ways. Corresponding to our three modes of transparency discussed above, eight questions varied based on the insertion or deletion of relatively difficult vocabulary; 15 questions varied by including or not including AOTA or AOTA/NOTA answer options; and 12 questions varied based on whether or not additional cues and context were added to help students recall relevant marginal knowledge.

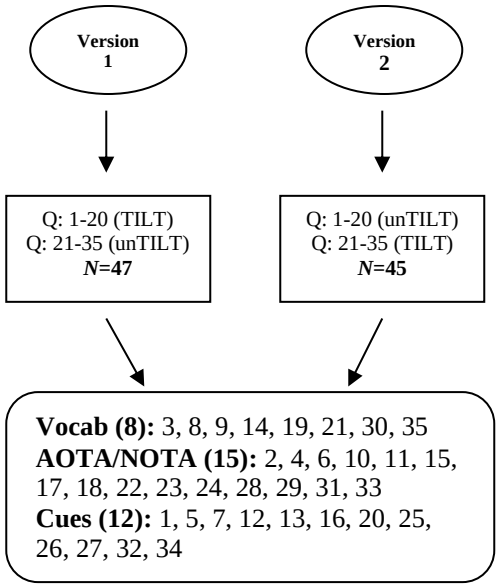

Figure 1: Multiple-Choice Exam Format

Data from these tests were collated to allow statistical comparison of the average scores of TILTed vs. unTILTed versions of each question. For each individual question, we ran a two-sample t-test (assuming unequal variances) to determine (a) whether students scored better on the TILTed versus unTILTed versions of questions at a regular and statistically significant rate; and (b) if this were true for some questions, but not others, whether one could glean useful information about what kinds of transparent strategies were most impactful for our students.

Results

Table 1 shows the statistical results. The first column identifies the relevant question identifier, and the second column identifies the type of transparent change made to the question (A = AOTA/NOTA; C = cues/context; V = vocabulary). The succeeding columns compare the mean and variance of student scores on the TILTed vs. unTILTed versions of each question (1 = correct, 0 = incorrect), with *p* values to indicate statistical significance of the means differences.

A $p \leq .05$ value is generally recognized as the standard for statistical significance. However, we agree with a recent statement by the American Statistical Association that "Practices that reduce data analysis or scientific inference to mechanical 'bright-line' rules (such as '$p < 0.05$') for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making," and that while "Pragmatic considerations often require binary, 'yes-no' decisions…this does not mean that *p*-values alone can ensure that a decision is correct or incorrect" (Wasserstein & Lazar, 2016, p. 131). In our judgment, the question at hand concerns not only whether our results achieve bright-line thresholds that are, perhaps, more appropriate to fields like experimental psychology,

biomedical research, or other medical or hard sciences (see Benjamin et al., 2018); but also, what level of statistical significance would justify recommending, adopting, or even rejecting (if results are negative) the kind of intervention we are proposing. Given the minimal risks that we perceive with our TILT intervention, and the potential it has for helping students immediately, for pedagogical purposes we are comfortable making practical recommendations at a level of $p \leq .10$, although only results at $p \leq .05$ can be said to be statistically significant.

Thus, in Table 1 where $p \leq .10$, the number is marked with a single asterisk to indicate practical significance; where $p \leq .05$, the traditional line of statistical significance, there is a second asterisk. All values in Table 1 were rounded to the nearest .01.

Table 1
Comparing TILTed vs. unTILTed Questions

Ques.	Type	TILTed		unTILTed		
		M	*Var*	*M*	*Var*	*p*
1	C	0.60	0.25	0.49	0.26	.15
2	A	0.45	0.25	0.33	0.23	.13
3	V	0.62	0.24	0.40	0.25	**.02****
4	A	0.60	0.25	0.44	0.25	**.07***
5	C	0.89	0.10	0.58	0.25	**.00****
6	A	0.79	0.17	0.67	0.23	**.10***
7	C	0.32	0.22	0.4	0.25	.21
8	V	0.21	0.17	0.13	0.12	.16
9	V	0.87	0.11	0.84	0.13	.35
10	A	0.66	0.23	0.33	0.23	**.00****
11	A	0.60	0.25	0.53	0.25	.28
12	C	0.70	0.21	0.62	0.24	.21
13	C	0.91	0.08	0.80	0.16	**.06***
14	V	0.68	0.22	0.60	0.25	.21
15	A	0.70	0.21	0.58	0.25	.11
16	C	0.64	0.24	0.56	0.25	.21
17	A	0.66	0.23	0.42	0.25	**.01****
18	A	0.87	0.11	0.73	0.20	**.05****
19	V	0.77	0.18	0.62	0.24	**.07***
20	C	0.60	0.25	0.60	0.25	.48
21	V	0.98	0.02	0.45	0.25	**.00****
22	A	0.76	0.19	0.77	0.18	.45
23	A	0.71	0.21	0.66	0.23	.30
24	A	0.56	0.25	0.43	0.25	.11
25	C	0.51	0.26	0.43	0.25	.21
26	C	0.64	0.23	0.77	0.18	**.10***
27	C	0.42	0.25	0.32	0.22	.16
28	A	0.71	0.21	0.47	0.25	**.01****
29	A	0.67	0.23	0.68	0.22	.44
30	V	0.64	0.23	0.57	0.25	.25
31	A	0.71	0.21	0.77	0.18	.28
32	C	0.82	0.15	0.74	0.19	.19
33	A	0.69	0.22	0.66	0.23	.38
34	C	0.42	0.25	0.45	0.25	.41
35	V	1.00	0.00	0.85	0.13	**.00****

Provisionally, we observe the following: For 28 out of 35 questions, students performed better on the TILTed vs. unTILTed versions of the question. Of these 28, 12 showed differences in performance at a $p \leq .10$ level, with eight of these reaching a $p \leq .05$ level. Among the 12 questions for which TILTing vs. unTILTing appeared to have a positive practical impact, four involved "vocabulary" (V), six were "AOTA/NOTA" (A), and two involved additional "cues or context" (C). Thus, 50% of vocabulary questions (4 of 8), 40% of AOTA/NOTA questions (6 of 15), and 17% of cues/context questions (2 of 12) showed a meaningfully positive impact when using TILTed vs. unTILTed strategies.

Vocabulary

Questions 3, 19, 21, and 35 were the four vocabulary questions to show meaningful positive results at *p* values of .02, .07, .00, and .00, respectively. Each of these questions is provided in the Appendix below (absent answer choices).

AOTA/NOTA

Out of the 15 AOTA/NOTA questions, six were found to have statistically meaningful positive mean differences between their TILTed vs. unTILTed forms, with four reaching statistical significance. Questions 4, 6, 10, 17, 18, and 28 saw *p*-values of .07, .10, .00, .01, .05, and .01, respectively.

Cues and Context

The addition of cues and contexts did not have as much of a positive impact on student performance as expected. In some cases, the impact was negative. Question 5 showed a large improvement in student performance from TILTing the question (M(TILTed) = .89 vs. M(unTILTed) = .58, $p < .05$). Meanwhile Question 26 showed the reverse effect, with students scoring significantly better on the question's unTILTed version (M(TILTed) = .64 vs. M(unTILTed) = .77, $p \sim .10$).

Discussion

What does one make of these numbers? On first glance, it appears that both the vocabulary and AOTA/NOTA interventions had a meaningful impact on students' ability (or inability) to correctly answer test questions. In many instances, replacing unnecessarily difficult, non-course-related vocabulary, as well as eliminating AOTA/NOTA options, seemed to better enable students to translate their knowledge into correct answers.

Unfortunately, the same cannot be said—at least not with as much confidence—for adding additional cues or context to prompt marginal knowledge. Not only were there fewer positive results in this category, but among the seven questions for which students scored at least as well if not better on the unTILTed version than the TILTed version (contrary to our expectations), four were based on added cues and context (the other three were AOTA/NOTA), and one of these (Question 26) was even significant at $p \sim .10$ (without rounding, $p = .103102$), suggesting that the additional cues in that question hindered students' ability to answer.

Vocabulary

Perhaps the most important conclusion to be drawn from this study is the importance of calibrating non-course-related vocabulary on multiple-choice tests so as to avoid creating obstacles for students who might otherwise know the answer to the question. Based on the inclusion or non-inclusion of the words "deleterious," "equalized," "preceded," and "subordinate," the average difference in correct answers on Questions 3, 19, 21, and 35 was 22%, 15%, 53%, and 15%, respectively. Each of these reached or approached statistical significance.

On the other hand, "vocabulary" words that did not make a meaningful difference, in Questions 8, 9, 14, and 30, were "peculiar," "exacerbated," "suffrage" (as opposed to "voting rights"), and "oversees," respectively.

In sum, instructors writing multiple-choice tests would do well to scan their tests for potentially problematic vocabulary terms. This is a quick and effective way to eliminate an unnecessary barrier to answering correctly.

AOTA/NOTA

To interpret the AOTA/NOTA results, it is important to highlight that these questions in their unTILTed form were presented in three different ways. There were 15 AOTA/NOTA questions in total: in 10 of these, the unTILTed form included one AOTA response, plus four concrete answers (AOTA-only); in four of these, the unTILTed form included separate AOTA and NOTA responses, plus three concrete answers (AOTA+NOTA); and one included an experimental "All or None" response plus four concrete answers.

Of the six AOTA/NOTA questions with statistically meaningful results, five of these were AOTA-only when unTILTed (Questions 4, 6, 10, 17, and 28), and one was AOTA+NOTA (Question 18). This means that 50% (5 of 10) of AOTA-only questions produced a statistically meaningful difference, while 25% (1 of 4) of AOTA+NOTA questions did the same.

Thus, while previous studies may be "inconclusive" regarding the dangers of using AOTA (as distinct from NOTA), particularly as an incorrect distractor, our study suggests that AOTA-only is indeed a significant problem for test-takers. At the same time, it is important to note that our study was limited to the use of AOTA (with or without NOTA) as a distractor, and not as a correct answer.

Cues and Context

Adding cues to prompt marginal knowledge was the least impactful method of making questions more transparent, although Question 5 saw a statistically significant positive effect, and Question 26 saw a statistically meaningful negative effect. As illustrative examples of how adding additional cues and context worked both positively and negatively, the Appendix includes Questions 5 and 26.

Question 5 shows how adding even minimal context can help students access marginal knowledge. The question asks students to identify the "First Governing document" of the United States, seeking the Articles of Confederation as an answer. The TILTed version adds the cue that this was a "failed" attempt to govern the thirteen states, and this cue appears to have made a significant difference in students' ability to answer the question. I do not believe that adding this cue undermined the rigor of the question—indeed, it did much to clarify it.

Question 26 asks about the transition of the federal bureaucracy from the old "spoils system" to the current "merit-based system," hoping students will identify the former as the precursor. The TILTed version of the question adds additional information that this transition happened "after a President was assassinated," and this information—designed to remind students of a particular classroom discussion at the point of the transition—appears to have distracted them in some way from the correct answer. Perhaps because two of the incorrect lures were the "Roosevelt system" and the "Jefferson system" (and half of the incorrect answers among this TILTed group were one of these two lures), I suspect that insertion of the word "president" in the question stem mistakenly drew them to those alternatives.

All of this suggests that adding information to question stems to cue prior or marginal knowledge can be a delicate process, with potential advantages and disadvantages. Literature suggests that,

"Excess material in the stem that is not essential to answering the problem increases the reading burden and adds to student confusion over what he or she is being asked to do" (Burton et al., 1991, p. 17)—and this would seem borne out by Question 26. Every word in a question stem will be read by students as a potential cue, and so it is important that every cue lead in the right direction, and not to incorrect distractors.

Conclusions

This study was an attempt to insert transparency into the mindset and strategy of multiple-choice test writing. Tests are often viewed as only an opportunity for students to communicate their knowledge to the professor, after the professor has communicated to them in lecture. But in a test, no less than in a lecture, it is the instructor who initiates the communication, and students can only respond to the questions and communication given to them.

As the results of this study show, even the smallest changes in that communication—a single word that a student can or cannot understand, a single cue that steers them in one direction rather than another—can drastically impact student performance on exams (for better or worse), notwithstanding their level of preparation. As instructors, this demands an extra amount of care on our part to ensure that we are communicating all that we hope to communicate, no more than what we want to communicate, and that we are communicating all of this to everyone in our classes. It also requires consciously designing questions to enable students to relay their knowledge to us without arbitrary barriers or distractions. At their best, tests are vessels for clear two-way communication.

In the narrow field of multiple-choice testing, there is of course no single-path solution for all classrooms and contexts. Nor will all transparent interventions succeed. But adopting a transparent philosophy at least attunes us to the small, daily steps we can take—perhaps changing one word on one question—that will make a difference to some.

References

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education: An International Journal*, *21*(2), 303-322. https://doi.org/10.1007/s11218-017-9412-z

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., …Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behavior*, *2*, 6-10. https://doi.org/10.1038/s41562-017-0189-z

Blendermann, M. F., Little, J. L., & Gray, K. M. (2020). How "none of the above" (NOTA) affects the accessibility of tested and related information in multiple-choice questions. *Memory*, *28*(4), 473-480. https://doi.org/10.1080/09658211.2020.1733614

Board of Regents, University System of Georgia (BOR). (2021). *Semester Enrollment Report, Fall 2021*. Atlanta, GA: Office of Research and Policy Analysis.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty* [Booklet]. Provo, UT: Brigham Young University Testing Services and The Department of Instructional Service.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941–956. https://doi.org/10.1002/acp.1239

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616. https://doi.org/10.3758/MC.36.3.604

Cantor, A. D., Eslick A. N., Marsh E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition*, *43*(2), 193-205. 10.3758/s13421-014-0462-6

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, *61*(7), 613-615. 10.1021/ed061p613

DiBattista, D., Sinnige-Egger, J.-A., & Fortuna, G. (2014). The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, *82*(2), 168-183. https://doi.org/10.1080/00220973.2013.795127

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*(4), 407-418. doi: 10.3758/MC.38.4.407

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, *4*(2), 115-124. https://doi.org/10.1207/s15324818ame0402_2

Georgia Southwestern State University (GSW). (2021). *Complete College Georgia – Campus Plan Update 2021*. https://completega.org/georgia-southwestern-state-university

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of "none of the above." *Evaluation and the Health Professions*, *17*(1), 123-126.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-

choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51-78. https://doi.org/10.1207/s15324818ame0201_4

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126-134. https://doi.org/10.3758/s13423-011-0181-y

Karras, R. W. (1978). Writing multiple-choice questions: The problem and a proposed solution. *The History Teacher*, *11*(2), 211-218. https://doi.org/10.2307/492246

Karras, R. W. (1993). A multidimensional multiple-choice testing system. In R. Blackey (Ed.), *History anew: Innovations in the teaching of history today* (1st ed., pp. 73-81). Long Beach, CA: The University Press, California State University, Long Beach.

Little, J. L., & Bjork, E. L. (2012). The persisting benefits of using multiple-choice tests as learning events. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 683-688. https://escholarship.org/uc/item/50d9x93k

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*(1), 14-26. doi: 10.3758/s13421-014-0452-8

Ma, P.-W. W. & Shea, M. (2021). First-generation college students' perceived barriers and career outcome expectations: Exploring contextual and cognitive factors. *Journal of Career Development*, *48*(2), 91-104. https://doi.org/10.1177/0894845319827650

Marsh, E. J., & Cantor, A. D. (2014). Learning from the test: Dos and don'ts for using multiple-choice tests. In M. A. McDaniel, R. F. Frey, S. M. Fitzpatrick, & H. L. Roediger (Eds.), *Integrating cognitive science with innovative teaching in STEM disciplines* (1st ed., pp. 37-52). St. Louis, MO: Washington University Libraries. http://dx.doi.org/10.7936/K7Z60KZK

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194-199. https://doi.org/10.3758/BF03194051

Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, *40*(1), 17-24. doi: 10.3928/0148-4834-20010101-06

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, *15*(8), 873-885. doi: 10.1080/09658210701746621

Roediger, H. L. III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L. III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory , and Cognition*, *31*(5), 1155-1159. doi: 10.1037/0278-7393.31.5.1155

Schimmelfing, L. C., & Persky, A. M. (2020). Activating prior knowledge using multiple-choice question distractors. *Medical Education*, *54*(10), 925-931. https://doi.org/10.1111/medu.14162

Scott, A. M. (1993). Life is a multiple-choice question. In R. Blackey (Ed.), *History anew: Innovations in the teaching of history today* (1st ed., pp. 59-72). Long Beach, CA: The University Press, California State University, Long Beach.

Stebleton, M. J., & Soria, K. M. (2012). Breaking down barriers: Academic obstacles of first-generation students at research universities. *Learning Assistance Review*, *17*(2), 7-20. https://hdl.handle.net/11299/150031

Unverferth, A. R., Talbert-Johnson, C., & Bogard, T. (2012). Perceived barriers for first-generation students: Reforms to level the terrain. *International Journal of Educational Reform, 21*(4), 238-252. https://doi.org/10.1177/105678791202100402

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Winkelmes, M.-A. (2019a). Introduction: The story of TILT and its emerging uses in higher education. In M.-A. Winkelmes, A. Boye, & S. Tapp (Eds.), *Transparent design in higher education teaching and leadership: A guide to implementing the transparency framework institution-wide to improve learning and retention* (1st ed., pp. 1-14). Stylus Publishing.

Winkelmes, M.-A. (2019b). How to use the transparency framework. In M.-A. Winkelmes, A. Boye, & S. Tapp (Eds.), *Transparent design in higher education teaching and leadership: A guide to implementing the transparency framework institution-wide to improve learning and retention* (1st ed., pp. 36-54). Stylus Publishing.

Winkelmes, M.-A., Bernacki, M., Butler, J., Zochowski, M., Golanics, J., & Weavil, K. H. (2016). A teaching intervention that increases underserved college students' success. *Peer Review*, *18*(1/2), 31-36. https://www.proquest.com/docview/1805184428

Winkelmes, M.-A., Boye, A., & Tapp, S. (Eds.). (2019). *Transparent design in higher education teaching and leadership: A guide to implementing the transparency framework*

institution-wide to improve learning and retention. Stylus Publishing.

Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, *3*(3), 140-152. https://doi.org/10.1016/j.jarmac.2014.04.003

JOHN LEJEUNE is an Associate Professor of Political Science at Georgia Southwestern State University. His primary research interests include political theory and the philosophy of education.

Appendix: Sample Test Questions

Note. In all questions below, version a) is TILTed, while version b) is unTILTed. Except for Question 26, answer choices are omitted.

Question 3 (Vocabulary)

a) As the new Constitution was being debated, which of the following would have had the most negative impact on the political power of small states?

b) As the new Constitution was being debated, which of the following would have had the most deleterious impact on the political power of small states?

Question 19 (Vocabulary)

a) If a Senate vote is tied, then who or what plays a deciding role?

b) If a Senate vote is equalized, then who or what plays a deciding role?

Question 21 (Vocabulary)

a) Which president served right before Donald Trump?

b) Which president preceded Donald Trump?

Question 35 (Vocabulary)

a) True (a) or False (b): The Supreme Court is below the circuit courts.

b) True (a) or False (b): The Supreme Court is subordinate to circuit courts.

Question 5 (Cues and Context)

a) The First Governing document in the United States—i.e. the first, but failed, attempt to establish rules to govern the relationship between the thirteen states—was:

b) The First Governing document in the United States was:

Question 26 (Cues and Context)

a) Which form of bureaucratic organization, which famously preceded the merit-based system used today, and which ended after a President was assassinated, was based largely on rewarding political supporters?

b) Which form of bureaucratic organization was based on rewarding political supporters?

Answer Options: spoils system, grant system, Roosevelt system, corporate system, Jefferson system