# LOGICAL MISCONCEPTIONS ON TRUE-FALSE ASSESSMENTS IN EARLY UNDERGRADUATE MATHEMATICS

SUSAN J. DURST AND SCOTT R. KASCHNER

Abstract. We explore student performance on True-False assessments with statements in the conditional form "If P then Q" in order to see whether logical misconceptions impede students' ability to demonstrate mathematical knowledge in the context of this kind of assessment. We administered an online assessment to a population of Calculus II students. Half of these students were given a standard True-False assessment, and half were given an experimental assessment with three choices. We find that students do make logical errors on True-False items of a certain logical form, and that these errors are unrelated to their calculus knowledge. However, this effect is eliminated by the experimental three-choice assessment. We believe that this provides an easy and practical alternative to True-False items in the problematic logical form for early undergraduate mathematics students.

## 1. Introduction

The use of True-False (T/F) questions to assess student understanding and knowledge has been studied extensively, and these studies have a rich and interesting history. While there is little consensus as to the best practice for using this form of assessment [3], there are many studies that suggest that well-crafted and well-administered T/F assessments can provide a valid measure of student understanding [7].

Problems associated with T/F assessments are well-established. Acquiescence, the tendency of examinees to answer T/F items as true, emerged in the literature as an observable phenomenon nearly a century ago [9], and larger studies have since verified the existence of this issue [5]. Another common problem that arises in T/F assessment is the effect of guessing, and this can be significant [4]. However, the effects of pure, arbitrary guessing are reduced in longer T/F assessments [7], and partial knowledge can greatly increase the quality of a students' guesses so that their performance correlates relatively well with their level of understanding of the material [10, 2]. In fact, there are many techniques for mitigating the problems associated to T/F assessment. In [1], it was demonstrated that indeed T/F assessment reliability can be improved by including more items on the assessment, and when guessing is be a serious issue, it can help to increase the number of possible responses—essentially changing to a multiple choice (MC) format. Alternate scoring formulas can also mitigate the effects of guessing [13], but they are often complicated to use in practice.

There are also many problems with T/F assessments that can be attributed to badly written items [3, 11, 6]. However it has been noted by many researchers, such as Robert Ebel, that while it is not easy, it is possible to write good T/F items. That said, even when great care is taken in writing T/F items, ambiguity

can remain an issue [7]. In crafting T/F items that attempt to assess less trivial content, the test-writer is more likely to introduce ambiguity into the item. As Ebel put it, should a test-writer "steer away from the Scylla of triviality, he is apt to be caught in the Charybdis of ambiguity." This ambiguity can also arise when considering the very nature of truth itself; Ebel notes [7] that "If the statement is not perfect in truth, if it has the slightest flaw, should it be considered false? Probably not." Outside the context of STEM (science, technology, engineering, and mathematics) fields, this is likely the correct conclusion. However, in a mathematics course a T/F assessment is often designed to test students' ability to distinguish between "perfect truth" and a statement with "the slightest flaw." The authors have found that for many young undergraduate students, this can be a stumbling block in displaying their content knowledge in a T/F assessment.

The purpose of this study is not to make general conclusions about T/F assessment. Rather, we focus on a specific type of ambiguity that arises commonly in undergraduate mathematics T/F assessment items. We identify a specific logical structure for T/F items and present evidence that this commonly-used logical structure for T/F assessment is ineffective at assessing student understanding in early undergraduate mathematics courses.

For our study we will focus on conditional statements—statements of the form "If P then Q." This is a common logical structure that appears in T/F items. A conditional statement of this form is deemed true if the consequent Q is true whenever the conditional P is true. However, there is a subtlety in this convention: if when P is true it is possible for Q to be either true or false, the entire conditional statement "If P then Q," is deemed false. For example, the conditional statement

$$\text{If the square of } x \text{ is } 4, \text{ then } x \text{ is } 2.$$

is false. Even assuming that the conditional "$x$ is 4" is true, the consequent "$x$ is 2" could be either true or false, since $x$ could also be $-2$.

While this convention is necessary for consistent logic, it can be a confounding factor in the assessment of mathematical content knowledge. In mathematics courses where definitions and mathematical properties have very fine subtleties, any layer of logical complication has the potential to cause students to make errors they otherwise may avoid. In this study, we quantify the prevalence of this phenomenon in T/F assessments for undergraduate students in Calculus II courses.

We also present an alternative to standard T/F assessment for conditional statements, and evidence that the alternative is more effective at assessing student understanding than standard T/F questions. Multiple choice (MC) is a common alternative to T/F assessment and has been shown to be more reliable than T/F [12, 1]. However, it has also been shown that MC is less reliable when the choices presented are independent T/F statements [8]. In alignment with these studies, we provide evidence that allowing students to assume a true conditional, and to declare the consequent to be "always true," "never true," or "sometimes true, sometimes false" mitigates the aforementioned logical ambiguity faced by undergraduate mathematics students.

Specifically, the purpose of this study is to answer the following research questions:

(1) Is the logical convention for assigning truth values to conditional statements confounding T/F assessment in early undergraduate mathematics?

(2) Is it more confounding for conditional statements of a particular logical form?
(3) Does an adjusted MC format improve the assessment for this population of students?

## 2. Methods

2.1. **Population.** To generate a population for this study, we recruited volunteers from Calculus II sections at the University of Arizona during the Spring semester of 2015. During the last week of class, emails were sent to the nineteen Calculus II instructors requesting their assistance in recruiting their students to participate in the study. Participating instructors sent an email to their students requesting their participation, and containing a link to an online assessment. Of the 788 students registered for Calculus II that semester, 107 volunteered to participate in the study, and 70 of these students completed the assessment. We only count the 70 who completed the assessment as participants. Participation in the study was entirely voluntary and had no impact on student grades, and we believe this contributed to the attrition. It also may have encouraged guessing, which we account for in the analysis.

2.2. **Instrument for Data Collection.** A ten-item assessment was developed consisting entirely of conditional statements. The content for the assessment was drawn from single-variable Calculus, the material typically covered in a two-semester sequence. Three different presentations of each item on the assessment were developed, each in the form "If P then Q." In the first presentation, T, the consequent Q is true whenever the conditional P is true. In the second presentation, F, the consequent Q is false whenever P is true. In the third presentation, B, the consequent Q is sometimes true and sometimes false when the conditional P is true. All three forms of each item are listed in Figure 6.

The assessments were constructed on SurveyGizmo, an online survey site. The ten different assessment items were presented to each participant in random order. For each item, the presentation was also randomly chosen, with with approximately 30% of participants seeing presentation T, 30% seeing presentation F, and 40% seeing presentation B. Participants averaged ten minutes to complete the assessment.

Two different forms of the randomized assessment were also implemented on SurveyGizmo. The first was classic T/F assessment, in which participants were given two response options for each item: True or False; we call this form (TF). The second assessment differed from the first only in that the response options given for each question were

- Always true,
- Never true,
- Sometimes true, sometimes false;

we call this form (TFB). The hyperlink sent to students in the recruitment email randomly sent half the participants to form (TF) and half the participants to form (TFB).

**(TF) Form**

| Item Presentation | T | F | B |
|---|---|---|---|
| Number Correct | 75 | 80 | 70 |
| Number Incorrect | 36 | 30 | 59 |
| Percent Correct | 67.57% | 72.73% | 54.26% |
| 95% Confidence Interval | 61.44-74.3% | 66.04-79.18% | 48.13-60.32% |

**(TFB) Form**

| Item Presentation | T | F | B |
|---|---|---|---|
| Number Correct | 53 | 57 | 80 |
| Number Incorrect | 50 | 44 | 66 |
| Percent Correct | 51.46% | 56.44% | 54.79% |
| 95% Confidence Interval | 43.5-59.02% | 48.07-64.33% | 46.35-62.77% |

FIGURE 1. Raw data for both forms of the assessment, separated by item presentation type.

## 3. Results

3.1. **Comparing Item Presentations Within One Assessment Form.** We wish to determine whether we observe a difference between participants' performance on items in presentation T, F, and B, and we test for these differences in both forms of the assessment. In Figure 1, we have recorded the results of both forms of the assessment, separated by item presentation. As our total sample includes only 71 participants, we also calculated a 95% confidence interval for these statistics using a numerical method called *bootstrapping*.

In bootstrapping, we imagine that the actual population consists of infinitely many copies of our sample. Then we essentially rerun our experiment arbitrarily many times by drawing samples of the same size with replacement from our original sample.

We used the programming language R to sample with replacement from our 35 (TF) forms, to produce 1,000 bootstrapped samples of size 35 and calculated the percentage of correct answers in each of the 1,000 samples. Taking the collection of all of these percentages and throwing out the 25 highest and lowest percentages gives us a 95% bootstrap confidence interval for the statistic. We did the same with our 35 (TFB) form assessments.

In the (TF) form of the assessment, participants performed significantly better on items in presentation T and F than they did on items in presentation B. Notice that the confidence intervals for these statistics do not overlap. In the (TFB) form of the assessment, this difference disappears.

In Figure 2, we have also recorded the 95% Bootstrap confidence intervals for differences between $P(\mathrm{T})$, $P(\mathrm{F})$, and $P(\mathrm{B})$, the percentage of correct answers to items in presentation T, F, and B respectively.

The only statistically significant difference that we observe is between items in presentation T and B, and items in presentation F and B on the assessment of type (TF). None of the differences in percentages on the (TFB) assessment achieve statistical significance.

**(TF) Form**

| | Original Data | 95% Bootstrap Confidence Interval |
|---|---|---|
| P(F)-P(T) | 5.16% | -4.27-14.76% |
| P(F)-P(B) | 18.47% | 10.73-26.71% |
| P(T)-P(B) | 13.31% | 5.07-21.72% |

**(TFB) Form**

| | Original Data | 95% Bootstrap Confidence Interval |
|---|---|---|
| P(F)-P(T) | 4.98% | -6.03-16.11% |
| P(F)-P(B) | 1.65% | -9.23-12.36% |
| P(T)-P(B) | -3.33% | -13.31-5.79% |

FIGURE 2. Bootstrapped confidence intervals for both forms of the assessment, separated by item presentation type.

| Item Presentation | Percentage, T/F | 95% CI, T/F | Percentage, T/F/B | 95% CI, T/F/B |
|---|---|---|---|---|
| T | 67.57% | 61.44-74.3% | 51.46% | 43.5-59.02% |
| F | 72.73% | 66.04-79.18% | 56.44% | 48.07-64.33% |
| B | 54.26% | 48.13-60.32% | 54.79% | 46.35-62.77% |

FIGURE 3. Comparison of item presentation type for both assessment types.

3.2. **Comparing Assessment Forms Within One Item Presentation Type.** We would also like to compare participant performance on a particular item presentation across the two forms of the assessment. As we can observe in Figure 3, the percentage of correct answers to problems in presentation T and F drop significantly between the two forms, and the percentage of correct answers to items in presentation B remain more or less constant.

It is likely that this drop is accounted for by the inclusion of three multiple choice options rather than two and the tendency of participants to guess. We expect participants who guess randomly on a (TF)-style assessment to guess correctly about 50% of the time, but participants who guess randomly on a (TFB)-style assessment should only guess correctly about 33.33% of the time. In this situation, the percentage of correct answers is not a good tool for comparing the amount of calculus knowledge demonstrated on the same type of item across different forms of the assessment.

For this comparison, we perform a $Z$-test on our results, comparing them against a null hypothesis of random guessing. The resulting $Z$-score gives us a measure of how unlikely it is that participants with no calculus knowledge would perform this well or better on this assessment, giving us a way of comparing how participants performed on the two different forms of assessment. The results can be seen in Figure 4.

The confidence intervals for these $Z$-scores look fairly similar, except the $Z$-score for problems of type $B$ on the (TF) assessment. The confidence interval for this $Z$-score spans zero, indicating that in our study, our participants' performance on items of this type did not differ from random guessing in a statistically significant way.

| Item Presentation | $Z$-Score, T/F | 95% CI, T/F | $Z$-Score, T/F/B | 95% CI, T/F/B |
|---|---|---|---|---|
| T | 3.7 | 2.36-5.2 | 3.9 | 2.2-5.54 |
| F | 4.77 | 3.32-6.18 | 4.93 | 3.12-6.64 |
| B | .97 | -.43-2.27 | 5.5 | 3.25-7.61 |

FIGURE 4. $Z$-scores and associated confidence intervals for both forms of the assessment, separated by item presentation type.

| Item Presentation | $Z_m$, T/F | $Z_M$, T/F | $Z_m$, T/F/B | $Z_M$, T/F/B |
|---|---|---|---|---|
| T | 2.32-4.93 | 2.41-5.12 | 2.21-5.53 | 2.29-5.74 |
| F | 3.22-5.87 | 3.36-6.12 | 3.14-6.61 | 3.2-6.9 |
| B | -.42-2.35 | -.45-2.49 | 3.14-7.1 | 3.34-7.54 |

FIGURE 5. 95% bootstrap confidence intervals for both forms of the assessment.

However, this is not a perfect tool for comparison—the $Z$-score is sensitive to sample size. The same percentage yielded by a larger sample will result in a higher $Z$-score. To control for changing sample size, we adjust our calculation of the $Z$-score. The $Z$-score is given by the formula

$$Z = \frac{X - np}{\sqrt{np(1-p)}},$$

where $X$ is the number of correct answers, and $p$ is the probability of a participant randomly guessing the correct answer on a single problem ($\frac{1}{2}$ for form (TF), $\frac{1}{3}$ for form (TFB)). We define an adjusted $Z_k$-score by

$$Z_k = \frac{k(X - np)}{n\sqrt{kp(1-p)}}.$$

This calculates the $Z$-score for an experiment in which a population of size $k$ attained the same percentage of correct answers as the actual population. Since this is meant to allow us to compare participant performance across the different forms of the assessment, we will calculate $Z_m$ and $Z_M$, where $m$ is the smaller sample size from our original data, and $M$ is the larger bootstrapped sample size. In Figure 5, we record the 95% bootstrap confidence intervals.

Here we can see that the $Z_m$- and $Z_M$-scores are similar across the two assessments for items in presentation T and F, but increase drastically for items in presentation B. This suggests that, even accounting for sample size, we see an improvement in participant performance on items in presentation B in the (TFB) assessment.

## 4. Conclusion

Students seem to struggle with T/F problems of the form "If P then Q," when neither Q nor its negation are logical consequences of P. We observed this difference clearly in the results of our T/F assessment. The students' performance on the questions of form B was almost indistinguishable from random guessing. This does not appear to be due to a misunderstanding of the underlying calculus concepts. Within the T/F assessment, we saw that students performed better on questions that tested exactly the same calculus concept, but were formulated so that either

P implied Q or P implied not-Q. Also, in our adjusted T/F/B assessment, where students were given the option of saying that Q would be true "always," "sometimes," or "never," student performance on questions of logical type B resembled their performance on questions of logical types T and F, and were far removed from random guessing.

This suggests that the poor performance on questions of type B in the T/F assessment are due at least in part to a lack of knowledge of predicate logic, not a lack of calculus knowledge. It would be interesting to have more qualitative data about our students' thought process in the T/F form of the assessment. Do they base their decision on a single example? If they do, then perhaps the improvement on the T/F/B form of the exam is due to the question drawing their attention to the fact that multiple examples need to be considered. The small amount of extra prompting brings to mind a piece of calculus knowledge that the student would otherwise have overlooked. Or perhaps they recognize that examples of the form "P and Q" and "P and not-Q" both exist. However, they do not recognize that this renders the "If P then Q" statement false, and proceed to guess randomly. In this case, the student would have complete knowledge of the calculus they need in order to solve the problem, and the incorrect answer would be entirely due to a logical error which would not occur in the T/F/B form.

This leaves us with an interesting choice to make as instructors. On the one hand, understanding the principles of logic is important to building a good strong understanding of mathematics. However, we are doing our students a disservice if we expect them to understand logical concepts that they have never been taught. And given the pressure in a calculus class to cover a large amount of material in a short span of time, it is not clear exactly when we would cover these logical ideas. Nonetheless, a calculus class which incorporates logical reasoning into its structure is an attractive idea.

This would need to be handled carefully. In our data we encountered an interesting outlier that is suggestive of what happens when these logical ideas are touched on, but not fully fleshed out. In Question 8, our statement of form B read:

$$\text{If } a_k \geq b_k \text{ for all } k \text{ and } \sum b_k \text{ converges, then } \sum a_k \text{ converges.}$$

This is a common example in Calculus II of a statement in logical form B, which is used to demonstrate that such statements are false. This question was answered correctly every single time, in spite of the fact that every other question that tested this calculus concept received both correct and incorrect responses. This suggests the possibility that rather than teaching students that statements in logical form B are false, we have accidentally been teaching them that *this particular* statement is false. Of course, our sample size is small, so we can only speculate, but it does highlight one of the possible pitfalls of attempting to cover logical ideas too quickly.

Another possibility is to avoid testing statements of the form "If P then Q" in a T/F context. The adjusted T/F/B assessment aligns better with students' intuitive understanding of logical implication, and it appears that students' performance is less random when the question is asked in this form. However, we also observed that it results in lower scores overall. Again, it would be interesting to obtain more qualitative data about our students' thought processes. It could be the case that students who are able to choose correctly between "true" and "false" shy away from making the more extreme-sounding assertion that a certain result will "always" or

"never" occur, and default to an incorrect "sometimes" option. Or it is possible that the drop in scores in the (TFB) form of the assessment that we observe in our data is purely the result of the decreased probability of guessing correctly. In this case, we could make an argument in favor of the T/F/B form on the grounds that we want to encourage our students to know the material, not to make random guesses.

What is clear from this data is that T/F questions of the form "If P then Q," need to be handled carefully by instructors teaching calculus students who have not had any training in logic. They are not a good tool for assessing student understanding of calculus concepts. If questions of this type are to be asked, students must have some sort of logical instruction to help them to prepare for assessments of this type. Otherwise, there are other types of assessment—in particular the T/F/B assessment—that will allow students more opportunity to demonstrate their calculus knowledge without obfuscating factors.

## References


[1] Richard F Burton. Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1):41–50, 2001.

[2] Richard F Burton. Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, 36(9):805–811, 2002.

[3] Richard F Burton. Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1):65–72, 2005.

[4] Richard F Burton and David J Miller. Statistical modelling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, 24(4):399–411, 1999.

[5] Lee J Cronbach. Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, 33(6):401, 1942.

[6] Steven M Downing. Validity: on the meaningful interpretation of assessment data. *Medical education*, 37(9):830–837, 2003.

[7] Robert L Ebel. The case for true-false test items. *The School Review*, pages 373–389, 1970.

[8] Robert L Ebel. The ineffectiveness of multiple true-false test items. *Educational and Psychological Measurement*, 38(1):37–44, 1978.

[9] Martin F Fritz. Guessing in a true-false test. *Journal of Educational Psychology*, 18(8):558, 1927.

[10] WA Gilmour and DE Gray. Guessing on true-false tests. *Educational Research Bulletin*, pages 9–12, 1942.

[11] Gareth Holsgrove and Margaret Elzubeir. Imprecise terms in uk medical multiple-choice questions: what examiners think they mean. *Medical Education*, 32(4):343–350, 1998.

[12] Albert C Oosterhof and Douglas R Glasnapp. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. *The Journal of Experimental Education*, 42(3):62–64, 1974.

[13] Frank Reid. An alternative scoring formula for multiple-choice and true-false tests. *The Journal of Educational Research*, 70(6):335–339, 1977.

| | |
|---|---|
| (T) | If $\lim_{k\to\infty} a_k \neq 0$, then $\sum_{k=1}^{\infty} a_k$ diverges. |
| (F) | If $\lim_{k\to\infty} a_k \neq 0$, then $\sum_{k=1}^{\infty} a_k$ converges. |
| (B) | If $\lim_{k\to\infty} a_k = 0$, then $\sum_{k=1}^{\infty} a_k$ converges. |
| (T) | If $a_{k+1}/a_k = r$ for all $k \geq 0$, then $\sum_{k=0}^{\infty} a_k = \dfrac{a_0}{1-r}$. |
| (F) | If $a_{k+1}/a_k = r$ for all $k \geq 0$, then $\sum_{k=0}^{\infty} a_k = \dfrac{a_0}{1+r}$. |
| (B) | If $a_{k+1}/a_k = r$ for all $k \geq 0$, then $\sum_{k=0}^{\infty} a_k = \dfrac{1}{1-r}$. |
| (T) | If $x$ is an inflection point, then $f''(x) = 0$. |
| (F) | If $x$ is an inflection point, then $f''(x) > 0$. |
| (B) | If $f''(x) = 0$, then $x$ is an inflection point. |
| (T) | If $f(x) > 0$ for any $x \in (0, 5)$, then $\int_0^5 f(x)dx > 0$. |
| (F) | If $\int_0^5 f(x)dx > 0$, then $f(x) \leq 0$ for any $x \in (0, 5)$. |
| (B) | If $\int_0^5 f(x)dx > 0$, then $f(x) > 0$ for any $x \in (0, 5)$. |
| (T) | If $\lim_{n\to\infty} f(x) = \infty$ and $\lim_{n\to\infty} g(x) = \infty$, then $\lim_{n\to\infty} \dfrac{g(x)}{f(x)}$ is an indeterminant form. |
| (F) | If $\lim_{n\to\infty} f(x) = \infty$ and $\lim_{n\to\infty} g(x) = 0$, then $\lim_{n\to\infty} \dfrac{g(x)}{f(x)}$ is an indeterminant form. |
| (B) | If $\lim_{n\to\infty} f(x) = \infty$, then $\lim_{n\to\infty} \dfrac{g(x)}{f(x)}$ is an indeterminant form. |
| (T) | If $f'(x) = 0$ and $f''(x) \neq 0$, then $f(x)$ is either a local minimum or a local maximum. |
| (F) | If $f'(x) \neq 0$, then $f(x)$ is either a local minimum or a local maximum. |
| (B) | If $f'(x) = 0$, then $f(x)$ is either a local minimum or a local maximum. |
| (T) | For any $n$, if $f(x) = x^n$, then $f'(x) = nx^{n-1}$. |
| (F) | For any $n$, if $f(x) = x^n$, then $f'(x) = \dfrac{x^{n+1}}{n+1}$. |
| (B) | If $f(x) = x^n$, then $f'(x) = x^{n-1}$. |
| (T) | If $a_k \geq b_k > 0$ for all $k$ and $\sum b_k$ converges, then $\sum a_k$ converges. |
| (F) | If $a_k \geq b_k$ for all $k$ and $\sum b_k$ diverges, then $\sum a_k$ converges. |
| (B) | If $a_k \geq b_k$ for all $k$ and $\sum b_k$ converges, then $\sum a_k$ converges. |
| (T) | If $\sum_{k=0}^{\infty} c_k x^k$ has radius of convergence $R = 1$, then $\sum_{k=1}^{\infty} c_k \dfrac{1}{2^k}$ converges. |
| (F) | If $\sum_{k=0}^{\infty} c_k x^k$ has radius of convergence $R = 1$, then $\sum_{k=1}^{\infty} c_k 2^k$ converges. |
| (B) | If $\sum_{k=0}^{\infty} c_k x^k$ has radius of convergence $R = 1$, then $\sum_{k=1}^{\infty} c_k$ converges. |
| (T) | If $\sum \lvert a_n \rvert$ converges, then $\sum a_n$ converges. |
| (F) | If $\sum a_n$ diverges, then $\sum \lvert a_n \rvert$ converges. |
| (B) | If $\sum a_n$ converges, then $\sum \lvert a_n \rvert$ converges. |

FIGURE 6. The ten conditional T/F items and their three presentations: True (T), False (F), and Sometime true/sometimes false (B).