Revised Efficacy Study Proposal for Norton's *Interactive Psychology*

Introduction

Interactive Psychology, an experiential Introductory Psychology textbook, is designed to facilitate student learning through embedded assessments that promote engagement with the material. Initial qualitative evidence suggests that students enjoy using the interactive features of *Interactive Psychology*, and instructors report noticing increased engagement among their students. However, rigorous testing of the potential impacts that these embedded assessments have on student learning and engagement is needed. This evidence-based approach will help Norton build a community of scholars who can improve their teaching through feedback of student performance and who gain sizable benefits from using an online experiential textbook.

This proposal first mentions analysis of existing data to inform the rest of the proposal (Appendix I). This proposal then presents an experimental and correlational approach (Appendix II) to understand the impacts of the embedded assessments on student performance, curiosity, and engagement.

Methods

In Appendix I, I describe analysis of existing data from Dr. Hard's Introductory Psychology class. This lets us know about how "big" an effect we can expect the embedded assessments to have on student learning.

In Phase I of Appendix II, to quantify the effects of the embedded assessments on performance and motivating other behaviors, I propose a controlled experiment where we recruit four groups of participants who either complete the interactives within a study unit or create an outline and who either do Inquizitive or reread the study unit. All students answer some test bank questions at the *start* and *end* of the study so that we can quantify their effect on student learning of the material (i.e., score on questions at *end* minus at *start*). We would also collect pre- and post-measures of curiosity and engagement/interest in the material.

Phase II of Appendix II is meant to complement Phase I. There, I suggest we recruit instructors with large Introductory Psychology courses who will use the textbook in fall 2020. If they assign the interactives for some lessons and not others, we could quantify the effect of interactives on actual student grades within the classroom, helping us to generalize the benefits we would observe in Phase I.

Budget and Timeline

Timeline: I would like to run the controlled experiment highlighted in Phase I of Appendix I ASAP or during summer 2020. Phase II of Appendix I mentions the possibility of recruiting Fall 2020 instructors for naturalistic data collection, so most analyses would be done in the fall, but recruitment would occur in parallel with Phase I.

Personal Compensation: [redacted]

Hours Needed:

1. Rather than apportion hours per phase in the project, I would like to commit up to 100 hours on the project initially and then check back in to announce what progress has been made. At that point, I would have a better estimate of how much work would be left and how many more hours are needed. This is [redacted] upfront, with more costs if 100 hours is not actually sufficient.
2. I'm open to other options. I am used to thinking about work as an hourly rate, and in my experience, unexpected difficulties usually arise in the research process, so pinning down the exact number of hours per project or project phase is hard.
    a. For example, I am not entirely certain of the logistics (e.g., will I recruit instructors for Phase II of Appendix II, or can that be folded into prior discussions with participating instructors?) and existing resources at Norton (e.g., is Appendix II even possible?).

Participant Compensation: Each component of the project has different associated costs. Of the options for participant recruitment during Phase I of Appendix II, I would suggest the Qualtrics Panel ($6/participant, 1,576 participants), and Phase II of Appendix II would cost ~$750-1000 if the other instructor incentives do not work.

Appendix I: Analysis of Existing Data

Part I: Introductory Psychology (PSY101) by Dr. Bridgette Martin Hard (a textbook author)

Introduction

Dr. Hard used *Interactive Psychology* for her PSY101 class in spring 2020. She frequently incorporates "bonus point surveys" into her class – students can earn modest extra credit for completing one of these surveys. As part of her midsemester survey, she asked students for their consent to analyze their course data, and she asked them about behaviors related to their textbook use. I use these data as a means of understanding what we can expect in the controlled and naturalistic studies proposed in Appendix II.

Method

235 students filled out the midsemester survey and consented to have their data from this course analyzed. The midsemester survey included unrelated items (e.g., feedback for teaching assistants) as well as questions related to the use of *Interactive Psychology*:

- How CONSISTENTLY do you do each of the following behaviors WHILE reading the textbook? (1 = never; 2 = sometimes; 3 = about half the time; 4 = most of the time; 5 = always)
    - I complete the "check your understanding" questions at the end of each study unit
    - I complete the interactive features within each study unit
    - I take notes on what I'm reading
    - I think about how what I'm reading relates to, but is also different from, what is covered in lecture.
    - I complete the Inquizitive quizzes assigned for each chapter soon after reading it
- How much did you use each of the available resources specifically to STUDY for your last exam in PSY 101? (1 = not at all; 2 = a little; 3 = a moderate amount; 4 = a lot; 5 = a great deal)
    - Inquizitive quizzes
    - Check Your Understanding questions at the end of each chapter
    - Chapter summaries at the end of each chapter
    - Bolded key terms in each chapter
    - The interactive figures in each chapter

Scores from their first exam as well as their initial Inquizitive scores were analyzed.

Results

How do textbook behaviors impact Inquizitive scores? Although consistently completing the check your understanding (CYU) questions was associated with higher Inquizitive (IQ) scores ($r = 0.16$, $p = .016$), both the CYU measure and IQ scores were highly skewed. This means that most students endorsed "always" consistently completing the CYU questions ($N = 118$, i.e., 50.21%) and most students got 100% on IQ ($N = 177$, i.e., 75.32%). Having a higher frequency of students towards "one end" of the scores on both measures means that a certain group of

students could be driving any associations that we observe, which is not ideal.

Since this proposal is aimed at understanding the impact of the embedded assessments on learning, I separated students into 2 groups: one with students who endorsed "always" or "most of the time" (N = 163, 69.36%) completing the CYU questions and the other with students who endorsed only completing the CYU questions "about half the time," "sometimes", or "never." I also excluded students who scored above 95% or scored 0% on Inquizitive, leaving only 39 students with eligible IQ scores.

Students who usually complete CYU questions (M = 77.38, SD = 18.02, N = 21) do not differ from students who rarely complete CYU questions (M = 77.39, SD = 18.02, N = 18) in terms of IQ scores. For these 39 students, there's also no relationship between completing the interactives or completing the IQ quizzes soon after lecture and their eventual IQ scores. Note that this is a very small portion of students, so we cannot conclusively say much about IQ scores and the embedded assessments.

How do textbook behaviors impact exam scores? Students who usually complete the CYU questions (M = 88.68, SD = 7.54, N = 163) had higher exam 1 scores on average than students who rarely completed the CYU questions (M = 83.42, SD = 11.16, N = 72; $F(1,233) = 17.89$, $p < 0.001$)**.** Consistently completing the CYU questions resulted in a five-point increase on exam 1 and explains about 7% of how widely distributed exam 1 scores are. Accounting for completing the interactives, this relationship is still significant, but then completing CYU questions only explains about 5% of the variance in exam 1 scores.

How much of a "boost" did CYU questions add on top of IQ? Now, let's only analyze the data for students who got 100 on their IQ scores (N = 177). Students who usually complete the CYU questions (M = 89.80, SD = 7.00, N = 130) have higher scores than students who rarely complete the CYU questions (M = 86.00, SD = 9.54, N = 47; $F(1,175) = 8.33$, $p = 0.004$). Even when students had the same score on Inquizitive, consistently completing the CYU questions resulted in a four-point increase in exam 1 scores and explains about 5% of how widely distributed exam 1 scores are. Accounting for completing the interactives, this relationship is still significant, but then completing CYU questions only explains about 3% of the variance in exam 1 scores.

How does Inquizitive impact exam scores? If we only look at the 39 students who scored between 1% and 95% on IQ, there is no relationship between IQ and exam 1 scores, and there is no relationship between completing the IQ soon after lecture and exam 1 scores.

Does consistency matter? Do the same relationships exist when you look at *how much* students relied on a resource before their exam? Not really. Reading the chapter summaries was also "skewed" with most students relying on summaries to study, but there's not a clear pattern in the data. If anything, when students endorse using the interactive figures, IQ, or summaries more to study, their exam 1 scores are worse, but this seems driven by the few students who did poorly.

Caveats: Note that students who complete the CYU questions may differ on a number of other factors that also impact their exam scores. This is shown by how little completing CYU questions explains the spread (3%-7%) of exam 1 scores. Also, these analyses artificially group

students based on self-reported behaviors, and this is a relatively small sample from someone who is also known to be a *really good* instructor.

Conclusion: We have evidence for a small benefit of *consistently* completing the CYU and interactives on exam scores in an Introductory Psychology course. CYU and interactives may also provide a boost not covered by Inquizitive alone.

Appendix II: Complementary Experimental and Naturalistic Testing of the Benefits Derived from Using *Interactive Psychology*

The following appendix presents a detailed plan for how to run complementary controlled and naturalistic studies that test the benefits of using *Interactive Psychology*.

Phase I: Controlled Experiment Testing Benefits of Interactives

Introduction

Interactive Psychology is designed to enhance student learning by "chunking" material into study units, incorporating active practice recalling old material with embedded assessments, including questions at the start of each unit as thematic guides for the material, and more. However, what differentiates *Interactive Psychology* the most from other textbooks is its incorporation of these embedded assessments as a part of the online learning experience. Do these assessments result in better student learning, enhanced motivation, and increased engagement?

Initial evidence from Dr. Hard's Introductory Psychology 101 class in the spring 2020 semester has suggested that exam scores are higher when students complete the check your understanding questions and complete the interactive figures. However, these data are correlational only, and the data are self-reported from students. Previous work (Tingley, 2017) has also suggested that Inquizitive can result in enhanced learning for students, but there was no comparison group, which makes its benefit unclear relative to normal student behaviors. That is, is an 18% boost in grades above average or what you might expect anyway?

By running a controlled experiment, we can probe benefits of using the embedded assessments within *Interactive Psychology* beyond performance metrics (e.g., effects on curiosity). We can also specifically conclude that completing the interactives *caused* benefits to student performance, whereas naturalistic studies (e.g., Dr. Hard's analyses) can only conclude that there is an *association* between completing the interactives and performance metrics.

Method

The check your understanding questions, interactive figures, and Inquizitive are all features designed to test student knowledge of the material they read or previously engaged with. Most studies that have examined the benefit of "testing" (retrieval practice; Dunlosky et al., 2013, *Scientific American MIND*) have incorporated a "reread/study" group (e.g., "study practice", Smith et al., 2016, *Science*). I therefore suggest the following procedure:

As with Tingley (2017), all students will answer a random selection of 3-5 exam-related questions (from the test bank) *before* and *after* completing or not completing interactive features. This will allow us to assess objective *baseline* knowledge of the material and knowledge subsequently gained. At these time points, we will also assess self-reported curiosity, prior knowledge, and confidence about the correct answers. For example, even if there are no performance benefits, teachers may still like to know that their students become more curious about the material after using the interactive features.

- After each test question (Wade and Kidd, 2019, *Psychonomic Bulletin & Review*), students could estimate how close their answer was to the correct answer on a scale of 1 (not close) to 7 (very close), whether they think their answer is correct, and how curious they are about the correct answer on a scale of 1 (not curious) to 7 (very curious).
- Students can also fill out engagement questions, such as those asked by Dr. Hard in her "course engagement" scale (catered to the study unit topic) or modified questions from other scales (e.g., Burch et al., 2015). The latter asks about emotional and cognitive engagement for a class/topic within and outside of the classroom.

Once students answer these questions, they will read the study unit. One group of participants will use the dynamic online version and be asked to complete the interactives included in the lesson, while another group of participants will be given a static pdf that does not include the interactives. This comparison group of participants will create an outline of the material they read; this is a frequent student behavior used as an alternative to testing and should also equate the amount of time that students across groups spend with the material.

Next, students will either do Inquizitive or be asked to reread the study unit. Rereading is a common comparison to "testing" again on the material. Inquizitive in this case is both an outcome variable (e.g., does completing the interactives result in better Inquizitive scores?) and another interactive manipulation designed to show the benefit of using this textbook.

In other words, there are four groups:

1. Pretest questions, self-report curiosity/prior knowledge -> complete interactives -> do Inquizitive -> post-test questions, self-report measures
2. Pretest -> make an outline of the study unit -> do Inquizitive -> post-test
3. Pretest -> complete interactives -> reread study unit -> post-test
4. Pretest -> make an outline of the study unit -> reread study unit -> post-test

By comparing Groups 1 and 4, we will be able to see how much doing *all* of the interactive features within *Interactive Psychology* compares to doing none and falling back on "bad" study behaviors. By comparing Groups 1 and 2, we will be able to see how much completing the interactive features *within* each study unit impacts Inquizitive scores and overall improvement from the pre-test to post-test. By comparing Groups 1 and 3, we will be able to see how much of a benefit Inquizitive adds to someone already using the text interactives. Comparing Groups 2 and 4 should replicate Tingley (2017) and show benefits of using Inquizitive. Finally, by comparing Groups 3 and 4, we will be able to see how much completing the interactive features *without* doing Inquizitive helps performance.

Qualitative survey data suggest that 7/14 instructors were assigning Inquizitive for points, 3/14 were assigning Inquizitive for self-study only, and 4/14 were not assigning Inquizitive at all, so it is important to quantify the testing benefits of the interactives *with* and *without* Inquizitive use.

Most educational studies, in real-world settings of the classroom, have at best an effect of 0.2 (Dynarski, 2017; Yeager et al., 2019), meaning an increase of 0.2 in GPA, for example, per standard deviation unit. This means that when comparing mean scores between two groups, about 57.9% of one group will have a higher score than the other group. Analysis of the data

within Dr. Hard's class (Appendix I) supports the idea that the benefit of the embedded assessments is likely small.

For us to confidently assess whether completing the interactive features result in differences in our variables of interest (80% power for a two-tailed *t-test*), it is suggested that we recruit at least 394 participants per group. Below I include options for recruiting participants.

1. Which Participants Should We Recruit? Recommended: A.
    a. Qualtrics Panel (~$6/participant; $500 minimum)
        i. Estimate from Mitch Chidester (Qualtrics contact) for a 30 minute study where students aged 17-21 who have not taken Introductory Psychology read some sections of a textbook, answer questions about the material that they read, and self-report motivation, engagement, etc.
        ii. Quote may change depending on details
        iii. No additional time needed to recruit participants, but requires inputting the experiment into Qualtrics.
    b. Amazon Mechanical Turk ($3-10/hour)
        i. Crowdsourcing platform of independent contract workers. This would rely on participants to follow instructions and not do the study if they are not undergraduate students, which most are not.
        ii. Costs an extra 20% to add "premium qualifications" (e.g., age), and costs 20% on top of what you pay participants ($3 means $3.6/participant)
        iii. Would also require posting an online link to the experiment, such as Qualtrics
    c. Duke SONA participants (credit)
        i. These are students who are already in a psychology course and need to do studies to earn credit. They may not be the target audience, especially if they already have some psychology knowledge.
        ii. We can put a pre-screen or qualification so that they cannot have taken Introductory Psychology. However, this pushes the timeline of the study until fall 2020, since the last day of classes at Duke is April 22nd.
    d. Duke Interdisciplinary Behavioral Research Center ($10/hour)
        i. Like MTurk/B, this is a community sample of Durham residents, so we would have to ask that participants not sign up unless they were students
        ii. Generally, in my experience, the data quality has been similar to that of B, and B, C, and D rely on recruiting participants through an online platform, whereas A has the company recruit the target demographic

2. Which Study Units / Interactives Should Be Used?
    a. Favorite Interactives as reported by qualitative data
        i. A word cloud count of the 800+ responses to the "favorite interactive figure" question on the qualitative surveys showed that "Brain" or "brain" was mentioned 205 times. Looking through the initial data, this may have been the 3D brain that lets students know about each region's function
    b. Select at least one study unit from each subarea of Psychology (Cognitive/Social/Developmental/Clinical)

        i. Psychology instructors have their own specialties and this allows marketing to that instructor.
        ii. Also makes it less likely that "topic" itself impacts the results, whereas if "brain" is the most commonly mentioned interactive, the study units would be biased towards people who enjoy the neuroscience component
    c. Study units based off the review of *Interactive Psychology* vs. other textbooks
        i. If the material is important enough to be included in multiple textbooks, then it would be considered critical to any Introductory Psychology curriculum.
        ii. This is also a direct comparison to other (less interactive) textbooks
    d. These options may not always be mutually exclusive

Phase II: Recruitment of Instructors in Fall 2020 for Naturalistic Study

Introduction

Psychology is currently undergoing a "credibility revolution" (Vazire, 2018): researchers are trying to promote higher standards for the evidence with which they make strong claims and greater transparency in the research process. As a part of this credibility revolution, researchers have been encouraged to preregister their hypotheses, stating in advance what they expect to find in a particular study and what analyses will be run.

I here propose that in summer 2020, based on the results from Phase I, we privately pre-register an efficacy study plan in which we recruit instructors who will use *Interactive Psychology* in fall 2020. This large naturalistic study will involve a community of teachers willing to engage in scholarship of teaching and learning, which means that we would have to develop an Institutional Review Board protocol where students consent to sharing their grades etc.

The purpose of Phase II is three-fold. First, by recruiting instructors in the research process, this may solidify their partnership with Norton through increased agency. Second, by preregistering hypotheses, we will have increased the credibility of the efficacy study in the eyes of other psychologists, especially if we actually replicate the results of Phase I with Phase II. Finally, Phase II allows us to look at classroom-relevant variables such as grades that are not clearly applicable to a controlled experiment and promotes greater generalization of Phase I results.

Method

I propose that we strategically recruit instructors who have large Introductory Psychology courses. For example, if an instructor has about 300 Introductory Psychology students, we would then need to only recruit 3-4 other instructors for our sample to be similar to that in Phase I.

Some of the Phase II procedure depends on the results of Phase I – for example, if we find that the embedded assessments increase curiosity or persistence, we may want to include measures aimed at assessing these constructs. However, one of the easiest study designs would involve the instructors requiring students to complete the interactives for some lessons (e.g., some of the favorite interactives mentioned in Phase I), but not others. We could then compare performance

(e.g., grades on exam questions, Inquizitive) on the lessons where completing the interactives was and was not required. This repeated-measures within-subjects design allows us to account for naturally occurring, but irrelevant differences in students and may match what instructors normally do with a textbook (i.e., selectively choose which components to assign).

Cost:

1. *Incentive for instructors*.
    a. Some instructors may participate for the sake of SoTL research, particularly if the research is preregistered and could result in a publication of some sort.
    b. Others may want a monetary incentive to participate. In recruiting instructors for other ongoing research at Duke, we have in the past transferred $250 to their discretionary accounts. Teachers use this money for things like conference travel.
    c. Yet another option for incentivizing the instructor would be to allow the students of the instructors' class to use the textbook for free.