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 presents an experimental and correlational approach (Appendix I) to understand the impacts of the embedded assessments on student performance, curiosity, and engagement. The proposal also presents the benefits of enhancing and strengthening Norton's existing framework for naturalistic data collection (i.e., whether students complete the interactives) and the additional insights we could generate from these data (Appendix II).

Methods

In Phase I of Appendix I, 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 I 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.

Finally, Appendix II discusses additional analyses that could be completed with increased collection of user data. For example, we could look at the efficacy of individual interactives and their ability to predict performance, allowing Norton to determine which interactives would need to be updated in the next textbook edition.

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. Finally, collection of user data (Appendix II)

depends largely on programming from the site engineers.

Personal Compensation: This work is akin to a "User Researcher role," and a summer internship as a UX researcher (for PhD students) at [redacted], for instance, pays [redacted] plus the cost of room/board for three months in high rent cities [redacted]. I therefore ask for compensation at a rate of [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 I, 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 I, I would suggest the Qualtrics Panel ($6/participant, 1,576 participants), and Phase II of Appendix I would cost ~$750-1000 if the other instructor incentives do not work. None of the suggested analyses in Appendix II cost additional money for participant recruitment.

Appendix I: 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 1) complete the check your understanding questions; 2) complete the interactive figures; 3) think about how the textbook material relates to the lecture; and 4) do well on the Inquizitive quizzes. 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. 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 three 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.

Appendix II: Using User Data from *Interactive Psychology* for Future Research

Introduction

My current understanding is that, as users use *Interactive Psychology*, data are primarily obtained through Google Analytics. Google Analytics, as deployed on a typical blog, can answer the following questions:

- How do you acquire users? (As in site referrals)
- Where are your users?
- When do your users visit?
- What pages do your users visit?
- How are your active users trending over time?
- What are your top devices?
- How well do you retain users?

In other words, Google Analytics primarily tracks things like session duration (how long students have the browser open), where they're accessing the browser from, etc. These data usually exist on the back-end of most Learning Management Systems (Sakai, Blackboard, etc.) that teachers use. While useful, the data are not immediately relevant to the teachers who simply want to know about the habits of their students. For instance, based on overall session duration, it is hard to know if your student spent that much time on the textbook or was multitasking and left the tab browser open.

If Norton engineers can build a back-end database to store whether its student users have completed the check for understanding question and interactive figures, whether they use the highlight and note-taking tools, and more, this gives Norton immense power to market the textbook to teachers and release subsequent editions with improved embedded assessments. For example, teachers will be able to understand more about student behaviors that facilitate their understanding and help teachers improve their lesson plans (e.g., all the students completed Interactive 5.1, but had to complete more Inquizitive questions on the material; perhaps more time is needed in the lesson plan on this material). Norton will also have data that can complement that from Appendix I, showing how individual embedded assessments relate to performance within each study unit.

Method

If Norton collects user data from the textbook beyond Inquizitive and Google Analytics, several additional research-related questions can be answered. For instance:

- *Is completing the embedded assessments associated with increased subsequent persistence*? Here, we treat completion of the interactives as an independent variable that predicts the likelihood that student users continue to answer Inquizitive questions. We could define persistence as the number of questions they answer beyond the required level for their course or the number of times they repeat either individual questions or questions that comprised the required level to advance in their course.
- *Is completing the embedded assessments associated with increased sustained attention*? Sustained attention could be captured by looking at the times associated with when the

students interact with the page. In other words, do students get distracted and turn to another browser, or do they continue to read on in the material (i.e., timings spaced close together)? If testing increased engagement with the material, it is possible that students stay "on-task" more.

This is in addition to the basic questions addressed in Appendix I about relationships between performance (e.g., Inquizitive scores) and completing check for understanding questions and the interactive figures. For example, we could look at whether completing the interactives is associated with benefits at all levels of the Inquizitive scores or their confidence ratings on questions ("metacognition").

By tracking how users use the textbook, we could also quantify "comparison" groups of students – i.e., students who only highlight the text they are reading but don't complete the interactives, students who don't do anything with the text despite being registered for a login, students who read the text "all at once" vs. consistently interact with the text, etc.

Moreover, the statistical relationships between the variables mentioned above may be stronger for some and weaker for others. This approach gives Norton quantifiable data on which interactive figures could potentially be improved, increasing their efficacy with regard to student performance. Future editions could be released not only with updated references and material, but with updated interactives that promote student learning more. For example, if an interactive is only weakly or not at all related to performance metrics like Inquizitive scores, the authors may wish to consider redesigning the interactive.

Timeline. The timeline largely depends on how difficult it would be to store the data for future research. Once stored, the analyses mentioned above are similar to those in Phase II of Appendix I. The **costs** involve the personnel analyzing the data and creating the database.