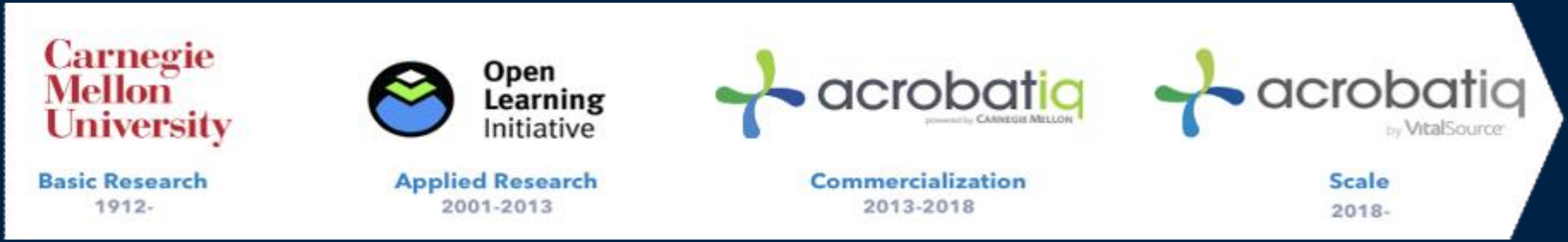
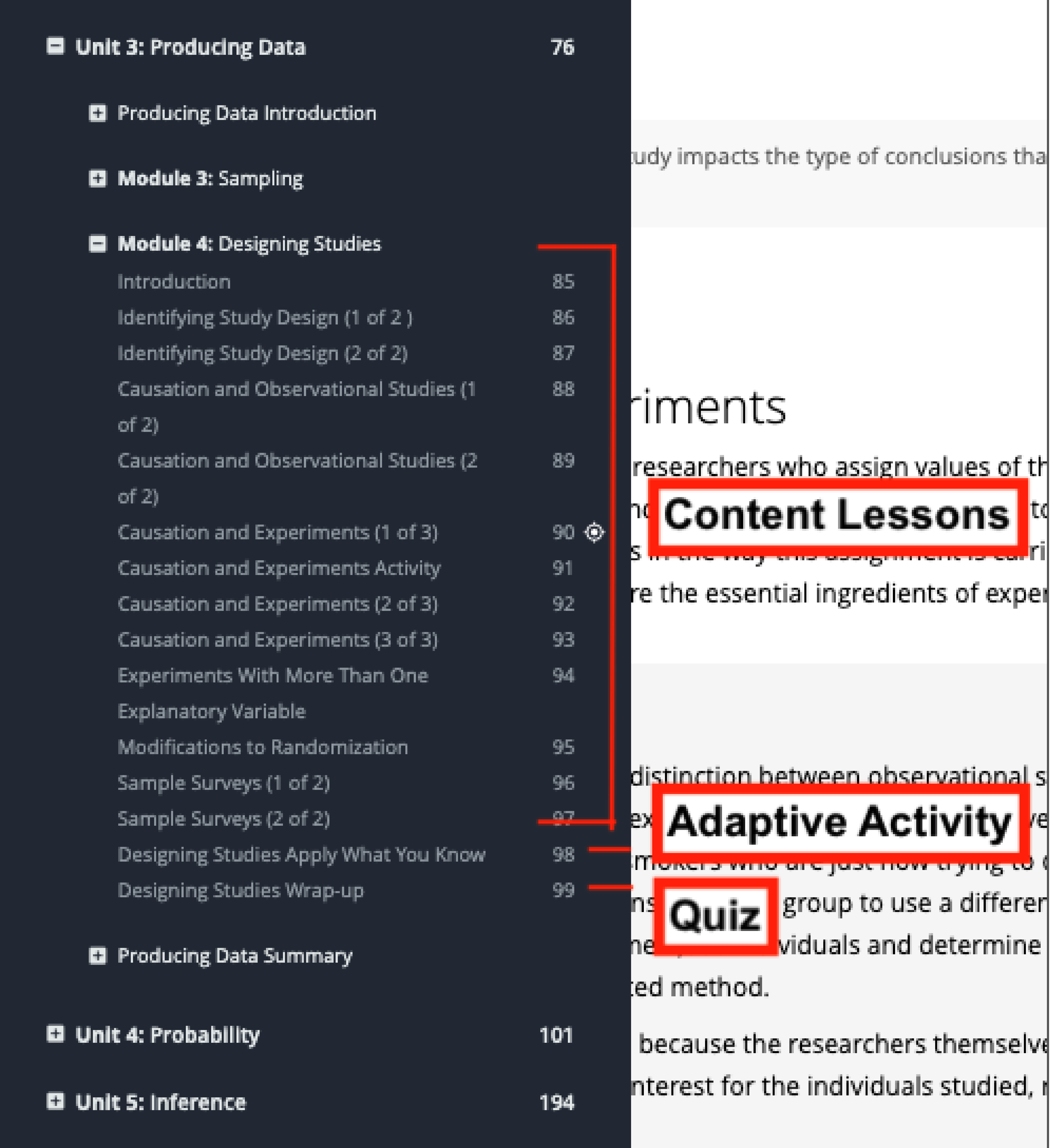
The Doer Effect



acrobatiq by VitalSource

Dr. Benny Johnson
Rachel Van Campenhout
Research and Development
VitalSource Technologies

Courseware



Unit 3: Producing Data 76

Producing Data Introduction

Module 3: Sampling

Module 4: Designing Studies

Introduction 85

Identifying Study Design (1 of 2) 86

Identifying Study Design (2 of 2) 87

Causation and Observational Studies (1 of 2) 88

Causation and Observational Studies (2 of 2) 89

Causation and Experiments (1 of 3) 90

Causation and Experiments Activity 91

Causation and Experiments (2 of 3) 92

Causation and Experiments (3 of 3) 93

Experiments With More Than One Explanatory Variable 94

Modifications to Randomization 95

Sample Surveys (1 of 2) 96

Sample Surveys (2 of 2) 97

Designing Studies Apply What You Know 98

Designing Studies Wrap-up 99

Producing Data Summary

Unit 4: Probability 101

Unit 5: Inference 194

udy impacts the type of conclusions tha

riments

researchers who assign values of th

re the essential ingredients of exper

distinction between observational s

group to use a differen

viduals and determine

ted method.

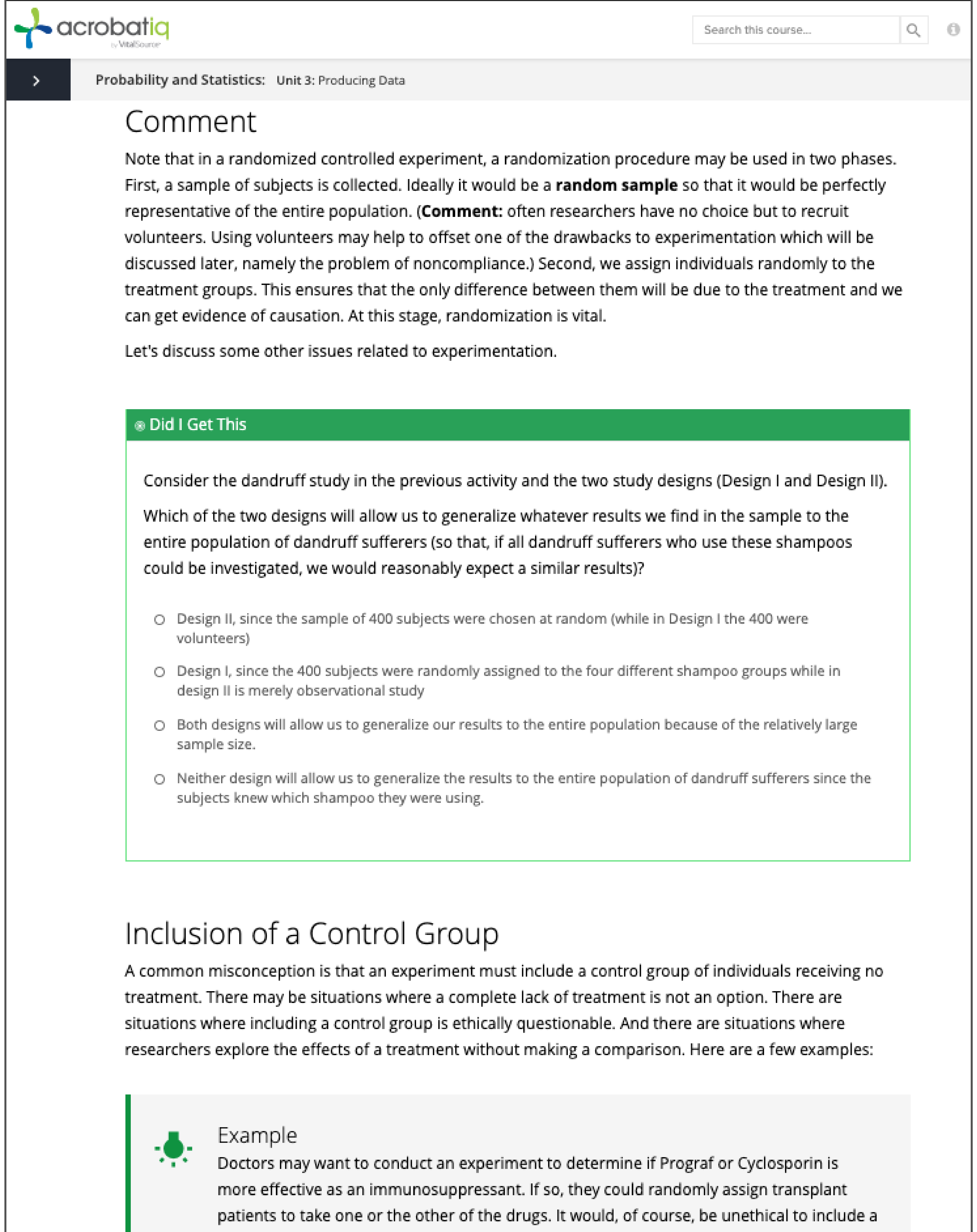
because the researchers themselve

nterest for the individuals studied,

Courseware is a comprehensive learning environment that provides lessons of content interleaved with formative practice, followed by an adaptive activity and a graded quiz.

Learn by Doing

The primary method used in the courseware is learn by doing: integrating formative practice questions into the text at frequent intervals.



Comment

Note that in a randomized controlled experiment, a randomization procedure may be used in two phases. First, a sample of subjects is collected. Ideally it would be a **random sample** so that it would be perfectly representative of the entire population. (**Comment:** often researchers have no choice but to recruit volunteers. Using volunteers may help to offset one of the drawbacks to experimentation which will be discussed later, namely the problem of noncompliance.) Second, we assign individuals randomly to the treatment groups. This ensures that the only difference between them will be due to the treatment and we can get evidence of causation. At this stage, randomization is vital.

Let's discuss some other issues related to experimentation.

Did I Get This

Consider the dandruff study in the previous activity and the two study designs (Design I and Design II).

Which of the two designs will allow us to generalize whatever results we find in the sample to the entire population of dandruff sufferers (so that, if all dandruff sufferers who use these shampoos could be investigated, we would reasonably expect a similar results)?

○ Design II, since the sample of 400 subjects were chosen at random (while in Design I the 400 were volunteers)

○ Design I, since the 400 subjects were randomly assigned to the four different shampoo groups while in design II is merely observational study

○ Both designs will allow us to generalize our results to the entire population because of the relatively large sample size.

○ Neither design will allow us to generalize the results to the entire population of dandruff sufferers since the subjects knew which shampoo they were using.

Inclusion of a Control Group

A common misconception is that an experiment must include a control group of individuals receiving no treatment. There may be situations where a complete lack of treatment is not an option. There are situations where including a control group is ethically questionable. And there are situations where researchers explore the effects of a treatment without making a comparison. Here are a few examples:

Example

Doctors may want to conduct an experiment to determine if Prograf or Cyclosporin is more effective as an immunosuppressant. If so, they could randomly assign transplant patients to take one or the other of the drugs. It would, of course, be unethical to include a

The Doer Effect

The doer effect is the learning science principle that the amount of interactive practice a student does (such as answering practice questions) is much more predictive of learning than the amount of passive reading or video watching the student does. [1]

Doing **practice** has

6x

the effect size

than **reading** alone.

The Doer Effect

The doer effect was investigated at Carnegie Mellon University by Koedinger et al. and was shown to be causal. [2, 3]

Doing more practice caused better learning.

The regression model controls for the amount of reading, watching, and doing in outside units, to control for a third variable [2].

Within Reading	Outside Reading
Within Watching	Outside Watching
Within Doing	Outside Doing

The Doer Effect:
Replicating Findings that Doing Causes Learning

Rachel Van Campenhout & Benny G. Johnson
Research and Development
VitalSource Technologies
Pittsburgh, USA

Jenna A. Olsen
Learning Analytics
Western Governors University
Salt Lake City, USA

The Goal of this Study

This paper aims to replicate previous causal doer effect research to:

- Identify if a similar learning environment using the same learning by doing methods can produce similar results
- Extend the external validity of these learning methods
- Provide additional evidence that this learning science principle should be scaled

Methods

- 3,120 students included from a Macroeconomics course from March 2017 to April 2019
- 6 course competencies are used as the unit, with 47 learning objectives mapped to the competencies
- Final exam questions were similarly mapped to the 6 competencies

WGU®
WESTERN GOVERNORS UNIVERSITY

Results

Mixed effects linear regression model

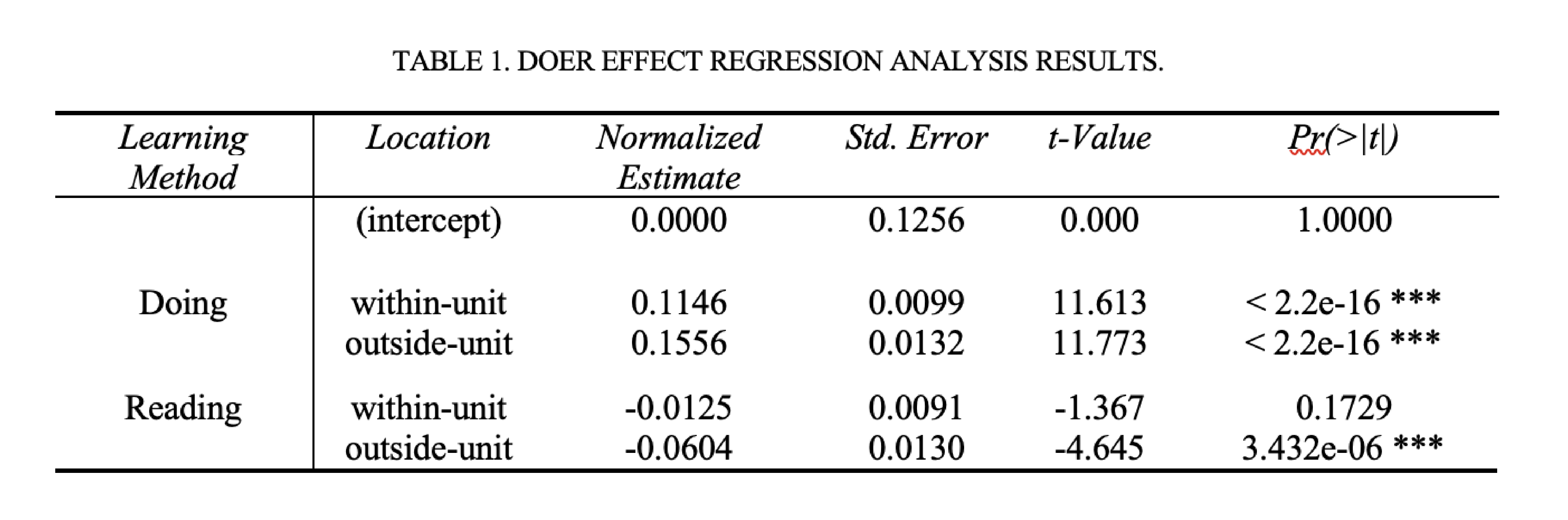


TABLE 1. DOER EFFECT REGRESSION ANALYSIS RESULTS.

Learning Method	*Location*	*Normalized Estimate*	*Std. Error*	*t-Value*	*Pr(>|t|)*
	(intercept)	0.0000	0.1256	0.000	1.0000
Doing	within-unit	0.1146	0.0099	11.613	< 2.2e-16 ***
	outside-unit	0.1556	0.0132	11.773	< 2.2e-16 ***
Reading	within-unit	-0.0125	0.0091	-1.367	0.1729
	outside-unit	-0.0604	0.0130	-4.645	3.432e-06 ***

- Both within-unit doing and outside-unit doing were strongly, positively significant.
- We would likely expect outside-unit doing to almost always be significant (regardless of whether the doer effect is causal), as it is well known that students who do more practice tend to get better outcomes.
- What matters is that within-unit doing is additionally significant, which means the relationship of within-unit doing to its own unit’s assessment score cannot be accounted for by the amount of outside-unit doing, indicating that relationship is causal in nature.

Conclusion

- **This analysis confirms that even when controlling for an outside variable, doing the formative practice within the courseware caused better performance on an external final exam.**

- **Doing practice *causes* better learning.**

What’s Next?
Automatic Question Generation

The doer effect is proven to be causal, but this method requires hundreds to thousands of practice questions.

We have been working on using artificial intelligence to generate practice questions from textbook content in order to create “base” courseware nearly instantaneously.

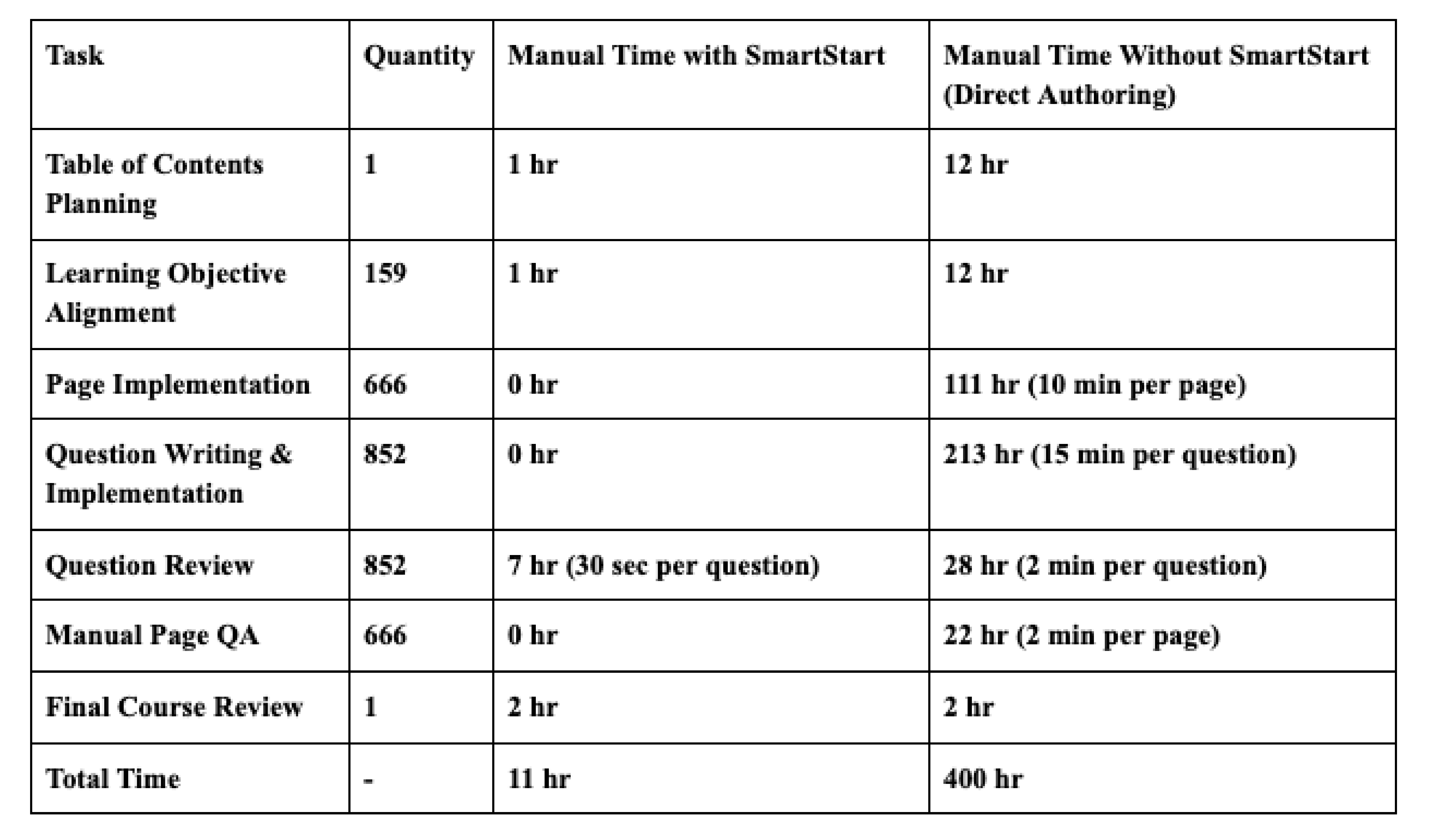
acrobatiq
by VitalSource

Transforming Textbooks into Learning by Doing Environments:

An Evaluation of Textbook-Based Automatic Question Generation

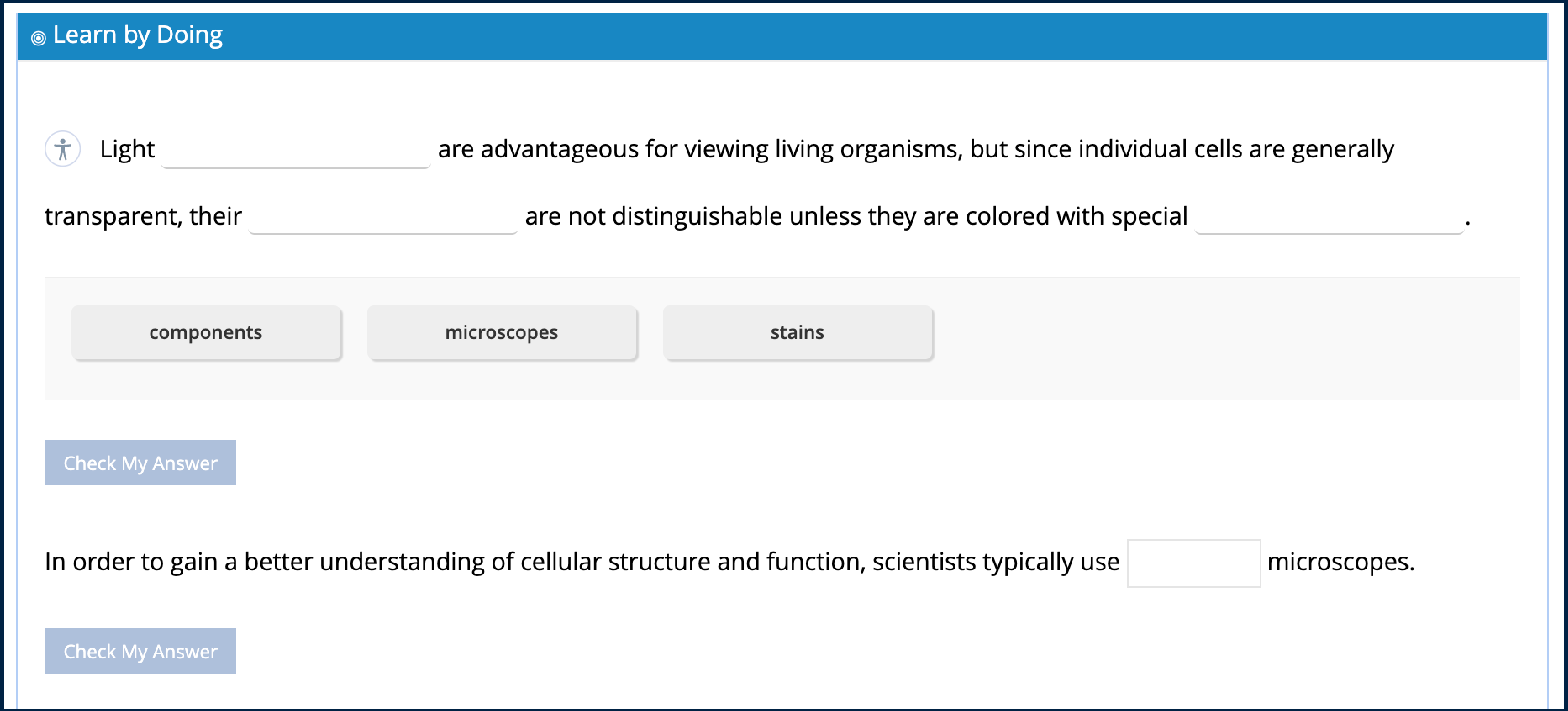
Scaling the Doer Effect with AI

Formative practice is good for students, but costly to create.
Automatic question generation can solve this problem.



Task	Quantity	Manual Time with SmartStart	Manual Time Without SmartStart (Direct Authoring)
Table of Contents Planning	1	1 hr	12 hr
Learning Objective Alignment	159	1 hr	12 hr
Page Implementation	666	0 hr	111 hr (10 min per page)
Question Writing & Implementation	852	0 hr	213 hr (15 min per question)
Question Review	852	7 hr (30 sec per question)	28 hr (2 min per question)
Manual Page QA	666	0 hr	22 hr (2 min per page)
Final Course Review	1	2 hr	2 hr
Total Time	-	11 hr	400 hr

AG Questions



Light ______________ are advantageous for viewing living organisms, but since individual cells are generally transparent, their ______________ are not distinguishable unless they are colored with special ______________.

In order to gain a better understanding of cellular structure and function, scientists typically use ________ microscopes.

The State of Research

In Kurdi et al.'s 2020 systematic review of research, there are gaps identified in the current research:

- Only 1 study evaluated AG questions in a classroom setting
- Only 1 study generated feedback
- Only 1 study identified Bloom's level
- Only 14 studies evaluated question difficulty
- **There is no clear "gold standard" identified**

Evaluating Questions

We compare our AG questions to HA questions in the same course using a mixed-effects logistic regression model.

In the same course, how do our AG questions compare to HA questions on:

- Engagement
- Difficulty
- Persistence

The Data: 786,242 total observations

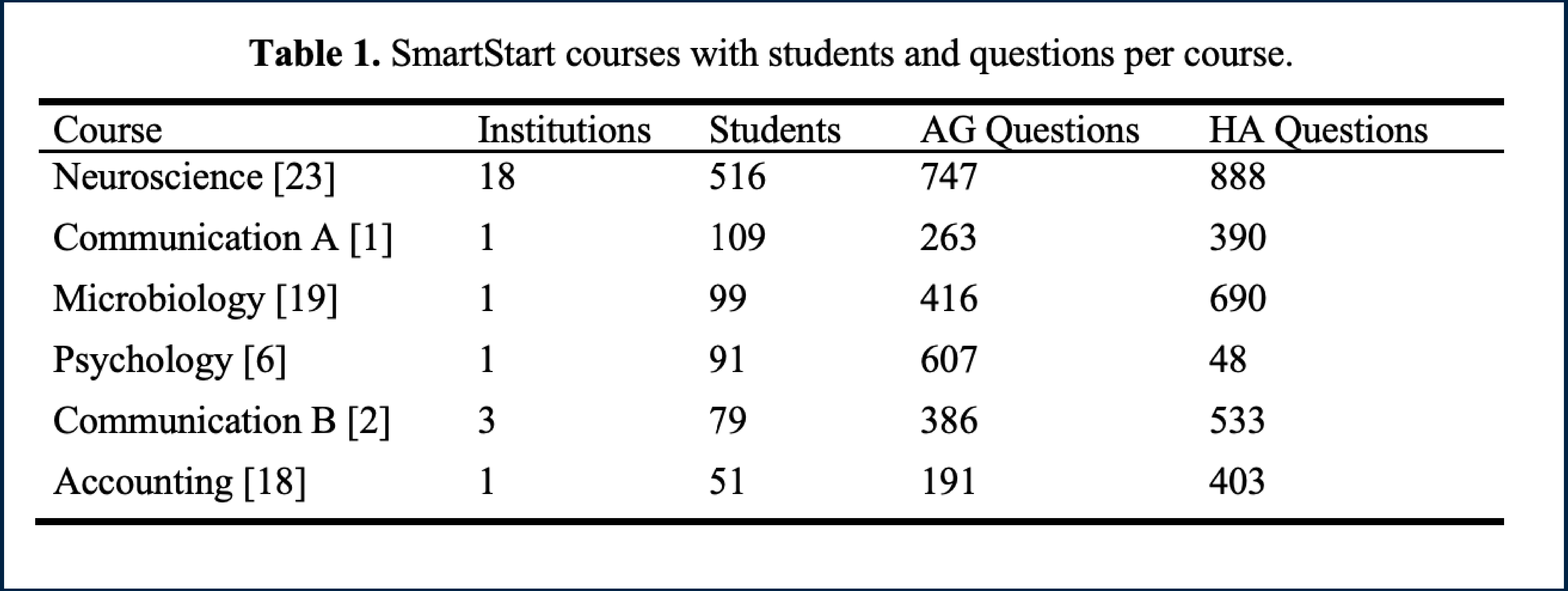


Table 1. SmartStart courses with students and questions per course.

Course	Institutions	Students	AG Questions	HA Questions
Neuroscience [23]	18	516	747	888
Communication A [1]	1	109	263	390
Microbiology [19]	1	99	416	690
Psychology [6]	1	91	607	48
Communication B [2]	3	79	386	533
Accounting [18]	1	51	191	403

Engagement

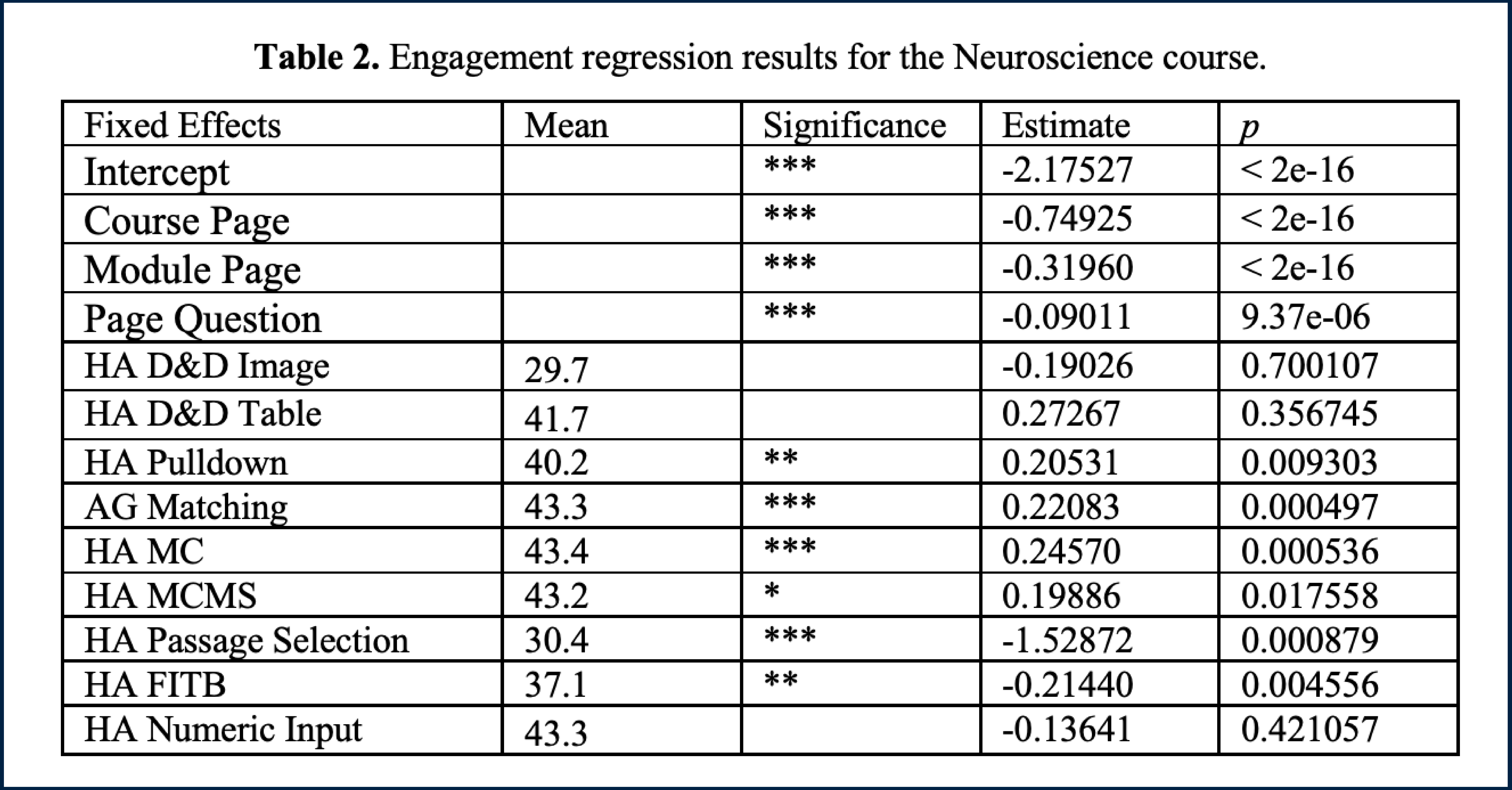


Table 2. Engagement regression results for the Neuroscience course.

Fixed Effects	Mean	Significance	Estimate	*p*
Intercept		***	-2.17527	< 2e-16
Course Page		***	-0.74925	< 2e-16
Module Page		***	-0.31960	< 2e-16
Page Question		***	-0.09011	9.37e-06
HA D&D Image	29.7		-0.19026	0.700107
HA D&D Table	41.7		0.27267	0.356745
HA Pulldown	40.2	**	0.20531	0.009303
AG Matching	43.3	***	0.22083	0.000497
HA MC	43.4	***	0.24570	0.000536
HA MCMS	43.2	*	0.19886	0.017558
HA Passage Selection	30.4	***	-1.52872	0.000879
HA FITB	37.1	**	-0.21440	0.004556
HA Numeric Input	43.3		-0.13641	0.421057

Engagement

To summarize the engagement analysis:

- For all questions, the location of the question in the course mattered.
- The recognition question types had higher engagement than the recall question types.
- The AG recognition type is similar to HA recognition types, and the AG recall type is similar to the HA recall type.
- There is no indication that students found the AG question types problematic in general and chose to answer them less frequently.

Difficulty

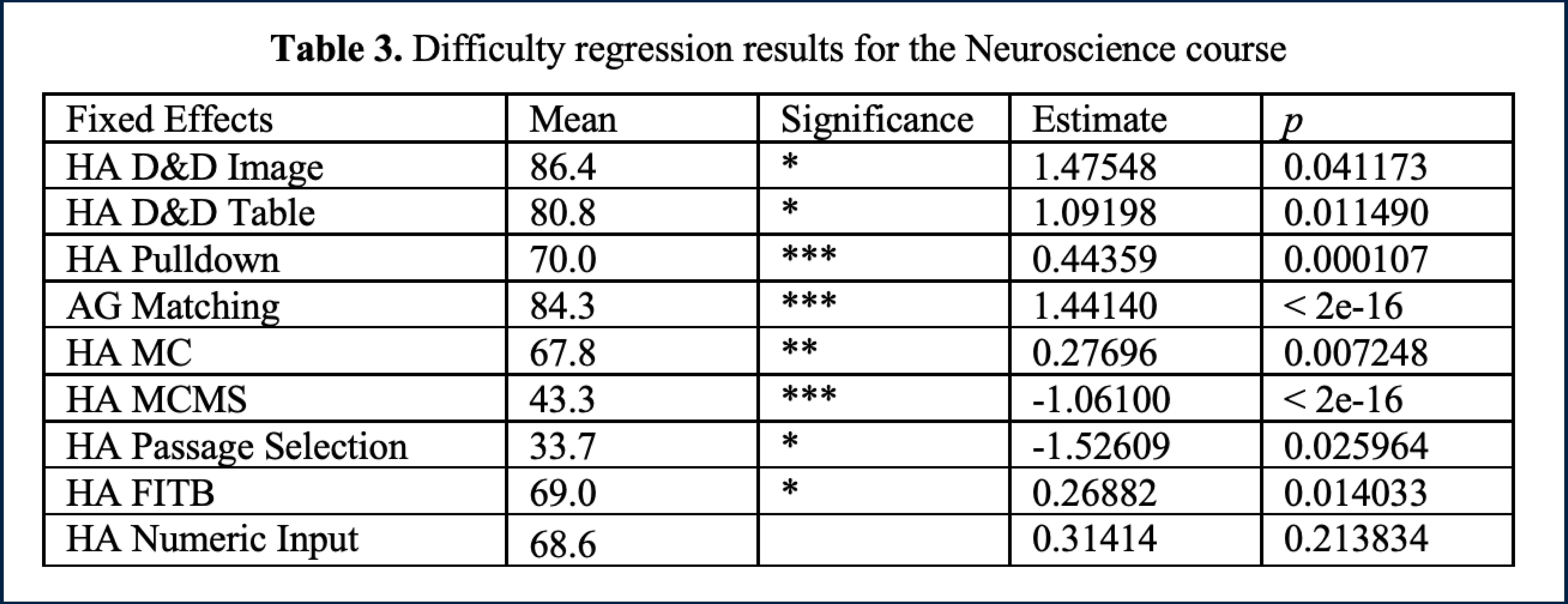


Table 3. Difficulty regression results for the Neuroscience course

Fixed Effects	Mean	Significance	Estimate	*p*
HA D&D Image	86.4	*	1.47548	0.041173
HA D&D Table	80.8	*	1.09198	0.011490
HA Pulldown	70.0	***	0.44359	0.000107
AG Matching	84.3	***	1.44140	< 2e-16
HA MC	67.8	**	0.27696	0.007248
HA MCMS	43.3	***	-1.06100	< 2e-16
HA Passage Selection	33.7	*	-1.52609	0.025964
HA FITB	69.0	*	0.26882	0.014033
HA Numeric Input	68.6		0.31414	0.213834

Difficulty

To summarize the difficulty analysis:

- AQ matching were often the the easiest, but similar to other HA recognition types.
- AG FITB were similar to several other HA questions, which were recognition types.
- HA FITB were often marginally to significantly more difficult than AG FITB.
- Questions varied in difficulty across courses, showing the impact of content.

Persistence

To summarize the persistence analysis:

- AG FITB had statistically different persistence from all questions except HA MCMS.
- HA FITB generally had lower persistence than AG FITB.
- AG matching had similar persistence to other HA questions.

In Summary

Students were not deterred by AG questions.

Engagement was impacted by placement in the course, and recognition vs recall types.

AG questions were in the difficulty range of the HA questions.

Easy questions did not deter persistence, but very difficult questions did.

References

1. K. Koedinger, J. Kim, J. Jia, E. McLaughlin, and N. Bier, “Learning is not a spectator sport: doing is better than watching for learning from a MOOC.” In: Learning at Scale, pp. 111–120, 2015. Vancouver, Canada. http://dx.doi.org/10.1145/2724660.2724681
2. K. Koedinger, E. McLaughlin, J. Jia, and N. Bier, “Is the doer effect a causal relationship? How can we tell and why it’s important.” Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, LAK 2016, pp. 388-397. http://dx.doi.org/10.1145/2883851.2883957
3. K. R. Koedinger, R. Scheines, and P. Schaldenbrand, “Is the doer effect robust across multiple data sets?” Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, pp. 369–375.
4. Van Campenhout, R. Johnson, B. G., & Olsen, J. A. (2021). The Doer Effect: Replicating Findings that Doing Causes Learning. Presented at eLmL 2021 : The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning. ISSN 2308-4367, pp. 1–6. Retrieved from: https://www.thinkmind.org/index.php?view=article&articleid=elml_2021_1_10_58001
5. Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: an evaluation of textbook-based automatic question generation. In: Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education. CEUR Workshop Proceedings, ISSN 1613-0073, pp. 1–12. Retrieved from: http://ceur-ws.org/Vol-2895/paper06.pdf

Thank You!