NUANCED VIEWS OF PEDAGOGICAL EVALUATION

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Computer Science

---

by
Kevin M. Storer
August 2017

Accepted by:
Dr. Jacob Sorber, Committee Chair
Dr. Kelly Caine
Dr. Bart Knijnenburg

ABSTRACT

Novel technologies and teaching methods are being integrated into post-secondary classrooms at unprecedented rates. Educators must be certain implemented changes provide equitable alternatives to traditional pedagogical practices, before migrating to modern paradigms. However, changes in classroom practices each introduce unique psychological influences into the classroom, which may influence traditional metrics of pedagogical success. Evaluation methods for assessing classroom changes should evolve with pedagogy, to accurately measure its effects.

In this paper, we share our experience exploring the equity of replacing traditional paper-and-pencil testing with digital examinations, to demonstrate the complexity of modern educational assessment. We observed a variety of variables, many of which often go unmeasured in pedagogical evaluations, influenced our metrics of success, and altered our initial conclusions. Further, our analysis suggests classroom changes may influence demographic groups differently, and these effects may be hidden, if success is only measured within the aggregate student body. Evaluations that do not explicitly consider and identify these effects may not be able to observe them. Our experience indicates that evaluations that do not account for demographics are incomplete, at best. More importantly, our results suggest these evaluations may draw, and support, invalid conclusions.

By describing the difficulties we encountered in evaluating the equity of digital examinations, we invite educators and education researchers to move beyond simplistic evaluations, consider the underlying factors influencing traditional metrics of success, and adopt a more nuanced view of pedagogical evaluation practices.

ACKNOWLEDGMENTS

Completing this thesis, attaining this degree, and maintaining my sanity, would not have been possible alone. I want to specifically thank:

My advisor, Jacob Sorber. You have given me your time, your advice, and your encouragement, without hesitation. Thank you for being a surrogate parent, a genuine friend, and an astounding mentor. Most importantly, thank you for believing in me and teaching me to keep "swinging for the fence." This journey would not have been possible without your support, and guidance.

My committee members, Kelly Caine and Bart Knijnenburg. Thank you for fostering my interest in human-centered computing, and human subjects research. Without your guidance, I would not have the appreciation for HCC that I have today, or have made the decision to pursue my PhD in Informatics. I am incredibly grateful for your willingness to share your knowledge, provide references, and connect me to the larger HCI community. I am looking forward to many future collaborations.

My friend, Josiah Hester. Thank. You. For. Everything. These last two years were quite an adventure, and having your support through the failures, encouragement toward successes, and friendship in a foreign place, were central to this experience being formative. Your commitment to science demonstrates what an honor it is to be an academic. Your passion for your research encouraged me to search for a field that made me feel equally passionate. I am a better scholar because of the example you set, and am both grateful and proud to know you. Let's go kill it.

My parents, Jim and Karen Wicker. Thank you for your constant support and encouragement, and for shaping who I am today. I would not be here, on Earth or at Clemson, without you. I am truly lucky to have such supporting, compassionate people in my life. You have always made big dreams feel possible. Thank you for sharing in my joys, disappointments, and victories.

My rock, Kenneth. Thank you for forcing me to take a break, reassuring me that everything will work out, and affirming me when I did not believe in myself. Your patience, kindness, and flexibility have kept me afloat, more frequently than you know. I cannot wait to embark on this new chapter with you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

0.1 Introduction

Integrating modern technology into the classroom creates opportunities for novel teaching methods and pedagogical practices. Already, web-mediated discussions, digital homework assignments, and online courses are common in universities across America. In disciplines like Computer Science, where course content is easily translated to a digital medium, the potential for educational innovation is even greater.

Current research identifies unique psychological effects created by technology. For instance, factors like technology anxiety and conceptions of computer self-efficacy influence individuals differently [5, 4]. Typical computer usage, and prior experience with technology affect feelings of competency and comfort. In America, computer usage and access to technology are closely related with household income, gender, and race. The effects of technology anxiety may be higher, and computer self-efficacy may be lower, for underrepresented groups. This may contribute to the disproportionately low number of women and people of color seen in computing professions.

Computer Science educators are aware of the need to broaden participation of underrepresented students in Computing, and research in engagement is growing. Still, the effects of everyday educational practices on underserved demographic groups often remain unmeasured in pedagogical evaluations. Research that does include these observations is frequently conducted for the express purpose of engaging underrepresented students in Computing. Of the 105 papers accepted to SIGCSE 2016 [1], 91 included human participants in their evaluations. However, only 35 reported either the racial or gender composition of their participant group. When these factors are ignored, the impacts of classroom practices may be minimized, or go unnoticed. Identifying the effects of educational systems on underrepresented groups, and ensuring equity in the classroom, is important. Including these observations in standard evaluation practices may provide a more complete picture of the impact of pedagogical changes, in addition to helping broaden participation of underrepresented students.

Educators who want to ensure novel teaching practices are equal alternatives to traditional practices should be certain these systems meet their goals, before adopting them in their classroom. Likewise, education researchers should thoroughly evaluate the equity and effectiveness of pedagogical changes, before encouraging their widespread use. However, effectively evaluating modern teaching methods is complex, and requires a deeper understanding of the effects of technology, like technology anxiety and computerphobia.

To demonstrate the complexity of pedagogical evaluations, and to show that real effects of educational practices may be obscured by observing only effects on the aggregate student body, we share our experience measuring the success of digital examinations in a Junior-level Operating Systems course. To determine whether digital testing had any advantage or disadvantage over traditional paper-and-pencil testing, we examined 93 students' exam scores and anxiety levels for one exam of each type.

At the end of the semester, our results indicated changing examination medium did not affect students' mean exam scores, when measured within the aggregate group. While we might have concluded digital exams are an equitable alternative to paper-and-pencil, examining performance within certain demographic subgroups of our classroom suggested otherwise. In our classroom, changes in examination medium appeared to affect women's performance, and differently influence state anxiety at test time low- and high-anxiety students. Without observing these groups individually, we would not have observed these effects, and would have considered our pedagogical change a success.

Considering and understanding the effects of demographics on metrics of success may change the conclusions of many educational evaluations, as shown in our example. When they do, it is impossible to ensure an evaluation's results are generalizable to classrooms whose demographic makeup differs from the evaluated course. Demographic factors should be collected, accounted for, and reported, in all pedagogical evaluations, so educators can understand the applicability of conclusions to their classrooms.

Evaluations not accounting for demographic factors contributing to students' success may leave important conclusions undiscovered. We share our experience to encourage educators and researchers to adopt a more nuanced view of assessment, that considers the many variables influencing student outcomes, in future pedagogical evaluations.

0.2 Background

Many factors influence educators' decisions to implement digital examinations in their classrooms. Digital examinations can be automatically graded, and are easily randomized, better preventing cheating, or strategic gaming of the exam. However, the time spent creating a secure, digital testing environment can be greater than writing a paper examination. Frequently, computer-equipped lab spaces hold fewer students than a typical classroom. It may be impossible to test the whole class concurrently.

Context-dependent memory is perhaps the most pervasive and compelling argument against the use of digital testing [14]. Memory for a specific task or topic is stronger when used within the same context, and digital examinations often do not occur in the lecture space. Students' recall of course content is hurt by this shift in context.

Digital examinations may hurt high-anxiety students. Test anxiety, and its impact on students' exam performance, has been well explored [10, 20, 8, 3], and show that anxiety and emotional self-regulation are affected simply by being in a testing environment [11]. Changing the context or procedure of exams, like testing medium, may affect students' anxiety levels, and performance. Further, tech anxiety, which more strongly affects individuals with less prior experience and engagement with technology, may inherently increase students' anxiety on a computer-mediated test [5].

Digital examinations may hurt women. The gender disparity in computer usage and comfort has been decreasing slowly, but computer usage and comfort are still significantly higher in men, suggesting they may receive an advantage on digital examinations [6]. Women rate their computing skills and technological competence lower than men [13], and are more likely to attribute failures in computing tasks to their own incompetence [7]. Women rate their computing self-efficacy lower than men[4], and experience higher levels of tech anxiety and computerphobia [16]. Further, women may be less comfortable engaging with a digital examination, and may be more comfortable engaging with paper media. They spend more time reading, read more classical literature, and are more likely to read long books [17].

This existing literature suggests digital media and technology introduce unique psychological influences into the classroom. Further, these phenomena may more significantly impact high-anxiety students, and underrepresented groups including women and students of color.

Because these factors may influence students' success, we explored the impact of implementing digital exams in the classroom, by collecting demographic and performance data three times, over a 16-week course, from 99 participants in a Junior-level, Operating Systems course, at Clemson University.

0.3 Method

We administered five surveys over the course of a 16-week semester, to 99 students enrolled in a Junior-level, Operating Systems course, at Clemson University. In this course, there are two examinations – a traditional paper-and-pencil midterm, and a digital final.

We gathered demographic information early in the semester, including major, class standing, gender, cumulative and major GPAs, expected final course grade, reason for enrollment, typical study and sleep habits, anticipated excitement, confidence, interest, and difficulty of the course. These surveys also provided the State-Trait Anxiety Inventory (STAI) [15], and the Positive and Negative Affect Scale (PANAS) [19], to establish a baseline for students' tendency toward anxiety, and anxiety and mood on a typical lecture day.

Prior to both the midterm and final examinations, students completed surveys, to report their anticipated exam grade, hours spent studying for the exam, and hours slept the night prior. Additionally, they completed the PANAS and the state anxiety form of the STAI.

After each examination, students completed another survey, reporting the grade they expected to receive on the exam, given their performance. Again, students completed the PANAS and the state anxiety form of the STAI.

At the end of this course, we linked students' responses to their exam scores as a percentage. In total, we collected data from 99 unique participants. Of these, 87 completed the initial demographic survey. 93 students completed both the survey prior to the midterm exam, and the survey following it. 74 students completed the survey prior to the final examination, and 73 completed the survey following it. 63 students completed all five surveys. Of the students who completed all five surveys, 56 identified as men, and 7 identified as women.

We analyzed our data set using dependent and independent *t*-tests, and subsequently multiple linear regressions and mediation analysis. Initially, we were interested in identifying whether the medium of the test affected students' exam scores and state anxiety at test time. Upon further analysis, we became interested in identifying whether the medium of the test equally affected students with high and low trait anxiety, whether the medium of the test equitably affected men and women, and which collected demographic factors influenced students' outcomes.

We use these observations to show conclusions of pedagogical evaluations may differ, depending upon the evaluation techniques employed, and invite Education researchers to consider the variety of factors that may influence results.

Observed Predictors of Success Metrics			
Metric	*Predictor*	*p*	R^2
Exam Score	Cumulative GPA	<0.001	0.236
	Major GPA	<0.001	0.235
	Expected Course Grade	<0.001	0.124
	Expected Exam Grade	0.031	0.045
	Confidence	0.043	0.033
	Typical Hours Slept	0.064	0.033
	Major	0.066	0.033
	Hours Studied	0.089	0.031
State Anxiety	Affect	<0.001	0.622
	Trait Anxiety	<0.001	0.437
	Expected Exam Grade	<0.001	0.087
	Hours Studied	0.016	0.055
	Gender	0.021	0.031
	Typical Hours Slept	0.023	0.046
	Expected Course Grade	0.058	0.026

Table 1: Our observed predictors of metrics of success, Exam Score and State Anxiety Prior to Testing, identified using linear regression, with p-values $<$ 0.10, ordered by significance and effect size. Many variables were related each outcome, and underlying factors, including participants' demographics, affected the conclusions of our evaluation.

0.4 Results

We first analyzed our data for significance and effect size, using both independent and dependent *t*-tests as appropriate, to determine whether exam medium affected mean exam scores and state anxiety at test time. To account for data mortality, as many students withdrew from the course during the semester, we report only results observed within the group of 63 students who completed all five surveys. While this relatively small sample size limits the power to which we may observe the effects of examination medium, we seek only to use these results as an example of the potential for analysis methods, and consideration of demographics, to change the conclusions of a pedagogical evaluation. Because of this, we treat all effects observed with p <0.1 as findings meriting discussion.

Measured within the aggregate student sample, examination medium did not affect average test performance (*t*(62) = -1.462, p = 0.147, 95% CI [-8.206, 1.253]), or state anxiety at test time (*t*(59) = -1.230, p = 0.223, 95% CI [-7.524, 1.794]). However, we observed a small increase in anxiety, as compared to a typical lecture day, immediately before and after testing in both paper and digital mediums (Respectively: *t*(59) = -7.681, p <0.001, 95% CI [7.863, 13.403]; *t*(62) = 7.589, p <0.001, 95% CI [7.576, 12.995]).

We split students into high and low trait anxiety groups, where our high-anxiety group included all students whose trait anxiety was higher than 39 on the STAI, a standard level for this distinction. We observed no difference in state anxiety prior to testing in the paper and digital mediums, for high-anxiety students (*t*(32) = -1.219, p = 0.2319, 95% CI [-6.315, 1.588]). However, we observed a significant difference in anxiety level for low-anxiety students, where higher state anxiety was observed prior to the digital final (*t*(23) = 1.868, p = 0.074, 95% CI [-0.313, 6.147]). Our high-anxiety students experienced significantly higher levels of state anxiety in the daily classroom, prior to the midterm, and prior to the final, than their low-anxiety peers (Respectively: *t*(54.983) = 4.358, p <0.001, 95% CI [5.659, 15.296]; *t*(52.668) = 6.119, p <0.001, 95% CI [10.306, 20.360]; *t*(56.763) = 3.639, p <0.001, 95% CI [4.183, 14.423]).

In our sample, gender had the largest disparities in performance and state anxiety, across exam medium. Women and men experienced no significant difference in state anxiety, prior to the paper-and-pencil midterm (*t*(5.885) = 0.792, p = 0.460, 95% CI [-10.060, 19.615]). Conversely, women reported significantly higher levels of state anxiety prior to the digital final exam than men (*t*(9.727) = 2.283, p = 0.046, 95% CI [0.149, 14.529]. Our women achieved higher scores on the paper exam, than on the digital final, at a nearly significant rate (*t*(6) = 1.930, p = 0.102, 95% CI [-3.971, 33.588]). For men, there was no significant difference in exam scores across medium (*t*(55) = 0.843, p = 0.403, 95% [-2.836, 6.955]). Additionally, women were observed to outperform men on the paper-and-pencil midterm (*t*(6.982) = 2.317, p = 0.054, 95% CI [-0.241, 23.148]). However, there was no significant difference in performance between men and women on the digital final (*t*(7.067) = -0.111, p = 0.915, 95% CI [-18.838, 17.143]).

0.5 Discussion

We use our findings to demonstrate the importance of accounting for demographic factors in assessment of pedagogical success. We observed no effect of examination medium on students' performance, or anxiety at test time, measured within the aggregate student body. For many education researchers, this simple assessment would be sufficient. However, our results suggest evaluations that end here are incomplete, since effects may be observed only within subsetted demographic groups, like women and high-anxiety students. Many factors influenced our metrics of success, as shown in Table 1. Further, we were only able to observe the most interesting and informative effects of changing examination medium when measuring within demographic groups.

We were surprised to find that exam medium had no significant effect on state anxiety for our high-anxiety students, and that the digital final increased state anxiety for our low-anxiety students, as compared to the paper examination. Given the findings on tech anxiety discussed in the background of this paper, we expected digital examinations to increase state anxiety for high-anxiety students. This effect may be caused by the fact that the digital examination occurred later in the semester, and was cumulative. Students with low trait anxiety levels may be more confident entering the first examination than their high-anxiety counterparts. They may prepare less, and perform more poorly than they expected. Performing poorly on the midterm may lead to an increase in state anxiety in these low-anxiety students prior to the final exam. The opposite could apply for high-anxiety students, who may prepare more and score higher than they expected on the midterm, providing security when entering the final examination. Whether or not this speculative hypothesis is correct, we use this finding as evidence that pedagogical evaluations are nuanced, and many factors contribute to success.

Regardless, we are not primarily interested in identifying the relationship between trait anxiety and exam medium on state anxiety at test time. Rather, we use this observation to show pedagogical changes may differently affect high- and low-anxiety students. Trait anxiety, and its relationship to educational practices, may need to be studied as a demographic issue. High- and low-anxiety students are not commonly considered unique demographic groups, but studying the effects of trait anxiety requires a similar approach to studying engagement of underrepresented groups. Researchers must be aware that changes in educational practices may differently influence this group, and intentionally observe their response, to identify these effects. Again, this highlights the importance of understanding participants' demographics, when assessing pedagogical success.

We were not surprised to observe that women in our classroom performed better on traditional paper examinations, but we did not anticipate the magnitude of this difference. Our women's mean score on the paper exam was more than a full letter grade higher than our women's mean score on the digital exam, and our men's mean score on the same paper exam. Further, women in our sample experienced significantly higher levels of state anxiety than men prior to the digital final, but not prior to the paper-and-pencil midterm.

We are, again, not primarily interested in identifying the relationship between exam medium, gender, and student outcomes, or in making evaluative judgments regarding the merits of digital examinations. But, we believe these findings demonstrate the need to account for demographic factors influencing traditional success metrics. Had our analysis not considered demographics, and other underlying factors, our conclusion would likely have been that paper-and-pencil examinations are equal alternatives to digital exams. However, our results demonstrate this is likely not true, in all regards.

Educators' goals vary widely. No single pedagogical measure or assessment can meet the needs of every educator. By traditional measures of success, we found no evidence to either encourage or dissuade Computer Science educators from using digital examinations in the classroom. Some Computer Science educators may be satisfied with the success of their students as an aggregate, but many prioritize broadening participation of underrepresented students. For these educators, our results suggest digital examinations were not equal alternatives to paper examinations, as they may negatively impact women's performance. Though not all educators are primarily concerned with broadening participation of underrepresented students, Education researchers should explore the influence of demographic factors in all evaluations, so educators can be sure pedagogical changes meet their unique goals.

Moreover, research conducted regarding equal alternatives to traditional educational practices is possibly invalid, if its success is determined without considering effects observable only within subsetted demographic groups. We encourage educators to consider the underlying factors at work in their future pedagogical evaluation, both to broaden participation of underrepresented students in Computing, and to provide a more complete picture of the implemented changes' effects.

In the following section, we provide a path model identifying all measured predictors of examination performance, and state anxiety at test time, created with multiple linear regression. After mediation analysis, we found underlying factors, including demographics, were the only significant predictors of student outcomes, further demonstrating the potential for demographics to influence the conclusions of pedagogical assessments, and the need to account for these factors in evaluation.

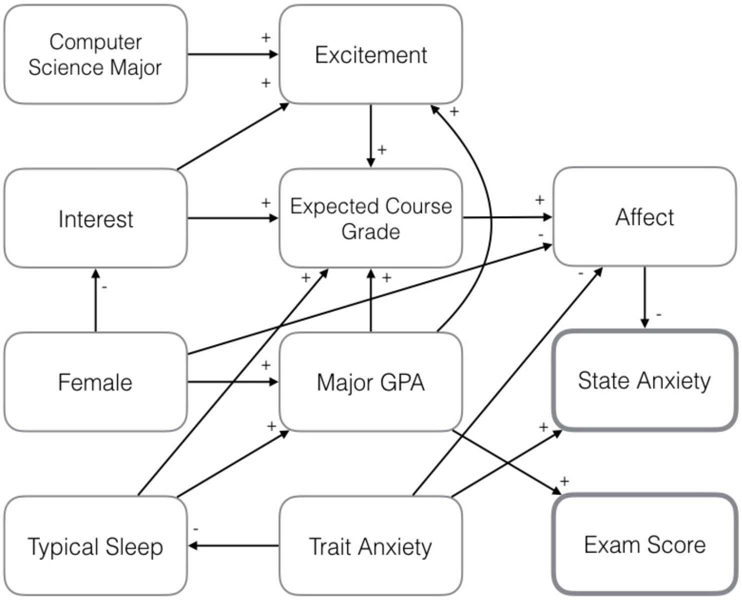

Figure 1: A path model of factors influencing our metrics of success, created through mediation analysis. All observed direct effects where $p < 0.05$ are shown. Causalities (indicated by directional arrows) within this model were drawn from our best, plausible inferences. Positive relationships are indicated by a +, and inverse relationships are indicated by a -. For example, being female was a significant predictor of lowered interest in course content, lowered affect at test time, and higher major GPAs. Mediation analysis showed that our tested treatment, exam medium, had no significant, observable effect on our outcome variables, state anxiety and exam score (shown in bold). However, we see there are many other factors predicting students' outcomes. While some of these factors, like gender, are more commonly measured, many are rarely included in evaluations. For instance, trait anxiety is not frequently measured in typical educational assessments, but here we see it is a significant predictor of our outcome variable, state anxiety at test time. Without considering the influence of these underlying, demographic factors, it is impossible to be certain that the conclusions of pedagogical evaluations are valid, and generalizable.

0.6 Path Model

We created a path model, shown in Figure 1, of all demographic factors observed to influence exam performance and state anxiety at test time with $p < 0.05$, to illustrate the complexity of identifying the real effect of pedagogical changes. This model shows the effects of factors used in our evaluation, on each other and on the outcome variables, identified with multiple linear regression and mediation analysis.

We do not use these observations to discuss the effect size and impact of any one factor on another, or on the outcome variables. Rather, we use this model as evidence of the variety of demographic factors, that are often unmeasured, that may impact student performance. This nuanced view of assessment should be taken in all future evaluations of educational systems, to provide a more complete picture of the effects of pedagogy.

Here, mediation analysis showed that our treatment (examination medium) had no direct or mediated effect on either performance or state anxiety at test time. Our outcome variables were only significantly influenced by demographic factors. Similar scenarios have likely occurred within existing pedagogical evaluations. Evaluations that do not account for demographics may inaccurately attribute success to pedagogy, where it is caused by the participants' demographic composition.

Moreover, we found demographic factors may influence outcomes, even when no observable direct effect exists. For example, in our classroom, Computer Science majors were not directly observed to show differences in state anxiety levels at test time, as compared to other majors. However, Computer Science majors reported higher levels of excitement about course content than other students. Students who were more excited about course content also expected higher overall course grades. Expecting high course grades positively influenced affect at test time, which, in turn, reduced anxiety. As such, Computer Science majors could be predicted to have lower anxiety levels at test time, than classmates of other majors. Major is not a widely considered demographic factor, particularly in upper level Computing courses, where students' majors are relatively homogenous. Here, major predicted different outcomes for Computer Science students, even varying from students with similar majors, like Computer Engineering.

Expectedly, our students with high trait anxiety were predicted to have higher state anxiety, and more negative affect at test time than low-anxiety students. High-anxiety students were also observed to sleep fewer hours per night, and students who slept more reported higher expected course grades and higher major GPAs. In turn, these factors affected anxiety at test time and exam performance.

Trait anxiety significantly influenced our students' outcomes. However, it is more difficult to measure than most other demographic information, and is not directly related to goals of broadening participation of underrepresented students. So, it is often left unmeasured in educational evaluations.

Here, we observed an even more complex relationship between gender and exam performance, than in our previous analysis. In our sample, major GPA was the only significant predictor of exam performance ($p = 9.15 \times 10^{-9}$, $R^2 = 0.235$), and female students were predicted to have significantly higher major GPAs than men ($p = 0.009$, $R^2 = 0.042$). So, this model suggests women will achieve higher examination scores than their male classmates. However, our prior analysis shows this was only true on the paper examination.

Further, despite being predicted to outperform men, women experienced higher levels of state anxiety at test time. Females' state anxiety at test time was mediated both by major GPA and interest in course content. Women's higher major GPAs should have reduced state anxiety levels at test time, by increasing expected course grade, and affect. However, women's state anxiety was increased by their lower interest levels, which reduce excitement, expected course grade, and affect. Cumulatively, women experienced higher levels of anxiety prior to testing than men. Rather, the positive effects of women's better past performance did not outweigh the effects of their lower interest.

We found the overall effect of these two paths to increase women's state anxiety levels, when expected course performance is introduced. Despite their higher predicted performance, women report lower expected overall course grades, at rates worth noting, although not statistically significant enough to draw definitive conclusions ($p = 0.126$, $R^2 = 0.009$). This aligns with prior research on feelings of self-efficacy and technological incompetence in women.

We cannot offer suggestions for aligning women's expectations with their real performance. However, we use the complexity of this observation to highlight the importance of taking a nuanced approach to evaluation. We were able to uncover this apparent conflict of women's anticipated and predicted performance, only by examining the mediated relationships of all our factors. We believe this merits further research, and cite this as an example that interesting and informative findings may be hidden by using simplistic pedagogical assessments.

Many evaluations do not consider the importance of demographic factors influencing their metrics of success. Degree major, trait anxiety, and gender, are only a few of the demographic factors in the classroom. We observed each to have real impacts on our measured outcomes. While our simplistic evaluation method, and our path model, both concluded examination medium did not impact student performance, our findings show our conclusions may have been different, if our participants' demographic makeup was altered. Evaluations not accounting for demographics cannot make generalizable conclusions, because observations depend on students' characteristics. Even when educational assessments consider these factors, recognizing their conclusions may be changed by their participants' demographics is important.

Identifying the complex structure of the demographic factors predicting success is critical to thorough future evaluations, to progressing our understanding of the real effects of pedagogical changes, and to furthering research in Computer Science Education as a whole.

0.7 Limitations

Before continuing, we wish to reiterate that the work represented in this paper is focused on exposing the potential for underlying factors, including demographics, to influence the conclusions of pedagogical evaluation. With this in mind, we recognize several limitations to our study.

Most importantly, we wish to acknowledge that a confound exists between time – midterm or final – and exam medium – paper or digital – in our study. While we attempted to mitigate the effects of this confound by examining only the group of students who completed all steps of this study, it would have been preferable to split our students into two groups, and mix the order in which students receive the digital and paper exam. Alternatively, this study could have been conducted across several classes of the same course, and each class provided a different order of treatments. Because our work does not attempt to evaluate the benefit or harm of digital examinations, and because it could have potentially been unfair to the students enrolled in this class to treat the two groups differently, we leave this issue to be solved in future work that speaks directly to the merits of digital exams.

Additionally, we recognize that our small sample of women reduces the power with which we are able to identify the effects of examination medium. Because of this small sample size, we discuss all effects observed with $p <0.1$. Moreover, we do our best to be explicit in not making claims about the benefit or harm of digital examination. Rather, we use these results only to show a single example of an evaluation in which considering demographics changes the evaluation's conclusions.

Additionally, we regret that our sample did not contain enough racial diversity to discuss the consideration of racial factors contributing to pedagogical evaluations in Computer Science classrooms. Similarly, we regret that the information gathered in the initial survey did not include socio-economic status, typical hours spent using a computer, whether the student's childhood home had a computer, and many other factors that may have influenced our conclusions. This work also contributed to a greater understanding of the underlying factors that influence student performance for the authors. In future work, we would be interested in collecting a much larger variety of demographic information.

0.8 Related Work

Evaluation methods in Computer Science Education research have been previously explored and critiqued. The focus of proposed improvements to scientific evaluations of educational practices vary. But, it is agreed rigorous pedagogical assessments are critical to the advancement of Computer Science Education as a scientific discipline.

Authors in [18] called for greater levels of experimental evaluation in Computer Science Education. This work reviewed 444 papers about CS1/CS2 accepted to SIGCSE, from 1984 to 2003. Of these, only 21% contained experimental evaluations, where the ”author made any attempt at assessing the ’treatment’ with some scientific analysis.” However, for this period, papers related to these introductory Computer Science courses only accounted for 25-30% of SIGCSE proceedings. Examining only papers within this category may inaccurately represent the number of SIGCSE publications that contain experimental evaluations.

Similarly, [9] reviewed Computer Science Education research, published at a number of venues, from 2000 to 2005. Of the 352 publications sampled, 123 contained behavioral, quantitative, or empirical research. But, 40% of these studies including human subjects reported only anecdotal evidence.

Authors in [2] sought to update the above work, and renew the call for improved evaluation methods in Computer Science Education. The authors found that for full paper publications in SIGCSE 2014 and 2015, the number of empirical evaluations had increased to over 70%. Here, empirical evaluation was defined as evidence provided from observation, based on the definition of evaluation in [12].

In this paper, we also show the need for more rigorous evaluation of educational practices. The above research indicates scientific evaluations in Computer Science Education are increasing, but it also demonstrates empiricism is not a strict requirement for publications in this area.

We agree with these authors, that assessment of pedagogy must continue to improve. Additionally, we assert accounting for demographic factors is a critical component in increasing empiricism in Computer Science Education research. However, to our knowledge, there is no existing literature review, or call to action, aimed at encouraging the collection and reporting of demographic information in pedagogical evaluations.

0.9 Conclusions

The research presented here shows evaluating the effects of pedagogical changes is neither simple nor direct. We have shown that the effects of educational systems may differently affect students of varying demographic groups. Further, these effects may be obscured by measuring only direct effects, or measuring only effects on the aggregate student group. Identifying the real effects of pedagogy requires explicit, intentional, observation of underlying factors, and the response of different demographic groups.

We use these findings to encourage researchers and educators to adopt a nuanced view of evaluation. Accounting for these factors in future evaluations is imperative to identifying the real impacts, and generalizability, of our educational practices, in addition to increasing participation of underrepresented students in Computer Science. We believe our results call for a fundamental shift in the evaluation practices of Computer Science Education researchers, that is critical to the progress of Computer Science Education as a discipline.

BIBLIOGRAPHY

[1] *SIGCSE '16: Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, New York, NY, USA, 2016. ACM.

[2] A. Al-Zubidy, J. C. Carver, S. Heckman, and M. Sherriff. A (updated) review of empiricism at the sigcse technical symposium. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 120–125, New York, NY, USA, 2016. ACM.

[3] J. C. Cassady and R. E. Johnson. Cognitive test anxiety and academic performance. *Contemporary educational psychology*, 27(2):270–295, 2002.

[4] S. Cassidy and P. Eachus. Developing the computer user self-efficacy (cuse) scale: Investigating the relationship between computer self-efficacy, gender and experience with computers. *Journal of Educational Computing Research*, 26(2):133–153, 2002.

[5] S. L. Chua, D.-T. Chen, and A. F. Wong. Computer anxiety and its correlates: a meta-analysis. *Computers in human behavior*, 15(5):609–623, 1999.

[6] A. Durndell and K. Thomson. Gender and computing: a decade of change? *Computers & Education*, 28(1):1–9, 1997.

[7] S. C. Koch, S. M. Müller, and M. Sieverding. Women and computers. effects of stereotype threat on attribution of failure. *Computers & Education*, 51(4):1795–1803, 2008.

[8] R. G. Paulman and K. J. Kennelly. Test anxiety and ineffective test taking: Different names, same construct? *Journal of Educational Psychology*, 76(2):279, 1984.

[9] J. Randolph, G. Julnes, E. Sutinen, and S. Lehman. A methodological review of computer science education research. *Journal of Information Technology Education*, 7(1):135–162, 2008.

[10] S. B. Sarason, K. S. Davidson, F. F. Lighthall, R. R. Waite, and B. K. Ruebush. Anxiety in elementary school children: A report of research. 1960.

[11] P. A. Schutz and H. A. Davis. Emotions and self-regulation during test taking. *Educational psychologist*, 35(4):243–256, 2000.

[12] M. Shaw. Writing good software engineering research papers: minitutorial. In *Proceedings of the 25th international conference on software engineering*, pages 726–736. IEEE Computer Society, 2003.

[13] M. Sieverding and S. C. Koch. (self-) evaluation of computer competence: How gender matters. *Computers & Education*, 52(3):696–701, 2009.

[14] S. M. Smith. Theoretical principles of context-dependent memory. *Theoretical aspects of memory*, 2:168–195, 1994.

[15] C. D. Spielberger. *State-Trait anxiety inventory*. Wiley Online Library, 2010.

[16] J. Todman. Gender differences in computer anxiety among university entrants since 1992. *Computers & Education*, 34(1):27–35, 2000.

[17] A. Uusen and M. Müürsepp. Gender differences in reading habits among boys and girls of basic school in estonia. *Procedia-Social and Behavioral Sciences*, 69:1795–1804, 2012.

[18] D. W. Valentine. Cs educational research: a meta-analysis of sigcse technical symposium proceedings. *ACM SIGCSE Bulletin*, 36(1):255–259, 2004.

[19] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[20] S. Zatz and L. Chassin. Cognitions of test-anxious children under naturalistic test-taking conditions. *Journal of Consulting and Clinical Psychology*, 53(3):393, 1985.