SYSTEMATIC REPLICATIONS AND STATISTICAL REPRODUCIBILITY OF EDUCATIONAL RESEARCH

BY

ROBERT WESLEY CRUES

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Carolyn J. Anderson, Chair
Assistant Professor Luc Paquette
Professor Michelle Perry
Professor ChengXiang Zhai

Abstract

Science is at a critical juncture: the findings of many studies are unable to be replicated and reproduced, while scientific output is growing exponentially and becoming easily accessible. The inability of findings to replicate has been particularly prominent in the field of psychology, where it has been estimated that less than half of findings are able to be replicated. A similar conclusion has been drawn about educational research. At the same time, thousands of papers are published making it increasingly difficult for researchers to know whether or not findings have replicated. This thesis addresses the replicability and reproducibility of educational research and proposes tools that could help researchers sift through large amounts of scholarly output. The first paper of this thesis differentiates between the ideas of replicability and reproducibility, and describes how educational researchers can design systematic replications and report the details needed to reproduce statistical analyses. The second paper examines the use of different text classifiers to extract details about the findings and contextual factors of published articles, where this information can be used by researchers to determine whether two papers are systematic replications of one another. The third paper develops text classifiers to identify the details needed to reproduce the statistical analyses in published papers. These three papers demonstrate there are many components needed to replicate and reproduce educational studies, and these details are sometimes easily identified by text classifiers.

Acknowledgements

First, I would like to acknowledge my advisor, Carolyn Anderson. Carolyn has guided me through the process of graduate school, entertained my ideas, and let me define my own path. She has spent so much time reading my work and providing feedback, and it has made all of it so much better. Additionally, my committee members, Luc Paquette, Michelle Perry, and ChengXiang Zhai, helped me to keep this project manageable, gave me ideas that significantly improved it, and provided invaluable advice about this thesis and my career.

In addition to Carolyn and my committee members, I've also had the privilege to work with the iLearn group during much of my time at Illinois. I learned a lot working with the group. I was able to analyze a lot of interesting data and present the results of these all over the world. It was great to have a set of others colleagues where I could talk through technical problems and learn about the psychological and educational theories to inform this work. Without iLearn, I would have missed out on many interesting research projects, and I definitely would not be as well-versed in data and text mining or technical writing without this experience.

I've also had the pleasure of interacting with a lot of other great students in QUERIES, Educational Psychology, and the College of Education. All of you–Jennifer T, Rebecca, Geneveive, Jeanne, Shereen, Martha, Ian, Jennifer M, and my "adopted cohort" (Allison, Arielle, Dawn, Jess, Kevin, Matt, Mai)–you have all been so helpful over the years I've been at Illinois. You made living in Champaign bearable!

Finally, I could not have completed this dissertation without the support of people outside of school. My parents and sister have always encouraged me to keep my eye towards the end, and have provided support in an innumerable number of ways. I could always call them and get a good laugh, and I often needed it! And, your skepticism about everything encouraged me to think about and do things for me and on my terms, not others. Lindsey, thanks for putting up with me! I cannot wait to see what the future holds!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The replicability crisis has created a major stir in the scientific community. While Ioannidis (2005) pointed out nearly 15 years ago that many published findings were implausible, the widespread realization that many published results in the social sciences came several years later. Pashler and Wagenmakers (2012) introduced a special issue of *Perspectives on Psychological Science* positing that there was a "crisis of confidence" in 2012. In 2015, the Open Science Collaboration published the *Science* article, "Estimating the Reproducibility of Psychological Science," where they claimed that less than half of studies published in several psychology journals were able to be replicated. Similar to psychology, education has also suffered from the non-replicable findings. Although no large-scale studies such as "Estimating the Reproducibility of Psychological Science" have been conducted to quantify the replicability of educational research, smaller scale studies have found that findings in educational research are not replicable. For example, Andres et al. (2017) considered 21 "rules" from eight studies that established behaviors predictive of success in massive open online courses (MOOCs). When testing whether these "rules" were generalizable to new courses, only nine held when using statistical significance as the rule for successful replication; however, of these nine, two were in the opposite direction. Others have estimated that the success rate is higher than found by Andres et al. (2017), but these are limited to a small set of studies (e.g., Andres et al. (2018)). Therefore, there is much ambiguity about whether the findings of many educational studies are able to be replicated.

Concurrently, scientific output continues to grow. Recent estimates suggest that scientific output (specifically, publications) is growing exponentially (Bornmann and Mutz, 2015). With scientific output growing at this fast rate, scientists might encounter significant difficulties when searching the literature to accumulate evidence to generate hypotheses or determine the credibility of published findings. In particular, scientists generally find that synthesizing information is a very time consuming task (Blake and Pratt, 2006), and increasing amounts of information make this task more difficult. Beyond the sheer quantity of publications, scientific information is increasingly becoming publicly available. The rise of open science movements (Spellman et al., 2017; van der Zee and Reich, 2018) encourages the sharing of experimental and observational data, and importantly, scientific publications. Likewise, funders of research, such as the National Institutes

of Health (National Institutes of Health, 2008), National Science Foundation (National Science Foundation, 2015), and the Institute of Education Sciences (Institute of Educational Sciences, 2016), have created rules to encourage or mandate the sharing of data and publications stemming from funded work. Additionally, some university systems, such as the University of California (Smith and Ventry, 2018), have put pressure on publishers to lower open-access fees and make publications freely available. The exponential growth of publications and recent shifts in public dissemination of research and data mean that scientists face an ever increasing deluge of information.

To help scientists with this deluge of information, several tools and approaches have been developed to assist scientists in sifting through mass quantities of information and hasten scientific discovery. These approaches have focused on ways to synthesize large quantities of information from multiple studies to draw new conclusions or tools to identify the most pertinent parts of scientific literature. The most prominent methods of synthesizing information from scholarly texts, meta-analysis and systematic reviews, afford scientists an overview of the scientific landscape, but have been criticized due to their lack of transparency and inconsistent reporting (Moher et al., 2007; Aytug et al., 2012) or drawing conclusions to irrelevant and impossible questions (Nelson et al., 2018). Notwithstanding these criticisms, systematic summarization of literature has led to scientific discoveries. Swanson (1988) found a link between magnesium deficiencies and migraines by manually and systematically searching vast amounts of scientific literature, and this relationship has since been validated clinically. Because scientists will encounter massive amounts of information during literature searches, automated tools have been developed to help scientists sort through information. These tools identify and extract details about study characteristics (Hansen et al., 2008; Boudin et al., 2010; Kiritchenko et al., 2010; Hsu et al., 2012; Groza et al., 2013) and about the findings of studies (Blake, 2010; Groza et al., 2013; Gabb et al., 2015), but have been primarily developed and tested in biomedical sciences. While these automated tools have been developed in response to the ever-growing amount of scientific output, they have not been explored as a tool to help scientists understand the replicability and reproducibility of scientific claims.

This dissertation explores the replicability and reproducibility of educational research. Social science and educational researchers need tools to provide succinct information about the large amount of published studies to help them gather evidence in support of particular findings and to judge the replicability of findings. To contribute to this effort, this dissertation reviews suggestions from social science and statistics to enhance the replicability and reproducibility of educational research, and proposes the use of tools that lay the groundwork for automated detection of published replication studies and the extraction of details needed to reproduce published statistical analyses. Together, these studies contribute to the ongoing conversation

to enhance the reliability and credibility of educational research.

1.1 Overview of Three Papers

The three papers in this dissertation leverage developments in mining biomedical texts and the revelations about scientific practices in the social sciences to generate methods to conduct and text mine journal articles. The first paper reviews literature from education, psychology, and statistics to draw distinctions between the ideas of *replicable* and *reproducible*, and identifies how educational researchers can conduct systematic replication studies and provide sufficient details so others can reproduce their statistical analyses. The second paper proposes the use of text classification strategies to identify the details needed to determine whether a finding has replicated, with attention towards understanding the context of the study. The third paper develops the use of text mining tools to extract sentences a researcher might need to reproduce the statistical analyses in a published paper.

Systematic Replications and Reproducing Statistical Findings in Educational Research

The first paper draws a distinction between the ideas of replication and reproduction of educational research. Once these definitions are operationalized within the context of educational research, I identify how educational researchers can design systematic replication studies. The key to designing systematic replications is that two studies are interested in the same hypothetical relationship, while attending to the contextual factors that could influence this relationship. In addition to outlining how to conduct systematic replications, I detail information needed to reproduce the statistical findings in published papers. This list begins with hypothesis formation and continues through data dissemination. I conclude this paper with a discussion of current scientific movements and the current barriers to optimizing the replicability and reproducibility of educational research.

Extracting Information from Articles to Help Scientists Detect Systematic Replications

The second paper proposes the use of text classifiers to identify information that scientists could use to determine whether two extant articles are systematic replications of one another. The classifiers explored in this paper identify sentences from published articles that communicate the findings and contextual factors in these articles. To determine the best classifier, experiments were conducted with various feature selection and classification strategies to optimize the harmonic mean of precision and recall, F_1. The classifiers were built and tested on 20 journal articles about what makes students successful in online classes. Out of

four classification tasks, the classifiers performed well for two of the classification tasks, but need further refinement for the other two. When classifier performance is higher, the information identified by these classifiers can be used to develop automated methods to detect systematic replications.

Reproducibility of Findings: The Details are in the Analyses

Finally, the third paper tests text classifiers that identify the details needed to reproduce statistical analyses in published papers, following from part of the list in the first paper. Some of these classifiers are a series of regular expressions, while others are statistical classifiers. The classifiers are built and tested on the set of 20 articles used in paper two. I found that the classifiers are often very accurate, but have somewhat low recall for some of the information needed to reproduce statistical analyses. Further refinement of the classifiers could provide scientists the information they need to reproduce statistical analyses to check the veracity of claims in published papers, given the data used to arrive at those findings.

1.2 Significance

The social science research community is at a critical juncture: the findings of studies are unable to be replicated and reproduced, and the quantity of studies that are published is growing exponentially. Current practices of replicating studies and disseminating statistical analyses are insufficient. Likewise, there are very few studies that employ text mining and natural language processing to help educational researchers with the deluge of information they face. This dissertation makes a step toward helping educational researchers test scientific claims and communicate their procedures so that others can readily replicate and reproduce their work. Furthermore, this dissertation develops tools that can help researchers with the deluge of information by developing tools that mine the text of journal articles to help researchers identify systematic replications and the details needed to reproduce the statistical analyses of published papers.

Chapter 2

Systematic Replications and Reproducing Statistical Findings in Educational Research

2.1 Introduction

Science has recently been faced with a startling revelation: the findings of studies are not able to be reproduced, let alone replicated. Headlines have claimed that over half of a set of published studies in psychology cannot be replicated (Open Science Collaboration, 2015) and that "most published research findings are false" (Ioannidis, 2005). Controversies stemming from fraudsters such as Diederik Stapel, who falsified data in social psychology experiments (Bhattacharjee, 2013), Anil Potti, who edited genomic data and was invesitgated by the Office of Research Integrity (Notice from Department of Health and Human Services, 2015), and websites such as `www.retractionwatch.com`, have generated public outcry over the deceitful choices of some scientists. While these examples span many scientific disciplines, the troublesome discovery that studies are not generalizable or repeatable has been of particular interest to the social science community.

Many of the headlines about non-replicability or reproducibility have been in psychology, but the field of education has not been immune to some of these controversies. In education, these have focused on the inability of findings to generalize and the lack of replication studies. Several studies have identified that claims in educational research are not generalizable (Malouf and Taymans, 2016; Andres et al., 2017, 2018), which is problematic if we seek to build a robust evidence base. Furthermore, it has been estimated that there are very few replication studies in education. It was hypothesized that only 0.13% of articles in top education journals were replication studies (Makel and Plucker, 2014). These are troublesome because the findings from educational research can have profound and potentially detrimental effects on students. New public programs and policy decisions are often the result of educational research; therefore, educational research should be robust and arrive at sound conclusions. To this end, the findings of educational research that result in transformative decisions about educational systems should be *replicable* and *reproducible*.

The ability to replicate and reproduce studies is paramount to science, as it is the way to test the validity and correctness of scientific findings (Schmidt, 2009) and develop scientific theories. The goal of this paper is to distinguish between *replication* and *reproduction*. To draw this distinction, I draw on literature from

various scientific disciplines and identify how these ideas can be operationalized in educational research. Once the distinction has been drawn between *replication* and *reproduction*, I propose recommendations to conduct replication studies and reproducible educational studies. The recommendations for replication studies focus on designing systematic replications, and the recommendations for reproducing studies focus on the details needed to reproduce statistical results. After delineating these recommendations, I conclude with a summary of concurrent proposed solutions to enhance the replicability and reproducibility of education research, as well as current barriers.

2.2 Replication versus Reproduction

Educational research exists in a complex environment. Because the research involves human subjects, there are many "inputs" and "outputs" that influence the research questions and outcomes of educational research. Berliner and Glass (2015) describe these complexities by stating that educational systems are comprised of

> "inputs we can't control (for instance, family wealth, parents' education, community support, and special needs of children); variables we can't easily identify or measure (such as competing school and district initiatives, classroom culture, peer influence, teacher beliefs, and principal leadership); and outputs we can neither predict nor easily measure (such as resilience, grit, practical intelligence, social intelligence, and creativity)" (p. 12).

These complexities make educational research different from the physical sciences, where the inputs and outputs of systems are more easily controlled and allows replication studies to be more straightforward, or psychology, where studies are amenable to experimentation.

Before identifying the specific differences between replicating and reproducing, it is important to note that these terms can have different definitions depending on the scientific discipline (Plesser, 2018). The general definitions used in this paper follow the same paradigm, discussed next, as others in the social sciences, including: Chhin et al. (2018), Stevens (2017), Zwaan et al. (2018), Coyne et al. (2016), and Schmidt (2009). However, it should be noted that replicating and reproducing are sometimes used synonymously (e.g., Open Science Collaboration (2015)), with little or no differentiation between these two related ideas, or even oppositely (e.g., Drummond (2009)).

To define *replicate* and *reproduce*, I draw on the "statistical" definitions proposed by Patil et al. (2016a). They define a *replicable claim* as, "given a population, hypothesis, experimental design, and analysis plan, you make an equivalent claim based on the results of the study" (p. 2), while they define the term *reproducible* as, "given a population, hypothesis, experimental design, experimenter, data, analysis plan, and code you

get the same parameter estimates in a new analysis" (p. 2). From these definitions, the major difference is the use of the phrases *equivalent claim* and *same parameter estimates*.

The notions of replicability and reproducibility advance the accumulation of knowledge by ensuring studies are generalizable and accurate. Studies should be able to be replicated, "to demonstrate that the same findings can be obtained in any other place by any other researcher [and] is conceived as an operationalization of objectivity" (Schmidt, 2009, p. 90). Simply, a scientific claim is replicable if it is observable across a wide spectrum of possible conditions without regard to a specific context. This implies that a replicable claim is generalizable across a wide set of potential cases, therefore, *equivalent* claims can be made across these different cases. While replicability focuses on the ability to make claims regardless of context, reproducibility focuses on single instances that allows us to make the same claims.

Reproducibility assumes that given a specific scenario with the exact same conditions and objects of study, we arrive at the exact same conclusion(s). In the context of this paper, we focus on the ability to reproduce the statistical findings of published studies. Because many studies suffer from errors in statistical reporting (Bakker and Wicherts, 2011; Nuijten et al., 2016; Aczel et al., 2018) and problematic practices with statistical and data analyses (Maxwell, 2004; Simmons et al., 2011; John et al., 2012; Gelman and Loken, 2014; Wang et al., 2018), specific details are required about how authors arrived at their conclusions so that others can reproduce the statistical findings of studies. Many papers fail to explain sufficiently how authors analyzed their data and arrived at their conclusions (Counsell and Harlow, 2017; Hardwicke et al., 2018), which is particularly problematic because authors can arrive at different conclusions even when answering the same question with the same data (Silberzahn et al., 2018). Therefore, reproducibility allows us to inspect the accuracy of statistical findings so that we are confident that the stated results are correct. It should be noted, however, that reproducibility does not imply replicability (Leek and Peng, 2015).

Beyond these definitions of replicable and reproducible used in this paper, others have developed "types" of replication and reproduction studies. These often take the forms of either "direct replications," where authors seek to mimic the original study as close as possible, and "conceptual replications," where there is some variation in the design and procedure of a study, but the overall goal is to learn whether an effect exists regardless of specific conditions (Schmidt, 2009). In addition to these often used ideas, others have proposed "close replications" (Brandt et al., 2014); "methods reproducibility", "results reproducibility", and "inferential reproducibility" (Goodman et al., 2016), all of which are variations on the theme of checking the viability and possible correctness of scientific claims. To this end, there have been debates regarding the superiority of one "type" of replication over another (see Simons (2014), Crandall and Sherman (2016), and Zwaan et al. (2018)). However it is often not fruitful to draw sharp dichotomies between types of

replications, especially in the behavioral sciences, because direct replications are very difficult to conduct in social settings (Lindsay and Ehrenberg, 1993; Schmidt, 2009; Tackett and McShane, 2018).

In particular, it is difficult and often infeasible for behavioral science researchers to replicate a study in the same way a scientist might replicate a chemical or physical reaction (Coyne et al., 2016). This is due to contextual factors that might influence humans (Berliner and Glass, 2015; Van Bavel et al., 2016; Coyne et al., 2016; Malmivaara, 2019). Because educational research involves humans, "it is not possible to run the same study twice" (Simonsohn, 2015, p. 567). Specifically, it is impossible for the same human to be in the exact same state at two different times. However, the findings of a study can be reproduced, as we only need the data collected and the procedure used to reanalyze the data. To understand these definitions in the context of education, consider the following examples.

2.2.1 Hypothetical Examples

Given these definitions and the complexities of conducting educational research, how can the ideas of replications and reproductions be operationalized? To make these definitions salient in the context of educational research, let us consider the hypothetical studies outlined in Tables 2.1 and 2.2. Suppose the goal of all of these studies is to identify empirically whether there is a relationship between student procrastination and course outcome, measured by student grades. At first glance, the the context of both studies in Table 2.1 seems similar, since both take place in college courses. However, the particular context and methods used to draw conclusions for these two studies are quite different. The population of interest is the same, college students taking introductory courses, but these courses are from different academic disciplines (i.e., accounting and chemistry) and might be quite different because they take place in different regions and different types of colleges (e.g., a large research university in Arizona versus a small liberal arts college in Massachusetts). Likewise, Study A uses a less-intrusive method of data collection (clickstream from an online course platform) and Study B used a survey to measure student behavior. Although there are many differences between these studies with respect to design, methods, and measurement, both studies arrive at a similar conclusion. Specifically, the (hypothetical) findings of both studies reveal that there is a statistical relationship between procrastination and course grades.

Now consider Studies C and D in Table 2.2. Studies C and D have the same goals as studies A and B in Table 2.1, but are more similar with respect to context, procedures, and measures. Specifically, the studies in Table 2.2 examine the relationship between procrastination and final grades in biology courses at two large, research universities. The measures of procrastination (count of how many seconds a student submits homework before the deadline) are the same for the two studies, as were the analytic techniques

Table 2.1: Two hypothetical research studies that consider whether student procrastination has an effect on success in introductory college courses.

	Study A	**Study B**
College Setting	Large research university in Arizona	Small, liberal arts college in Massachusetts
Course	Introductory Accounting course	Introductory Chemistry
Course format	Online	Face-to-face
Measure of procrastination	Count of how many hours a student submits homework before the deadline	A survey that inquires about student's study habits
Measure of success	Students average grade in the course, [0,100]%	Earning an A or B in the course versus earning a C, D, or F in the course
Analytic Technique	Beta regression	Logistic regression
Hypothetical Finding	Every hour prior to the deadline a student submits homework, her log-odds are higher for obtaining a higher grade in the course	Students who agreed that they frequently procrastinate had higher log-odds of earning a C, D, or F in the course

(decision trees) to arrive at the conclusion. The conclusions of studies C and D are slightly different in magnitude, however, the key result is the same: students who do not procrastinate (up to a certain point) earn higher grades than those students who do procrastinate. Studies C and D are more similar to each other than studies A and B: C and D use the same course, same measure of procrastination, and same analytic technique. Unlike studies A and B, there is clearly an element of replication between studies C and D, regardless of any temporal element (e.g., D came before C, therefore C replicates D).

Using our assumed definition of replication, these studies are replication studies and the finding replicated across the studies. Students who are proactive (versus procrastinating) generally perform better in the course. The claim is robust, regardless of the type of institution, subject or format of the course, how the independent variable (measure of procrastination) and dependent variable (measure of course outcome) were used. Whether the studies are quite similar (as is for studies C and D) or somewhat different (as for studies A and B), the four studies find that procrastination shares a statistical relationship with grade. If we wanted to reproduce the findings of any one of these studies, we would need the data used to arrive at these findings, and any steps the authors took to analyze the data and arrive at their finding. If we could verify the accuracy of the findings and the correctness of them using the same methods the authors used, then the we could say the studies are reproducible.

Now that the differences between replicating a study and reproducing the (statistical) findings of studies are defined, we need to consider, in the context of educational research, how researchers can design studies to maximize their replicability and reproducibility. In the next section, I describe how a replication study can be designed systematically within educational contexts.

2.3 Systematic Replications in Education

Broadly, a replication study is a scientific study that has similar characteristics to another study, where both studies seek to understand the same phenomenon or to answer a similar (or the same) question. There are many studies that consider the same or similar questions of interest, but a well-designed replication study conducts a follow-up study in a systematic fashion. By systematically varying certain conditions of a study, we can build a cumulative knowledge base. As Schmidt (2009) explains,

> "...a cumulative science should be built on its foundations in a systematic way. Adding a brick here and another brick there without much regard for the space between them may result in an unstable building with weak parts, leakages and unnecessary parts that will require a major effort later on to effect their removal" (p. 96).

Table 2.2: Two more hypothetical research studies that consider whether student procrastination has an effect on success in introductory college courses.

	Study C	**Study D**
College Setting	Large public, research university in California	Large public, research university in Michigan
Course	Introductory biology	Introductory biology
Course format	Online	Online
Measure of procrastination	Count of how many seconds a student submits homework before the homework deadline	Count of how many seconds a student submits homework before the homework deadline
Measure of success	Student earns an A versus not an A in the course	Student earns an A versus not an A in the course
Analytic Technique	Decision tree	Decision tree
Hypothetical Finding	Students who submit their assignments 150 minutes before the homework deadline are more likely to earn an A than students who submit homework less 150 minutes before the deadline	Students who submit their assignments 100 minutes before the homework deadline are more likely to earn an A than students who submit homework less 100 minutes before the deadline

Ideally, the goal of a replication study is to not just ensure that a finding is generalizable, but also, to develop a further understanding of the phenomenon of interest. To effectively design replication studies, we must consider a specific, hypothetical relationship of interest, and how or whether the location, subjects, or manipulations considered by the authors influence this relationship. By systematically varying one or some of where the studies take place, who participates in them, and any intentional or unintentional manipulations, researchers can carefully design systematic replication studies.

To guide the discussion of systematic replication study design, let us consider the flow chart in Figure 2.1, which describes a work flow of designing studies. In this figure, we assume that researchers use an underlying theory to inform a research hypothesis. This then assumes that the goal of many studies is to test whether a hypothesized relationship exists. This hypothesized relationship is the core concept that is under consideration for a study. With respect to determining whether one study replicates another, it is this relationship that we are most interested in understanding whether or not is replicated across different studies. To test the hypothesis, we assume some sort of analytic model and collect data that are observable (for example, treatment conditions, interventions, changes-over-time) and possibly related to latent ideas (for example, proficiency, intelligence, emotional well-being). The measured variables used to test the relationship might contain error. This error could be decomposed into two sources: systematic (bias) or random (sampling variance). Sources of this error might be the contexts of one particular study; namely, these are where the study took place, the subjects in the study, and any sort of manipulations the researchers performed (either by design or outside of the researchers' control, but could influence the findings).

To design a systematic replication study, the first element of design should be to define the relationship of interest. Often, these hypothesized relationships are ill-defined, latent, difficult to measure, and the theories used to describe them are imprecise (Gelman, 2015). Thus, a key step in designing a replication study is to consider carefully this hypothesized relationship, and whether there is a one-to-one relationship between the original study's hypothetical relationship and the replication study's hypothetical relationship. To establish whether a study is a replication of another study, the hypothesized relationships' behavior, magnitude, and direction, should be virtually the same across studies. Once the relationships are specified, there should be an explicit definition of how these relationships are actualized and tested by the studies. This would lead to a falsifiable statement that can be tested via the data collected. In the examples in Tables 2.1 and 2.2, the hypothetical relationship is that the more a student procrastinates, the worse their course outcome. Determining consistency between two studies' hypothesized relationship is a key step toward designing a systematic replication, as this is the core of the study.

Once congruence is established between the original study and replication study with respect to the

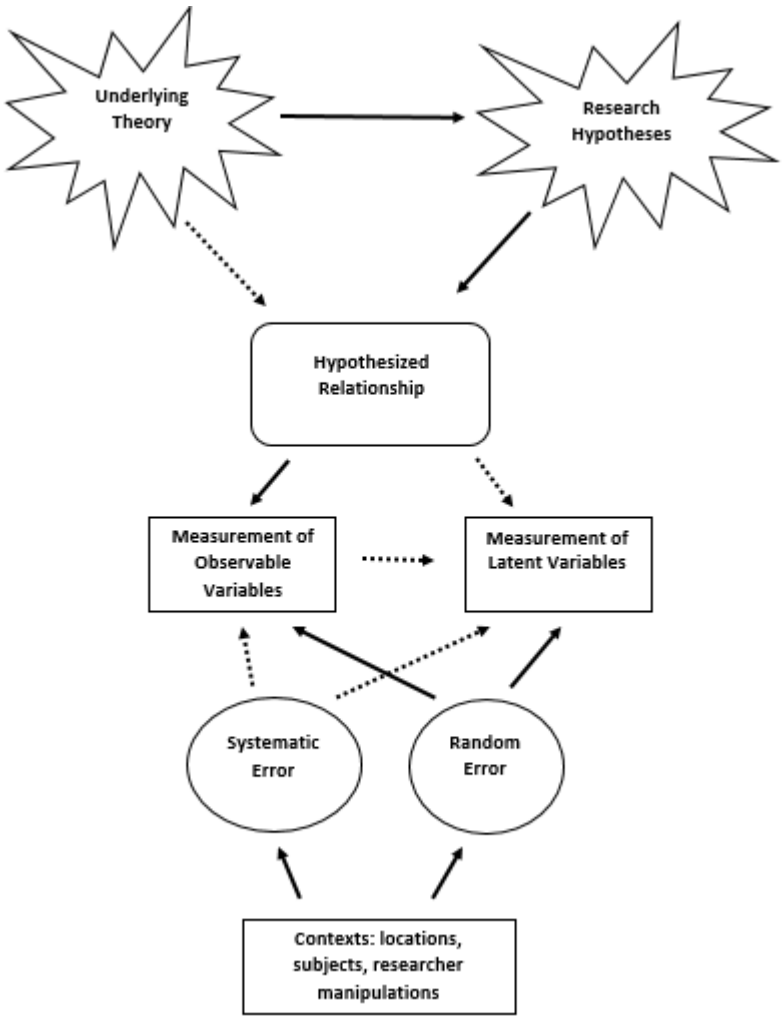

Figure 2.1: A hypothetical relationship is at the core of most studies, and this relationship can be informed by underlying theory or research hypotheses. The underlying theory might specify the behavior of the relationship (both its magnitude and relationship), or the theory might encourage hypothesis development that could define the actualization of this relationship. The actualization of the relationship is through observable and possibly latent variables. When measuring either type of variable, systematic or random error can enter into the measurement. The errors can follow from the various contexts of a study that influence the measurement of the observable and latent variables. Note that the dashed lines indicate that these relationships might not exist, whereas the solid lines indicate that these relationships are necessary.

relationship of interest, then it should be determined how the relationship will be measured via observable variables and whether latent variables are used. This relationship does not need to be one-to-one between the original and replication studies. However, in ideal cases, measurement of variables involved in the relationship will not introduce error beyond the contexts that might produce systematic or random error, which are discussed next. That is, measurement should not be biased (Flake et al., 2017). Further, researchers should be mindful that measurement of many human behaviors are context dependent (Flake et al., 2017; Gelman et al., 2018); therefore, researchers should communicate about how they measure the variables and carefully consider whether the measurement introduces error outside of the specific context of the study.

After the hypothetical relationships are explicated and specifications are made about how a set of observed and/or latent variables are related to each other, we must consider how the data are collected to test these relationships. Because educational systems co-exist among other social phenomena, we must consider how the context of the educational system could systematically or randomly introduce error into measurement of observable and/or latent variables. Specifically, we must consider how the location of studies, the subjects and participants, and any researcher manipulations could impact the findings of a study, also called "inputs" by Berliner and Glass (2015).

The location of a study can influence its outcome. For example, Van Bavel et al. (2016) suggests that the location of a study could impact its findings and replicability, after finding the studies considered in Open Science Collaboration (2015) were context-dependent. An example of this is the difference between conducting a study in a laboratory versus an authentic educational setting, where study participants might behave differently (Gelman et al., 2018). Furthermore, "cultural variability exists over time as well as space" (Greenfield, 2017, p. 767), especially in educational settings, where values and norms about education might be different depending on when and where a study takes place. To this end, geographical location and school climate are location-based sources of error that could influence the outcome of a study (Coyne et al., 2016). Therefore, a complete understanding of the location, and the implications it has on a study, are vital when designing a systematic replication. These implications should be carefully considered so that there is little, or a clearly identifiable, difference between the location of an original study and its systematic replication.

In addition to where a study takes place, we must also consider the role the subjects play when examining the hypothesized relationship. The particular background and characteristics of students and teachers in educational settings might influence the outcome of educational studies (Berliner and Glass, 2015). For example, in the field of medicine, knowledge about how patients were selected and their lifestyles (e.g., smoking habits, alcohol use, marital status) are vital to drawing inferences from studies (Malmivaara, 2019), and a similar analogy about student and teacher characteristics can be generated in the context of education.

An example in education might be to consider two classrooms of students, where one classroom has relatively high SES but another fairly low SES, as we would expect differences between these students. Although researchers cannot control student and teacher background, a clear understanding and description of the subjects of the study can guide the design of a replication study. If subjects are vastly different, this could influence the findings of the replication study; therefore, a clear description of the subjects and their characteristics can help us systematically design a replication study.

Finally, we must consider any manipulations, planned by the researcher or otherwise, that are different between the original and replication study. Particularly, manipulations capture variations between the original and replication study besides where the study took place or who participated in the studies. In general, these manipulations might be differences in stimuli, instructions, or procedures (Brandt et al., 2014). Within educational studies, these manipulations might be differences in content/skills, instructional design and delivery, and duration (Coyne et al., 2016). Other researcher manipulations might be deviations between the original and replication study that are unplanned, but should be documented, such as problems with data collection or modifications made on-the-fly. When designing a systematic replication, researchers should note how any manipulations they consider influence testing of the hypothesized relationship. For example, using different data collection strategies might mean different variables are observed, yet the relationship of interest is the same between the original and systematic replication. Accounting for these differences when designing a replication and considering their impact on the hypothesized relationship is paramount to designing a replication that systematically replicates another study.

Currently, there are no prescribed rules or theorems which state how much variation can exist between a replication study and an original study, whether or not the replication study intended to replicate another study. The difference between a replication and non-replication study is murky (Jamil, 2018). Generally, a study intended to be a replication should minimize the number of factors varied between the original and replication study (Schmidt, 2009; Coyne et al., 2016), in order to accumulate knowledge about a hypothesized relationship. The simplest solution is to sparingly vary meaningful factors beyond the natural variation that exists in educational contexts to design a systematic replication. By carefully and systematically varying single conditions, we can use the heterogeneity of irrelevancies (Shadish, 1995; Schmidt, 2009) to draw generalizations beyond one study by determining which factors do not introduce error or influence the relationship of interest. Thus, systematic differences between the original and replication study should be documented to allow us to understand whether a concept is generalizable and robust.

Once a replication study has been designed and completed, an intuitive next step would be to determine whether the replication was "successful." A criterion often used is whether statistical significance is achieved

in the same direction as the original study, such as in Open Science Collaboration (2015). More nuanced and exact ways of determining whether one study successfully replicates another have been proposed. Anderson and Maxwell (2016) proposed using effect sizes and confidence intervals of effect sizes to determine whether a successful replication occurs. However, statistical significance can result in meaningless claims (Simmons et al., 2011), as many studies are underpowered (Maxwell, 2004). Instead of using results of significance testing or the computations from one study, more robust conclusions can be drawn by understanding variability between and within studies (Patil et al., 2016b; Gelman, 2018a). Therefore, a pattern of results indicating success under many varied conditions is the most convincing evidence of an effect. A pattern of results, versus considering a single or few studies as evidence in support of an observed effect, can protect against random error.

If a well-designed systematic replication is deemed as "unsuccessful," then examination of the differences between the original and replication study can shed light on this finding, assuming the original study was not flawed or arrived at an erroneous conclusion. When only one meaningful factor is varied between studies, but a replication is deemed unsuccessful, this could give us insight about contextual influence on the tested relationship. As more factors are varied, determining why a finding did not replicate becomes more difficult, as there are more potential sources of error. At the start of this section, we noted that a replication study can give us further insight into a phenomenon. A replication failure is one way to do this. Failure to replicate gives us evidence of the generalizability of a finding, or that the original finding occurred via chance, and offers insight into factors that impact the finding's generalizability.

In sum, systematic replication studies allow us to gain insight in the robustness of a finding. A systematic replication requires a link between two studies on the basis of a hypothesized relationship that shares a one-to-one relationship with another study. Varying meaningful factors can allow us to understand the robustness of a finding (i.e., it generalizes more broadly) or limits of the finding (i.e., there are certain instances where it does not hold). In the next section, we turn to what authors need to include so that others may *reproduce* statistical and data analyses.

2.4 Details Needed to Reproduce Statistical Results

While designing a systematic replication study requires identification of a hypothetical relationship that is tested along with details about where it is tested, who it is tested on, and how it is tested, reproducing a study requires us to take a study, use its data, and arrive at the exact same results as the original authors stated. While all studies' procedures can be reproduced, I will focus on reproducing studies that use statistical

methods or data analytic techniques to arrive at conclusions and findings.

Reproducing a study allows us to evaluate the original study for accuracy and technical correctness. Ideally, independent researchers (i.e., those not affiliated with the original study) should be able to arrive at the exact same conclusions if they have the original data and procedures for analyzing the data. Being able to reproduce the statistical findings of a study allows outsiders the ability to check for problems such as *p*-hacking, where authors manipulate data to arrive at statistically significant conclusions (Simmons et al., 2011, 2012), whether analyses could lead to spurious findings due to the "garden of forking paths" (Gelman and Loken, 2014), or other questionable research practices (QRPs), such as excluding certain portions of data, underreporting certain findings, omitting corrections for multiple testing, or ignoring and violating statistical assumptions (John et al., 2012; Counsell and Harlow, 2017; Wang et al., 2018). Given the frequency of errors stemming from statistical analyses in the scientific literature, checking the accuracy and completeness of published findings should be a part of the scientific process.

To avoid problematic statistical findings in published articles, several approaches have been proposed to prevent them from appearing in the literature. For example, Lang and Altman (2013) and Schulz et al. (2010) have proposed guidelines for reporting statistical analyses and study designs in medical research, called the SAMPL guidelines and CONSORT, respectively. Additional reporting of study design and analysis in psychology was proposed by Simmons et al. (2012), who proposed the "21-word solution" to be included in manuscripts:

> "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (p. 4).

The "21-word solution" goes beyond the recommendations of the APA Task Force on Statistical Inference (Wilkinson and The Task Force on Statistical Inference, 1999) and the requirements in the APA publication manual. However, these recommendations are often prescriptive in the types of information authors should include or how their results should be presented, yet they do not protect against many of the problematic statistical practices that are likely causal factors to the "replicability crisis" (Pashler and Wagenmakers, 2012; Giner-Sorolla, 2012).

Further recommendations to improve data analyses to mitigate problematic practices have focused on the practice of sharing data and materials that are used to arrive at findings (Spellman et al., 2017; Cook et al., 2018; Martone et al., 2018; McBee et al., 2018; van der Zee and Reich, 2018). Some journals signal the availability of data for inspection and use by readers through badges, which has resulted in more data being made available (Kidwell et al., 2016). However, even when data are made publicly available, it can still be difficult or impossible to arrive at the exact statistical findings (Hardwicke et al., 2018). Likewise, many

qualified data analysts and statisticians can analyze the same data, yet use different methods and arrive at many different conclusions (Silberzahn et al., 2018), so it is imperative that authors provide the specific details of their statistical and data analyses so that the soundness and accuracy of claims can be verified.

To minimize the inability to reproduce the statistical findings to avoid problematic data analyses and their associated findings, the following pieces of information should be included in manuscripts or publicly available supplemental materials to allow others to reproduce the statistical analyses in published papers. I outline the needed information in the next subsections.

2.4.1 Study and Analysis Plan

1. **Falsifiable Hypothesis and Realizations of Tested Relationship:** A description of how the hypothesized relationship will be actualized (via an experiment) or derived (via observational data). This includes a complete description of how any observable and latent variables will be measured. For an experimental study, this should be a tool or scale that has been validated (Flake et al., 2017). For an observational study, this might consist of the specific variables collected (or specific function of variables) that are used to test this relationship. Along with this information, a falsifiable hypothesis in terms of the analytic technique to be employed to answer the research question should be specified. Ideally, the question is driven by an underlying theory that informs how the relationship is expected to behave.

2. **Sampling Strategy:** If the study is experimental, a complete description of the sampling strategy should be specified, with the sample size identified prior to data collection. If the collection of data is a convenience sample or otherwise not random, this should be specified. Likewise, a complete description of how data were collected when using existing data sets for an observational study should be specified; if this information is unavailable, then the lack of information should be noted. For either type of study, there should be a clear specification for how the data collected for the study generalizes to a population, if the goal is inference.

3. **Analytic Strategies:** Once the data have been collected, a complete description of the analytic strategies used to test the falsifiable hypothesis should be reported. This plan should have been specified prior to collecting data to avoid forking paths (Gelman and Loken, 2014) and p-hacking (Simmons et al., 2011); ideally, this plan should follow pre-registration protocols (Leek et al., 2017; Gehlbach and Robinson, 2018) and be made available. The specific analytic technique should be explicitly named.

4. **Transformation from Subject Data to Analysis Files:** After data collection, there should be traceable steps from subject responses to the raw data files, if the data were collected by the authors. For example, how were subject responses transcribed into a computerized format? Were the data collected via survey software or were responses entered by a human? When the data are analyzed from an existing data set, a description of where the data are stored, and any details relevant to how the raw data were collected and transformed should be acknowledged or referenced. For example, data sets from repositories or government agencies should be clearly referenced, so that it is apparent how data were transformed into computer files. At this stage, reference should be provided to any specific software, algorithm, or technique employed to transform raw data to a computer-based, analyzable format.

2.4.2 Transforming and Manipulating Data

1. **Data Preparation:** Once the data are transformed from raw data to an analyzable format, a description of any data processing should be provided. Specifically, it should be made clear how data were taken from a database or table and then transformed to a format usable for the specified data analysis software and analytic strategy. Additionally, if any data cleaning or transformations are required, a traceable script from the analyzable format to any transformations should be documented. If erroneous data is suspected, this should be noted, along with details on how this problem was remedied.

2. **Transformations:** Using the prepared data, there should be a summary of how the transformations done in the previous step impacted the data. Was any information lost in the transformations? For example, were continuous variables re-coded to discrete variables? If any data were removed due to quality problems, how does this change the summary information provided previously? Do changes in the summary information significantly change the original data? Further, were any of the transformations motivated by underlying theories?

3. **Descriptions of Variables:** At this stage, descriptions of all variables should be made available. These descriptions should include summary statistical information (e.g., measures of central tendency and variation) for quantitative data and a dictionary for qualitative data. Other data (e.g., free-form text) should be described to indicate what the data represent. In addition, a specification of any programs used to process data, along with documentation of all steps taken to process data should be reported.

2.4.3 Data Analysis and Reporting

1. **Data Analysis:** If null hypothesis significance testing (NHST) is used, then there should be evidence that the statistical claims are sufficiently supported by the observed data. Assumptions of any statistical testing or method should be verified, as this could impact the results of statistical tests (Hoekstra et al., 2012; Counsell and Harlow, 2017). Additionally, all statistical tests should be reported, not just those that are statistically significant (John et al., 2012; Wang et al., 2018), and statistical significance should be taken as only one piece of evidence about the findings (Wasserstein and Lazar, 2016; Leek et al., 2017). Further, reporting of interval estimates can give further insight into the magnitude and variability of an effect (Cumming, 2014). Including interval estimates along with raw descriptive statistics and claims of statistical significance allow for more transparent reporting. If other computational models are used, there should be explicit reasoning on why and how a certain criterion were used to arrive at the final model. These could be various error metrics (examples could include mean squared error, mean average error for quantitative data or precision and recall for qualitative data). Particularly, there should be clear sequential steps from how the analyzable data were transformed into a statistical test or computational model, and all details for redoing the analyses to arrive at the exact same parameter estimates should be provided (Stodden, 2015). This should include references to all software and settings used in the software, preferably with the exact code used to arrive at the findings. If algorithms are sensitive to starting values (or other random processes), there should be an indication that different values result in equivalent parameter estimates and findings, and the values used and tested should be denoted.

2. **Model Selection:** Once a final model or result is identified, researchers should describe how the model selected is superior to other statistical or computational models. Was the selection strategy automated? Were there out-of-sample validation strategies such as cross validation or training and test sets? There should be evidence that the results are accurate and remain consistent for the sample as a whole, as well as in pieces when aggregated (e.g., to avoid problems such as Simpson's paradox). This stage should be included in any analysis plans specified before collecting and analyzing data.

3. **Accuracy Checking:** After a final model is selected, there should also be checks to ensure the accuracy of the results. There should be checks to ensure data were properly read into software, data were pre-processed correctly, and that any statistical or computational models were correctly specified. Additionally, all interpretations and conclusions of any statistical or computational methods should be justified by the method and data. For example, causal language should not be used when the design

and methods yield only correlational or statistical relationships, and if NHST is used, results should be interpreted correctly (see Aczel et al. (2018) for discussion on errors in reporting NHST in psychology).

4. **Assessment of Alternative Approaches:** Even when defensible analytic choices are used to gain insight into data, results can be contradictory or vastly different (Silberzahn et al., 2018). Therefore, authors should acknowledge why their analytic procedures constitute a rigorous test of the specified hypothesis and how the results of the statistical analyses hold regardless of various analytic choices. It is not necessary (and likely impossible) to conduct every potential analysis of the data, however, some testing of different, defensible methods, should result in similar conclusions.

2.4.4 Storage and Accessibility of Data and Materials

Data & Code Availability: Along with including specific details to mitigate the risks of flexibility in data collection and analysis, authors should acknowledge where any data and statistical codes are stored, along with any technical details that are not relevant to the manuscript, but are needed to reproduce any analyses. Due to the sensitivity of educational data, it might be impractical or impossible to sufficiently anonymize and release data publicly (van der Zee and Reich, 2018). If this is the case, a few fictitious records could be generated to allow other researchers insight into how analyses were conducted and how they could reproduce the results on similar data. On the other hand, if the data are not sensitive or already publicly available, there should be a citation to the data's location. The raw data should be provided, so that any transformations or manipulations can be performed as done in the original analyses.

Providing all of these details, either in a manuscript or in online supplements, will aid in increasing the reproducibility of the statistical analyses reported from an investigation. While other lists have identified specific details to include (e.g., such as including R^2 when describing the fit of a general linear model or the exact p-values computed), the recommendations specified above focus on the full data analysis process, from planning the study to publishing and sharing the data and findings. Furthermore, providing a specific list of details is impractical for each possible method, as many new methods are developed (Epskamp, 2018); however, the intention of these recommendations is to encourage less "hand-waving" when analyzing data and providing concrete details about the analysis process and statistical findings.

From the recommendations on designing a replication study and reporting statistical analyses to maximize reproducibility, researchers are required to completely understand the context of their study and how the data were collected and analyzed. Importantly, the claims made from the study should be well supported by the data, where the techniques used to collect and analyze the data are well documented. Next, I discuss ways the recommendations from this review can be implemented into educational research.

2.5 Conducting Replicable and Reproducible Educational Research

The ability of findings to replicate and for results to be reproduced is vital to ensure the credibility and impact of educational research. To this end, there are several movements that are currently changing the landscape of educational research. Specifically, these movements are open science movements, pre-registration, registered reports, systematic replications, and complete reporting. However, there are currently barriers that could prevent these movements from gaining significant traction; these are the desire from researchers and funders for novelty, a dubious funding climate, and the uniqueness of educational research.

2.5.1 Proposed Solutions

There are several recommendations that have been proposed to enhance the replicability and reproducibility of educational research. These generally fall under the ideas of making educational research open, specifying and registering studies, and careful design and reporting.

"Open Educational Research": Part of the recommendations proposed to enhance the reproducibility of findings in this paper are to openly share data and research materials when publishing findings. These recommendations are part of the broader "open science" movement. Recent calls have focused on making educational research more open in certain areas such as gifted and special education (Cook et al., 2018; McBee et al., 2018). Broadly, open educational research can be implemented by opening up the design, data, analysis, and access to educational research (van der Zee and Reich, 2018).

Open science movements encourage researchers to be transparent about how they conducted their studies and arrived at their conclusions. Specifically, there have been calls to pre-register how a study will progress through data collection and analysis (McBee et al., 2018; Gehlbach and Robinson, 2018). The goal of pre-registering studies is to mitigate any risks for questionable research practices, while also promoting well designed studies; however, the complexity of educational research requires more careful consideration of how studies might be best pre-registered (Gehlbach and Robinson, 2018). Pre-registration is not an elixir, as Silberzahn et al. (2018) notes about the variability in conclusions drawn when teams of scientists analyzed the same data, "preregistration would not have prevented the observed variability in effect-size estimates across the teams in this study. Outcomes can vary as a result of different, defensible analytic decisions whether they are made post-hoc or a priori" (p. 16). Thus, opening up the design of a study allows outsiders to understand how a study was operationalized, which promotes reproducibility, but not necessarily replicability.

In addition to sharing how the study was operationalized by communicating the design, open science

movements also focus on the sharing of data and how the data are analyzed. Sharing data can be ethically challenging (discussed in the next section), but can help further scientific discovery. Specifically, data in either its raw form or as meta data, should be shared so that it can be found, interpreted, and analyzed (Martone et al., 2018). Likewise, sharing of data and how the data was analyzed can identify mistakes, which are prevalent in the literature (Bakker and Wicherts, 2011; Nuijten et al., 2016). Although awarding papers "badges" when they share data and materials has resulted in more open data (Kidwell et al., 2016), this does not necessarily make it easier to reproduce the findings of studies (Hardwicke et al., 2018). As such, the recommendations identified in this paper speak to how authors should report and disseminate their data analysis processes with the broader scientific community. Following these practices could mitigate questionable research practices and enhance the accuracy of published findings, while also sharing the burden of data collection amongst many researchers.

Finally, the open science movement calls for free and unencumbered access to scholarly findings. These could be either through pre-print servers or open access journals (van der Zee and Reich, 2018). Examples of pre-print servers for the social sciences include PsyArXiv, SocArXiv, and the Social Science Research Network (SSRN), where anyone with an internet connection can download scholarly papers. In addition to grassroot open science movements, policy makers have also encouraged open access publishing. Some US federal funders (e.g., National Institutes of Health) require public dissemination of work derived from certain funds (National Institutes of Health, 2008), and the University of California (which produces 10% of scientific publications in the US) is negotiating bulk open-access fees for affiliated scholars (Smith and Ventry, 2018). By allowing free access to published findings, practitioners and scholars without access to the expensive fees required by publishers can implement and build on the work of other scholars (van der Zee and Reich, 2018).

In sum, the open science movement can enhance the transparency of educational research. While it cannot directly influence the replicability of claims, many of these ideas promote reproducible science. The recommendations provided in this paper specifically identify ways educational researchers can communicate about portions of their studies towards these open science goals. In addition to providing specific and sufficient details to aid in the reproducibility of scientific claims, other approaches proposed in this paper and by others could test the replicability of claims.

Systematic Replication and Reporting: Open science practices are hypothesized to enhance the reproducibility of claims, but, being able to replicate scientific claims is no less important than being able to reproduce them. The recommendations identified in this paper suggest that systematic replication focus on testing a specific relationship of interest while remaining cognizant of how varying meaningful contextual

factors influence this relationship. Other proposals have focused on reporting effect sizes with confidence intervals (e.g., Cumming (2014)) and minimizing p-hacking (e.g., Wicherts et al. (2016)). These recommendations generally follow the traditional paradigm whereby authors can easily compute effect sizes or use NHST, but the recommendations provided in this paper follow a more general approach to research involving statistical data analysis.

To this end, systematic replication is needed to examine the veracity of scientific claims. The approach described in this paper offers a method to design replication studies in a systematic fashion. Ideally, scientists would use registered reports (Nosek and Lakens, 2014), where they can submit an introduction, literature review, and methods section for peer-review. So long as these portions of the paper pass stringent peer-review, the paper would be published regardless of whether the findings are negative or positive, avoiding "the file drawer" problem (Rosenthal, 1979) or the desire to report only positive results (Fanelli, 2010). With systematic replication and careful reporting of studies and their findings, educational researchers can determine the credibility of certain claims. However, there are current barriers to implementing these solutions, discussed next.

2.5.2 Barriers to Replication and Reproducing

Current barriers exist that do not overtly inhibit practices to enhance the replicability or reproducibility of research, but they do limit the potential of current movements and proposed solutions. Generally, these follow from the current research climate and the unique environment where educational research takes place.

Novelty: Replication research has often been considered uncreative, posing a high cost but low reward for researchers, and provides little status as an impactful publication, resulting in very few published replication studies (Madden et al., 1995; Makel et al., 2012; Makel and Plucker, 2014). Concurrent to the infrequent publication of replication studies is the exponential growth of publications (Bornmann and Mutz, 2015). While it cannot be claimed that there is a causal relationship between publication growth and the failure of many scientific claims to replicate, these could be correlated. To assuage the barrier to publish replication studies, some journals have encouraged publishing replications, but it is not a widespread practice (Martin and Clarke, 2017). Some journals in education, such as *Remedial and Special Education* and *Journal of Research in Mathematics Education* had special issues on replication in 2016 and 2018, respectively. However, overt replication is not mainstream in educational literature. Instead, there is still a desire to focus on telling a "good story" (Kerr, 1998; Simmons et al., 2011; Giner-Sorolla, 2012; Gelman and Loken, 2014; Gelman, 2015) versus confirming previously published findings and drawing reasonable conclusions from data.

Funding: Parallel to the problem of novelty as a requisite to publish (and implicitly, advance scientists'

careers), researchers must also secure funding to conduct research. Chhin et al. (2018) note that the Institute of Education Sciences "explicitly" supports replication research, but also that, "every dollar spent on a replication study is a dollar not spent on a new study" (p. 10). Thus, there is tension between support for replication research and the ability to secure funding to conduct such research, which could be a costly endeavor (Makel and Plucker, 2014). Without funding and support for educational research from federal agencies or private donors, widespread replication and developing a solid evidence base might not be obtainable.

Unique Challenges: Educational research exists in a social, contextual environment, which presents unique challenges. Different students, schools, and teachers bring many attributes that cannot be manipulated (Berliner and Glass, 2015). There are an innumerable number of contexts that could impact replication studies in educational settings (Coyne et al., 2016), and context could influence the ability of claims to replicate (Van Bavel et al., 2016).

Additionally, there are ethical concerns that must be considered when adopting these practices. Although published research should be approved through Institutional Review Boards (IRB), practices such as sharing data after collection should be carefully considered (Meyer, 2018). For sensitive data, such as educational data, even more stringent protocols to uphold rigorous standards of anonymity should be upheld (Joel et al., 2018; van der Zee and Reich, 2018), and researchers must abide by federal law. While other social sciences have ethical requirements to maintain participant confidentiality, educational researchers are bound to regulations such as the Family Educational Rights and Privacy Act (FERPA). Likewise, we must consider whether a study should be replicated, by minimizing the potential harm to subjects and providing useful knowledge (Zimbardo, 1973; Star, 2018). Furthermore, researchers should be keenly aware of how their moral and ethical duties play into replication studies, particularly when dealing with vulnerable populations (Howe and Moses, 1999).

2.6 Concluding Remarks

Replicating and reproducing educational research is an important endeavor to ensure the credibility and reliability of educational research. There have long been calls to address the lack of replication studies in educational research (Bauernfeind, 1968), as replication studies are needed to solidify our knowledge of educational phenomena (Star, 2018). Makel and Plucker (2014) summarize the importance of conducting replication studies by stating,

> "If education research is to be relied upon to develop sound policy and practice, then conducting

> replications on important findings is essential to moving toward a more reliable and trustworthy understanding of educational environments. Although potentially beneficial for the individual researcher, an overreliance on large effects from single studies drastically weakens the field as well as the likelihood of effective, evidence-based policy" (p. 313).

Thus, replication studies are needed so that policy makers can make sound decisions that impact millions of people. To this end, replication studies should be conducted systematically (Schmidt, 2009; Coyne et al., 2016).

Often, replication studies happen unsystematically in the social sciences (Hunt, 1975; Cook et al., 2016; Chhin et al., 2018), but they should be conducted methodically to build a robust base of knowledge. In this paper, I outlined a way of conducting replication studies in a systematic fashion. To conduct systematic replications, we need a one-to-one relationship between a prior study's hypothesized relationship of interest and the replication study's hypothesized relationship of interest. Beyond knowledge of the hypothesized relationship that is tested, we must be aware of the contextual factors of a study. To systematically conduct replications, the contextual factors should be carefully and minimally varied between the original and replication study (Schmidt, 2009; Coyne et al., 2016). Determining whether a relationship replicates requires a pattern of similar results across many different contexts. Beyond designing a replication study, this paper also discussed ways to make statistical and data analyses reproducible.

To reproduce a study, many details are needed about how the data were collected and analyzed. The recommendations about information to report and avoid questionable research practices identified in this paper speak to a wide range of methods, not just the often used NHST. To this end, these recommendations also speak to the open science movement, where scientists are encouraged to share their data, analysis, and materials for reuse by other scientists (Cook et al., 2018; Martone et al., 2018; McBee et al., 2018; van der Zee and Reich, 2018). Further, reproducible findings allow for error checking or reported analyses, which can bolster the credibility of published findings.

In sum, the ability to replicate and reproduce educational research is vital to build a strong evidence base of research about educational systems and their contexts. When findings are not reliable and incorrect, there could be wide implications about the usefulness and trustfulness of the field. Thus, scientists should carefully adopt ways to make their work reproducible, while also carefully designing studies to test the robustness of effects.

Chapter 3

Extracting Information from Articles to Help Scientists Detect Systematic Replications

3.1 Introduction

A pervasive problem in psychology and education has recently been identified: the findings of many published studies cannot be replicated (Open Science Collaboration, 2015; Andres et al., 2017, 2018), suggesting that there is a "crisis of confidence" in the social sciences (Pashler and Wagenmakers, 2012). Although scientists often have unobtainable expectations for replications (Patil et al., 2016b), the concern over the replicability of scientific claims cannot be ignored. In particular, it has been estimated that only 40% of findings from psychology are able to be replicated (Open Science Collaboration, 2015). Replication of results is paramount to scientific discovery, as it affords scientists evidence in support of particular hypotheses and allows for the accumulation of scientific knowledge (Schmidt, 2009).

Confounding the inability of findings to replicate from published studies is the deluge of information facing scientists. Science, and specifically, the way in which scientists communicate (i.e., publications) is growing exponentially (Bornmann and Mutz, 2015). This growth rate makes it difficult for scientists to remain abreast of recent results, and likely, to identify all of the relevant findings for their endeavors. While the number of scientific publications is growing exponentially, the number of replications published in journals is strikingly low. For example, Makel et al. (2012) found 1.07% of publications in psychology are replications, and in education, the rate is even lower at 0.13% (Makel and Plucker, 2014). There have long been calls to increase the frequency of replication in psychology and education (Bauernfeind, 1968; Smith, 1970; Berliner and Glass, 2015; Chhin et al., 2018), yet, explicit replications are infrequently published. Although explicit replications are infrequently published, it has been hypothesized that replications do occur frequently, albeit in disguise (Hunt, 1975; Schmidt, 2009). For example, Cook et al. (2016) found that 31% of studies from a set of articles they considered were replications, yet authors rarely identify their studies as replications. Likewise, Chhin et al. (2018) found that around half of funded projects between 2004 and 2016 by the Institute of Education Science (IES) were replication studies. Thus, it appears replications are occurring, but it is unclear whether these replications are systematic or unsystematic. To build robust knowledge,

replications should be systematic (Schmidt, 2009).

This paper proposes a text classification strategy that will provide scientists with succinct information from large quantities of text to help them determine whether systematic replications occur in the published literature. The method defined in this paper follows from the knowledge discovery for databases (KDD) process (Fayyad et al., 1996), where data are transformed and analyzed, with guidance toward interpretation. To test the viability of this approach, I used a set of 20 journal articles to build and test the classifiers. Specifically, I evaluate the classifiers using information retrieval metrics to gauge how well the classifiers identify sentences containing the information sought. In general, the goal of this approach is to provide a transparent way for scientists to extract information from published articles, and encourages them to compare studies to determine whether they are similar enough to be considered systematic replications. Importantly, it might help uncover whether certain findings are able to be replicated and build a robust base of knowledge. Before outlining the specifics of this strategy, I give a brief overview of the replication crisis and identify how scientists could use the extracted information to draw inferences from published articles.

3.2 The Replicability Crisis and Drawing Inferences From Studies

The inability to replicate the findings of studies has been attributed to many causes. Ioannidis (2005) boldly claimed that "most published research findings are false," and this has led to a keen focus on problematic practices with how scientists analyze data. The reliance on null hypothesis significance testing (NHST) has led scientists astray (Spellman et al., 2017), likely due to underpowered studies (Maxwell, 2004), and studies with noisy, small effects and small sample sizes (Gelman, 2018b). Likewise, flexibility in data analyses (Simmons et al., 2011; Gelman and Loken, 2014) and "questionable research practicies" (John et al., 2012; Wang et al., 2018) result in suspect statistical practices and (sometimes) ludicrious claims. To this end, it has been hypothesized that authors hypothesize after the results of a study are known (HARKing) (Kerr, 1998), report only positive or statistically significant findings (Fanelli, 2010, 2012; Pigott et al., 2013), all in an effort to meet "aesthetic benchmarks" (Giner-Sorolla, 2012, p. 556) that make scientific claims bolder than the evidence support. Therefore, there are many problematic practices that occur in scientific publications that could impact the replicability of scientific claims.

While many have examined potential causal factors that have led to the replication crisis, others have noted that replications are occurring in the literature, yet in disguise. It has been posited that around one-third to one-half of studies are (at least conceptual) replications of other studies (Cook et al., 2016; Chhin

et al., 2018), but these estimates do not consider whether the replications are systematic. For a replication to be systematic, the studies need to be quite similar, except for varying as few as possible conditions between the original and replication study (Schmidt, 2009; Brandt et al., 2014; Coyne et al., 2016). Particularly, contextual factors, such as who is being studied, where the study took place, and how the replicators collected data could impact the replicability of studies (Smith, 1970; Schmidt, 2009; Brandt et al., 2014; Berliner and Glass, 2015; Van Bavel et al., 2016; Gelman et al., 2018). Thus, these *conditions* of studies should be carefully examined to understand the replicability of a claim. Researchers often seek to draw generalizations from studies, regardless of the design (i.e., randomized, quasi-, or natural experiments), by taking a local instance where only certain facets are manipulated and draw conclusions and make generalities about larger concepts (Cook et al., 2002, p. 19-20). The goal of these studies, especially when employing statistical methods, is to take a sample from a population, then analyze the sample to make inferecnes about how a certain phenomenon behaves in the entire population. However, because of contextual factors or the *conditions* of a study, it is important to contextualize the findings.

One way to draw inferences between studies is to use the concept of the heterogeneity of irrelevancies (Shadish, 1995; Schmidt, 2009). When an underlying effect is tested on different subjects at different locations with various conditions, we can determine whether these variations share a relationship with the effect or whether they are irrelevant. Specifically, if a particular finding is supported regardless of where it is tested, who it is tested with, or how it is tested, this constitutes evidence of the replicability of the claim. However, determining whether one study successfully replicates another is not straightforward. The often used criterion of deeming success on the basis of statistical significance in the same direction is somewhat limited, so others have proposed examining whether effects are numerically similar (Anderson and Maxwell, 2016). Furthermore, others have proposed examining the success of replications on a continuum (Gelman, 2018a). Thus, the most convincing evidence of successful replication is when a pattern of results in the same direction with similar magnitudes and interpretations come from many studies.

Therefore, to determine whether studies are systematic replications of one another, we need the *conditions* of the studies and the *findings* of the studies. The specific conditions needed are the *location*, *subjects*, and *manipulations* in each study. Because of the large number of studies and their current rate of growth, we need a way to automate the task of identifying the conditions of studies and their findings. To do so, I propose and demonstrate the use of text mining tools.

3.3 Mining Academic Literature

Synthesizing information from scientific texts is an extremely time-consuming task for scientists (Blake and Pratt, 2006). When combining the findings from multiple studies, a large majority of the work is identifying relevant studies (Allen and Olkin, 1999). To help scientists sift through the large amounts of information published, several tools have been proposed to extract information from journal articles.

Many of the tools to help scientists parse information from the text of academic texts have focused on identifying "scientific artifacts" and the findings of studies. For example, Groza et al. (2013) used conditional random fields, support vector machines, and ensemble techniques to identify sentences containing information about the hypotheses, motivations, background, objectives, and findings of studies labeled using the CoreSC schema (Liakata et al., 2010). To identify sentences containing the information sought, they used locational and linguistic features of the sentences, but experienced mixed success on the basis of precision and recall. Others have focused on identifying the findings of studies and scientific claims. Blake (2010) proposed the Claim Framework, which examines the syntactic relationship between words and phrases in sentences to identify scientific claims from the full text of biomedical articles. As an alternative to syntactic structure, Gabb et al. (2015) used unigrams (Brown et al., 1992) and locational features with support vector machines and naïve Bayes classifiers to identify sentences containing the findings of toxicology articles. They found these classifiers had high accuracy and fairly high recall. For each of these approaches, the general goal is to take the entire text of journal articles and reduce them to the most pertinent parts to help scientists synthesize details and findings from published studies.

However, the aforementioned approaches have been explored on biomedical texts, which often follow a prescriptive structure (Purcell et al., 1997) and have defined relationships such as gene-protein or treatment-disease that are reported. These characteristics are inherently different from social science texts, where relationships are often not clearly defined (Gelman, 2015) and the structure of texts is not as homogeneous as they are in the physical and biomedical sciences (Sándor and Vorndran, 2009). The goal of this paper is to explore the feasibility of using text classification strategies to identify whether the findings and conditions of social science texts can be readily identified, following from the approaches explored in biomedicine. To explore whether the results or conditions of studies can be identified by text classifiers, I next discuss an experiment I conducted with 20 journal articles drawn from education research.

3.4 Method

3.4.1 Objective

Given that many studies likely test the same hypotheses (Hunt, 1975; Schmidt, 2009), even when certain conditions are varied, the goal is likely the same: make generalizations beyond one study's subjects to a larger population. Thus, systematic replications seek to understand the same (or very similar relationships), but might be in different locations, use different subjects, or use different manipulations or study conditions. If a pattern of related findings are found regardless of where the study took place, who was studied, or how the scientists collected and analyzed data, then we have evidence that the finding might be replicable and that the contextual factors might not influence these relationships.

To help scientists judge whether studies are systematic replications, the rest of this paper outlines a supervised text classification strategy that builds classifiers to identify sentences that could help scientists make these determinations. Specifically, classifiers are built to identify sentences describing

1. the findings of studies, or *results*,
2. who was studied, or the *subjects*,
3. where a study took place, or the *location*, and
4. how the authors conducted their measurements or performed any manipulations that could impact the generalizability or outcome of a study, which I label *manipulations*.

To test whether these details can be identified from the text of journal articles, I used a set of 20 journal articles from education that examine what makes college students successful in online learning environments. Four classifiers were built, one for each piece of information, to test the viability of this approach. To optimize performance of the classifiers, I tested various feature selection strategies and text classifiers and evaluated them on the basis of recall, precision, and F_1. In the following section, I describe the strategy used to identify studies and build the classifiers.

3.4.2 Identification and Selection of Studies

This study explored 20 articles that investigated what makes college students successful in online classes. The 20 articles used to build the classifiers follow from the citations of Cerezo et al. (2016). Once I examined all of these citations to determine whether they were relevant to the general theme, I examined the citations of the cited papers. Additionally, I considered papers that cited the citations of Cerezo et al. (2016). Further,

I considered two review articles, assuming they might contain relevant citations; these two review articles were Kauffman (2015) and Papamitsiou and Economides (2014).

This led to a collection of 65 candidate articles to be used for the rest of this paper. Because manual annotation of each sentence in the collection was required, a filter was used to pare down the number of articles to a manageable number for experimentation. To do this, I considered the educational discipline of the article and the year of publication. The oldest article considered was published in 2004, while the most recent was published in 2017.

Some of the candidate articles were about one educational discipline (e.g., a single course in business or accounting), while others were more broad, focusing on a set of students who were willing to answer a survey. I filtered the 65 articles to 20 by enacting the date restriction (later than 2004) and focusing on papers that were about students or classes in the arts, education, business, or multiple subjects. This affords an opportunity to be diverse within the discipline (education), in case different disciplines report findings or study conditions in different manners. The list of the 20 articles is reported in Appendix A.

3.4.3 Post-Study Selection Processing

All of the articles considered in this study were available through open-access journals or through standard library databases from the University of Illinois at Urbana-Champaign in Fall 2017 and Spring 2018. When retrieving these articles, all were in portable document format (PDF); however, to use automated methods to extract needed information, articles needed to be converted from PDF to plain text. The PDF files were in various formats, where some are two-column journal articles and others were not. Additionally, many of the articles had headers or footers. In either of these cases, automatic conversion from PDF to plain text can be problematic. Files were copy-and-pasted from PDF to plain text to avoid problematic transformations.

The next stage of processing the articles consisted of filtering out certain sections of the text. Journal articles generally follow a sequential format, and as Purcell et al. (1997) notes for a clinical research article, the following are key "contexts": title, authors, background, objective, methods, results, conclusions, limitations/biases, future work, acknowledgements/collaborators and references. These contexts are often headings of journal articles, which break the article into various sections. Although Purcell et al. (1997) specified these for clinical research articles, they apply to the domain explored in this study[1]. Many of these sections are required to form a coherent and complete journal article, but are not relevant to the study. As the goal of this investigation is to develop a method to detect systematic replications, the authors, back-

[1]This is not always the case for publications in the social sciences, particularly education (Sándor and Vorndran, 2009). However, for the empirical, quantitative studies considered in this study, the general contexts developed by Purcell et al. (1997) appeared to be quite prevalent.

Table 3.1: Distribution of sentences containing relevant information within the training set

Category	# of sentences with	# of sentences without	total
Result	543 (22.3%)	1892 (77.7%)	2435
Condition			
location/setting	31 (1.3%)	2404 (98.7%)	2435
subjects	157 (6.4%)	2278 (93.6%)	2435
manipulations	198 (8.1%)	2237 (91.9%)	2435

ground, acknowledgements/collabortors, and references are not relevant. When converting the articles from PDF to plain text, these sections were omitted.

Once articles were converted to plain text, I used the R pacakge `tokenizers` (Mullen and Selivanov, 2016) to detect sentence boundaries. Each article was split into sentences. Although authors communicate the study conditions and findings over multiple sentences, sentence level annotation was used for simplicity. For the rest of this investigation, I use sentences as the unit to be classified as containing the relevant information.

After articles were tokenized into sentences, each sentence was annotated. The sentences were annotated by denoting whether or not the sentence contained a *result* or finding, and whether or not the sentence contained information about the *conditions* of the study. A sentence was judged to contain a result or finding if it contained "factual statements about the outputs of an investigation" (Liakata et al., 2010, p. 2055). Likewise, a sentence was denoted as containing a condition of the study if it included "any operational aspect of the study, including the persons, times, settings, treatments, and observations used" (Shadish, 1995, p. 423). Sentences were labeled to denote if they contained information about who the *subjects* were for the study, the *location* or setting of a study, and any *manipulations* the authors performed. Sentences could receive multiple annotations, such that a sentence could be labeled as containing a finding and describing one of the conditions of a study or as containing multiple conditions.

Once sentences were annotated, I split the articles into training and test sets. The training set consisted of 80% of the articles (i.e., 16), and 20% (or, four) of the articles were for testing. Table 3.1 reports the number of sentences in the training set that have one of the targeted pieces of information. Because the sentences may be annotated into multiple categories, the sentences in each task are not mutually exclusive.

3.4.4 Text Representation and Feature Selection

The text classifiers in this study rely on where a sentence is located within an article (following from the section headings) and the words (specifically, unigrams) in sentences to detect whether a sentence contains a result or condition of the study. Before supervised methods can be used, the text must be transformed from

Table 3.2: Unique words in each training set when various pre-processing steps were applied.

Pre-Processing Steps	**# of Unique Terms**
No Stemming and Inclusion of Numerals	5448
No Stemming and Exclusion of Numerals	4767
Stemming and Exclusion of Numerals	3119
Stemming and Inclusion of Numerals	3800

Table 3.3: Two-way table used to estimate χ^2 statistics and mutual information.

Contents/Term	Has Information	Does not have information	Total
Term	A	B	# of sentences with term
No Term	C	D	# of sentences without term
Total	# of sentences with information	# of sentences w/o information	n

their form as sentences to a matrix representation. Specifically, I used a bag-of-words approach and construct the document-frequency matrix (DFM), which consists of the the sentences as rows and the columns as words. Element ij in this matrix is the frequency of word j in sentence i. To construct this matrix, I used the R package `quanteda` (Benoit, 2018). When constructing this matrix, all characters were transformed to lower case, and punctuation and URLs were removed from the sentences using the built-in features of `quanteda`. Additionally, English stopwords were removed using the list provided in `quanteda`. Table 3.2 shows the number of unique words for each classification task.

Because the classification task is supervised, and the DFMs are quite sparse, I used three supervised feature selection methods to select terms that separate sentences containing the targeted information from sentences that do not contain this information. To select which words to use, I used the same three strategies as Gabb et al. (2015): χ^2 statistics, mutual information, and information gain.

The χ^2 statistic is defined as (from Jiang (2012))

$$\chi^2(t) = \frac{nF(t)^2(p(t) - P)^2}{F(t)(1 - F(t))P(1 - P)}, \tag{3.1}$$

where n is the total number of sentences, $p(t)$ is the probability that word t appears in a sentence containing relevant information (i.e., a result or condition), P is the proportion of sentences containing the relevant information, and $F(t)$ is the proportion of sentences that contain word t. In practice, the χ^2 statistic can be estimated for each word t by creating a two-way table (as shown in Table 3.3) for each word t. Specifically, the χ^2 statistic can be estimated by computing

$$\chi^2(t) = \frac{n(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}, \tag{3.2}$$

where the quantities are from the table for word t. If $\chi^2 = 0$, then word t and the information being classified

do not share a relationship.

Information gain (IG) is defined as (from Yang and Pedersen (1997))

$$ig(t) = -\sum_{i=1}^{m} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{m} P(c_i|t) \log P(c_i|t) + P(\bar{t}) \sum_{i=1}^{m} P(c_i|\bar{t}) \log P(c_i|\bar{t}). \tag{3.3}$$

Note that this definition is general, such that there are m potential classification categories. In equation 3.3, $P(c_i)$ is the probability of category c in the collection of sentences to be classified, where there are $i = 1, ..., m$ possible categories, $P(t)$ is the probability of term t, and $P(\bar{t})$ is the probability the term is not present in the collection of items to be classified.

Finally, mutual information (MI) is defined (from Yang and Pedersen (1997)) as

$$mi(t, c) = \log \frac{P(t \wedge c)}{P(t)P(c)}. \tag{3.4}$$

However, MI can be estimated using the quantities from Table 3.3 by computing

$$mi(t, c) \approx \log \frac{A \cdot N}{(A + C)(A + B)}. \tag{3.5}$$

MI is equal to zero if the term and category are not related, as MI is a conditional probability.

I computed these three feature selection metrics for each of the four classification tasks. Words were then ordered from greatest to least for each selection strategy within each task, and then feature sets were created. Specifically, terms with the largest 50, 100, 150, and 200 scores were selected. For each feature selection strategy, I computed these metrics with and without stemming the terms and removing numerals.

3.4.5 Methods of Supervised Text Mining

Once word features were selected and the reduced DFMs were combined with the locational features (i.e., section headings, such as Introduction, Method, Result, Discussion, and Conclusion), I used five supervised classifiers to identify sentences containing results or the conditions of the studies. The five supervised classifiers were the support vector machine with a linear kernel (SVM), decision trees (DT), decision trees with bagging (BATR), decision trees with boosting (BOTR), and naïve bayes (NB). All of these classifiers were built in R using the following R packages: `e1071` (Meyer et al., 2018), `rpart` (Therneau and Atkinson, 2018), `ipred` (Peters and Hothorn, 2018), `fastAdaboost` (Chatterjee, 2016), and `e1071`, respectively. All defaults were used for the support vector machine, decision tree, bagging, and naïve bayes, and 10 (weak) classifiers for the boosted decision trees.

Table 3.4: General form of confusion matrix

Predicted (↓)/Actual (→)	Not in Target Class	In Target Class	Total Predicted
Not in Target Class	True Negative (TN)	False Negative (FN)	Predicted Negative (PN)
In Target Class	False Positive (FP)	True Positive (TP)	Predicted Positive (PP)
Total Actual	Actual Negative (AN)	Actual Positive (AP)	Size of Test Set (N)

3.4.6 Evaluation of Classifiers

The classifiers were built using the training set for each of the four classification tasks. To estimate the performance of the classifiers, I used three evaluation metrics when using the test set on these classifiers: precision, recall, and F_1. These metrics were calculated by creating the confusion matrix (as shown in Table 3.4) for each classifier using each feature selection strategy. From the quantities in Table 3.4, the evaluation metrics are computed as

$$\text{precision} = \frac{TP}{TP + FP}, \tag{3.6}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{3.7}$$

and,

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{3.8}$$

From these definitions, we can see that precision is the ratio of actual relevant sentences to those that are marked as being relevant, and recall is the quotient of sentences identified by the classifier as being relevant, divided by the total number of sentences actually containing the relevant information. F_1 is the harmonic mean between precision and recall.

Previous work (Gabb et al., 2015; O'Mara-Eves et al., 2015) has opined that recall ought to be maximized when screening potential information for inclusion in reviews of scientific literature. Maximizing recall will allow for a large proportion of the sentences containing results or conditions of the studies to be identified by the classifier, but it comes at the expense of the classifiers also identifying sentences that do not contain this information (i.e., thus resulting in lower precision). Because both are important for this task, we will consider F_1 as a global measure of performance.

Although accuracy can be considered a reasonable metric in some cases, it is not a reliable metric when the data are imbalanced (Manning et al., 2009). These data suffer from imbalance because most sentences do not contain the desired information (i.e., the results or conditions of a study). This could result in high accuracy even though it might miss many relevant sentences. As such, precision, recall, and F_1 are better suited for evaluating the classifiers used.

3.4.7 Summary of Experiments to Estimate Viability of Identifying Conditions and Results

To summarize, I built four classifiers to extract information that could help scientists determine whether studies are systematic replications of one another and whether the findings of these replications are successful replications. To determine which of the five classifiers optimize performance on the test set, I used recall, precision, and F_1. These classifiers were trained using unigrams from the sentences in the training set and where the sentence was located within each article (i.e., title, abstract, introduction, methods, results, and discussion/conclusion). The unigrams were selected using three supervised feature selection strategies (χ^2 statistics, mutual information, and information gain), where I considered the top 50, 100, 150, and 200 terms because more features might lead to overfitting of the classifiers to the training set (Gabb et al., 2015). Further, I considered whether including numbers as unigrams and stemming the terms optimized performance of the classifiers.

Following from these various considerations, I evaluated the performance of 480 classifiers for each of the four classification tasks. In total, 1920 total classifiers were evaluated to consider optimal performance on the test set, where the classifier with the largest F_1 was selected as the best performer. Figure 3.1 shows the process flow to identify sentences containing relevant information; this was implemented four times, once for each classification task.

3.5 Results

The main output of this study is to evaluate the performance of the four classifiers developed that identify sentences describing the conditions and findings of studies. The goal of the classifiers is to maximize precision and recall, generally, and F_1, particularly. If performance of the classifiers shows that many sentences containing this information are identified as such, then scientists can use this information to determine whether studies are systematic replications of one another and whether the findings replicate.

Beyond providing results of the classification task, I also present descriptive information about the features used in these classifiers: the location of the sentences and unigrams. Many studies that mine academic texts only use the title and abstract to determine whether an article is relevant or not, but I show that is insufficient for the tasks in this paper because authors present the conditions and results of their studies throughout their articles. Furthermore, I present the performance of the classifiers and which words were selected by the feature selection strategies and used in the classifiers.

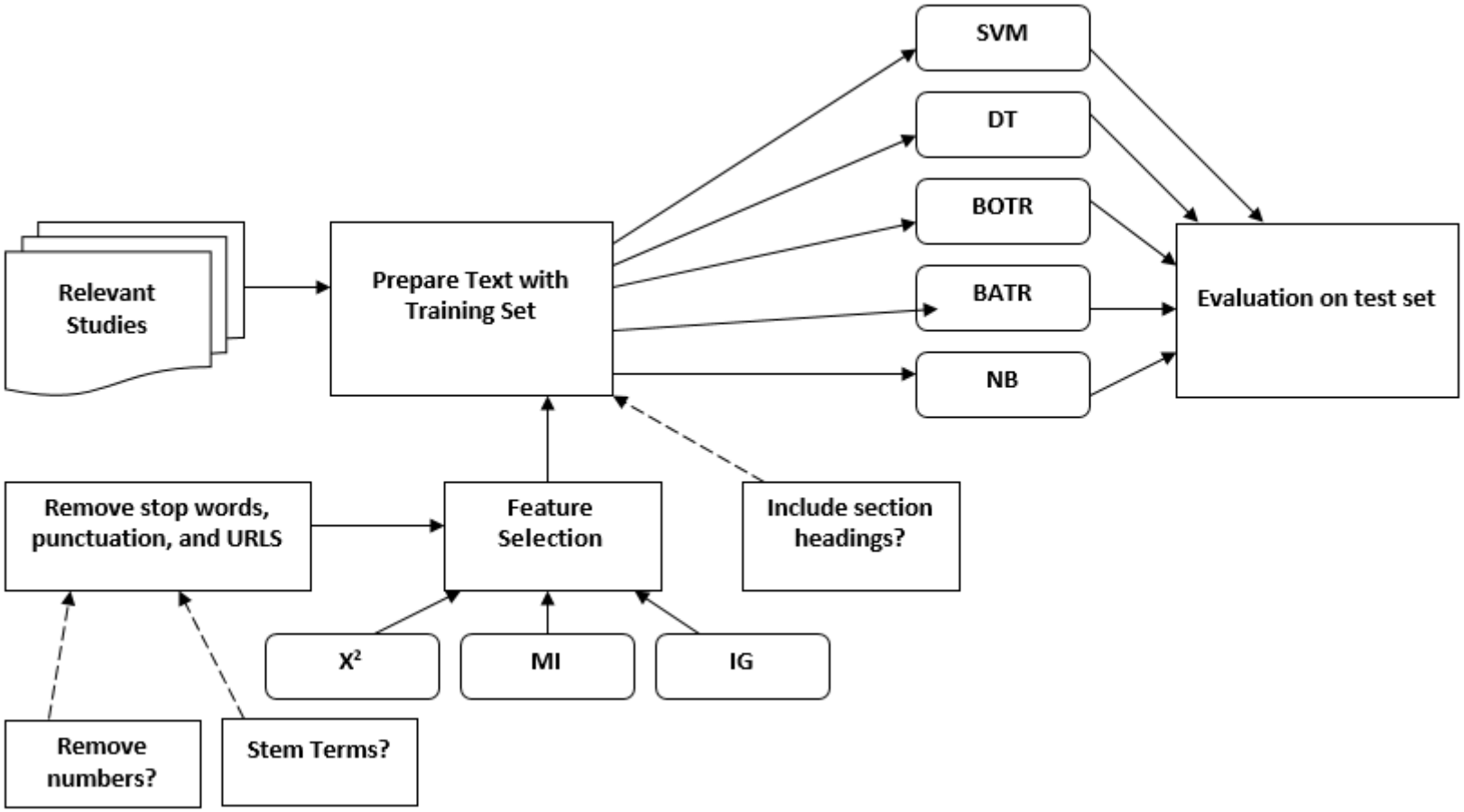

Figure 3.1: Process for identifying sentences containing relevant information from full-text journal articles. Note that SVM=support vector machine; DT=decision tree; BOTR=boosted decision tree; BATR=bagged decision tree; NB=naïve bayes; X^2=chi-squared statistics; MI=mutual information; IG=information gain

3.5.1 Where and how do authors present their findings and describe their studies?

In previous studies of mining academic texts, only the abstract of the text were considered (e.g., Matwin et al. (2010), Hansen et al. (2008)); however, ignoring the complete text of the article and considering only the title and abstract will not afford us an opportunity to thoroughly understand the outcomes and conditions of studies. This is evident from Figure 3.2. We see that authors do not limit themselves to just the result section of the papers when discussing the findings of their studies, but instead include results in the abstract, methods, results, and discussion/conclusion sections. By mining only the abstract or results section, we would miss many of sentences containing the findings of studies. This parallels the findings of investigations of biomedical texts: Gabb et al. (2015) and Blake (2010) found sentences containing results are sprinkled throughout the text of articles. Additionally, they found the ratio of sentences containing results to those not containing results is small.

Figure 3.3 shows where authors of articles describe where their study took place. Examples of descriptors of the location or setting where a study took place include:

- a mention of the name of an institution (e.g., "Rio Salado College" (p. 51) in Smith et al. (2012)), or
- a more vague description, as Holder (2007) writes, "a growing university in the Midwest" (p. 249).

Figure 3.3 shows for the entire collection of journal articles, very few sentences mention the location or setting of the studies. This is not surprising–authors might mention a detail about the location of where a study took place in the abstract or introduction, and again in the methods section, but will refrain from repeating this information. In this collection, the sentences containing details about the location of the study are in the title, methods, and results sections.

Unlike sentences containing the location and setting of studies, authors present descriptions of the subjects of their studies throughout the articles. Figure 3.4 shows that sentences containing information about the subjects of the studies were located throughout the articles; specifically, this information was present in the abstract, introduction, results, and discussion/conclusion sections of the articles. Examples of descriptors of sentences coded as containing information about subjects included:

- "Each section had similar student profiles in terms of the ratio of male to female students, representation of diverse ethnicities, average years of college study completed, and competency in using the computer and Internet as integral parts of their course work" (Kim et al., 2014, p. 155)
- "We recruited the participants through informal e-mail distribution lists and via a link placed on the

universities' home pages" (Klingsieck et al., 2012, p. 300)

Sentences describing manipulations in the studies are quite variable. They include information that might impact inferences drawn from the article, but are not obvious descriptors of where the study took place or who participated in the study. Two examples of sentences containing this information include:

- "The program is taught entirely online, and students are not required to ever visit campus in person" (Willging and Johnson, 2004, p. 110), and
- "Researchers utilized the data logging features of the Moodle learning management system and the Classroom Sense of Community Index" (Black et al., 2008, p. 65)

Most sentences containing descriptions of manipulations were located in the methods section of articles, as shown in Figure 3.5.

Along with a large variance in where and how authors present the findings of their studies and the descriptions of subjects, locations, and manipulations, there are many sentences that do not contain this information. From Table 3.1, it is evident that many sentences in the journal articles considered are not relevant to detecting replications. Likewise, authors present the descriptors of their studies and the findings of the studies throughout journal articles; therefore, the entire journal article must be considered if this information is given to scientists to help them determine whether studies are systematic replications or whether findings have been replicated. From Figure 3.6, we note that authors generally present the findings of their studies in two sections, the result and discussion/conclusion sections, and authors often present the conditions of their studies in the methods section. However, searching only sections where the majority of the information lies would result in missing some details that could aid scientists.

3.5.2 Which word features were used in these models?

In addition to including where a sentence was located within an article, the majority of features used in the classifiers were unigrams. Table 3.5 shows the the terms with the top 10 largest χ^2 scores for the results classifier, and the terms with the largest 10 information gain scores for the three classifiers that identify sentences containing the conditions of the study. These feature sets correspond to the classifiers that maximized F_1 in Table 3.6.

In Table 3.5, the highest-scoring words for the results classification task are not surprising. Authors often use inferential statistics to analyze their data and draw conclusions, and this is evident with terms such as *significant*, *p*, *01*, *f* (likely corresponding to the $\mathcal{F}$ distribution). Other terms in the list likely correspond to disseminating descriptive statistics (e.g., *variance* and *m*, the APA abbreviation for mean), while others are

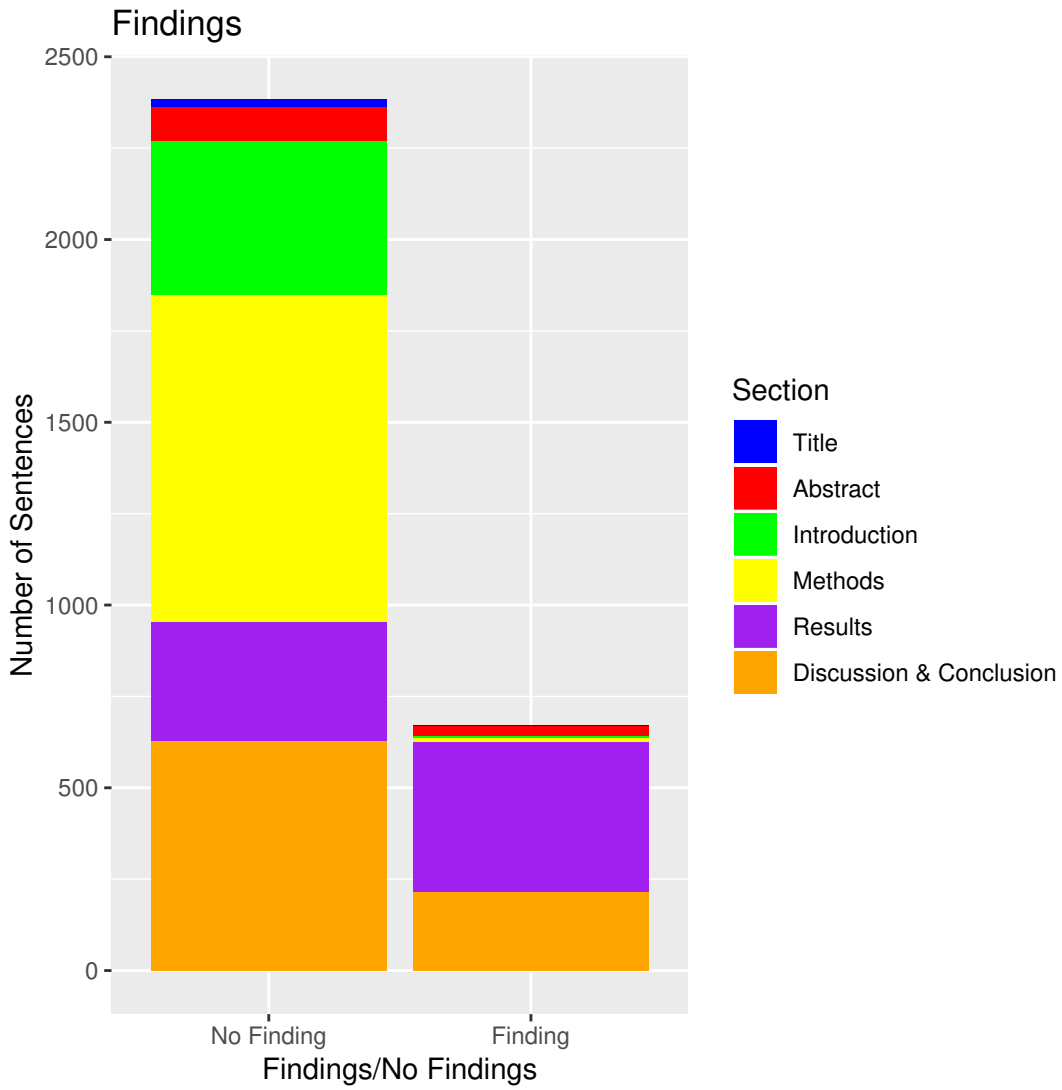

Figure 3.2: Where are the findings of studies in the 20 articles?

Table 3.5: Top 10 terms based on feature selection strategies when the classifiers in Table 3.6 maximized F_1.

Classification Task	**Top 10 Words**
Results	significant, p, 01, variance, predictor, f, showed, m, explained, revealed
Condition	
location/setting	university, undergraduate, college, master's, offers, located, illinois, online, korea, conducted
subjects	students, sd, average, accounting, participants, master's, taiwan, pursing, undergraduate, degree
manipulations	group, cours, session, board, signific, chat, entir, discuss, lectur, taught

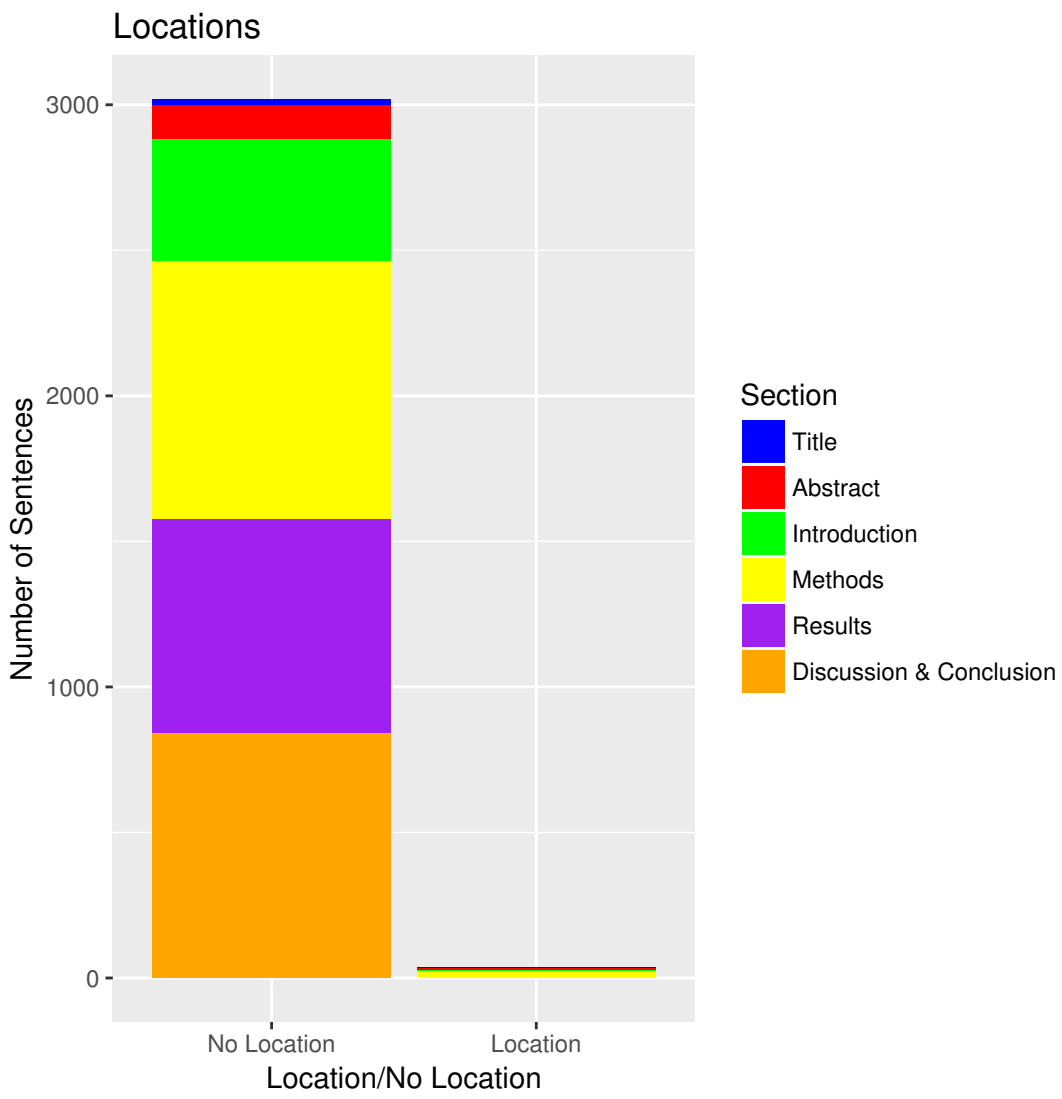

Figure 3.3: Where do authors discuss where the studies took place?

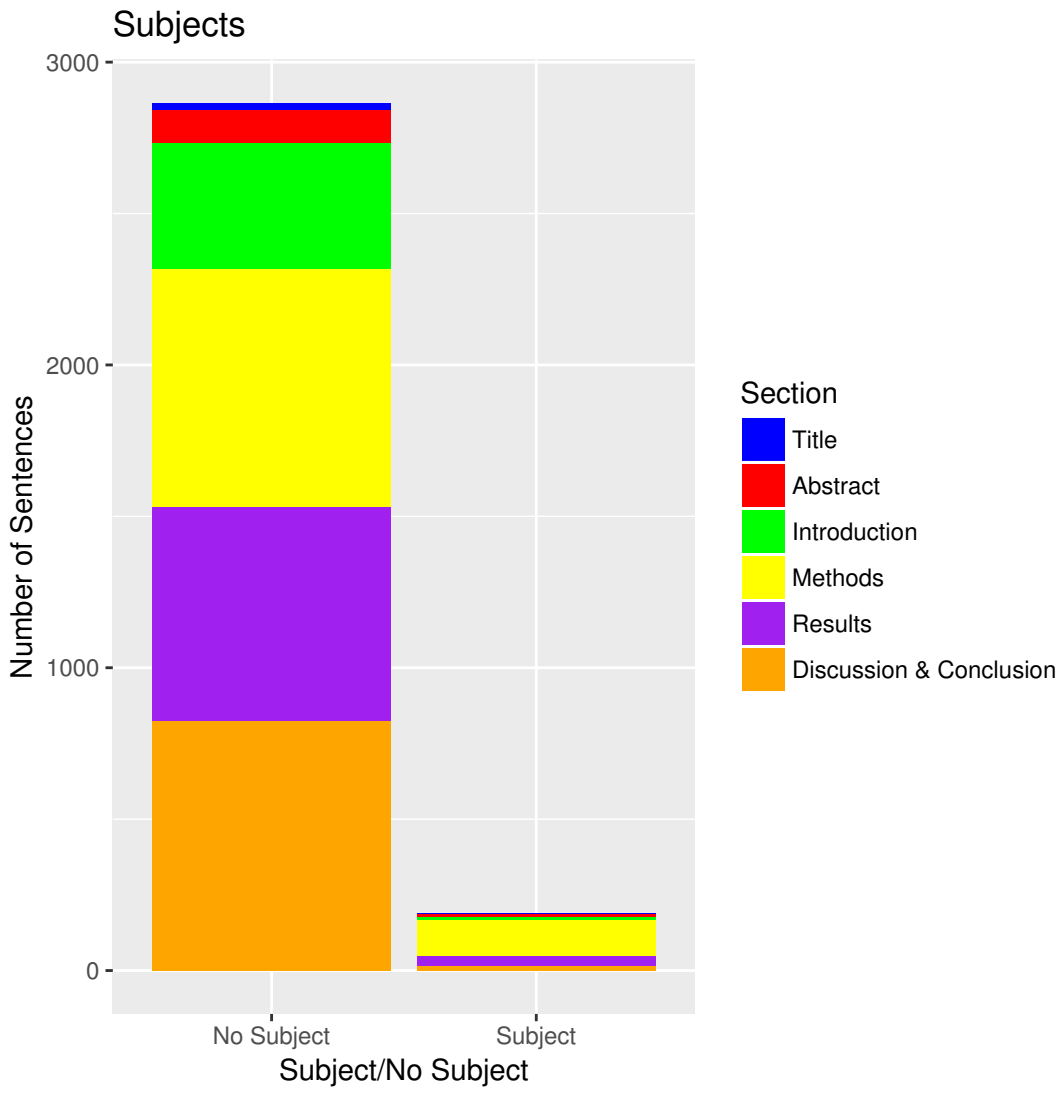

Figure 3.4: Where do authors discuss the subjects or participants of their studies?

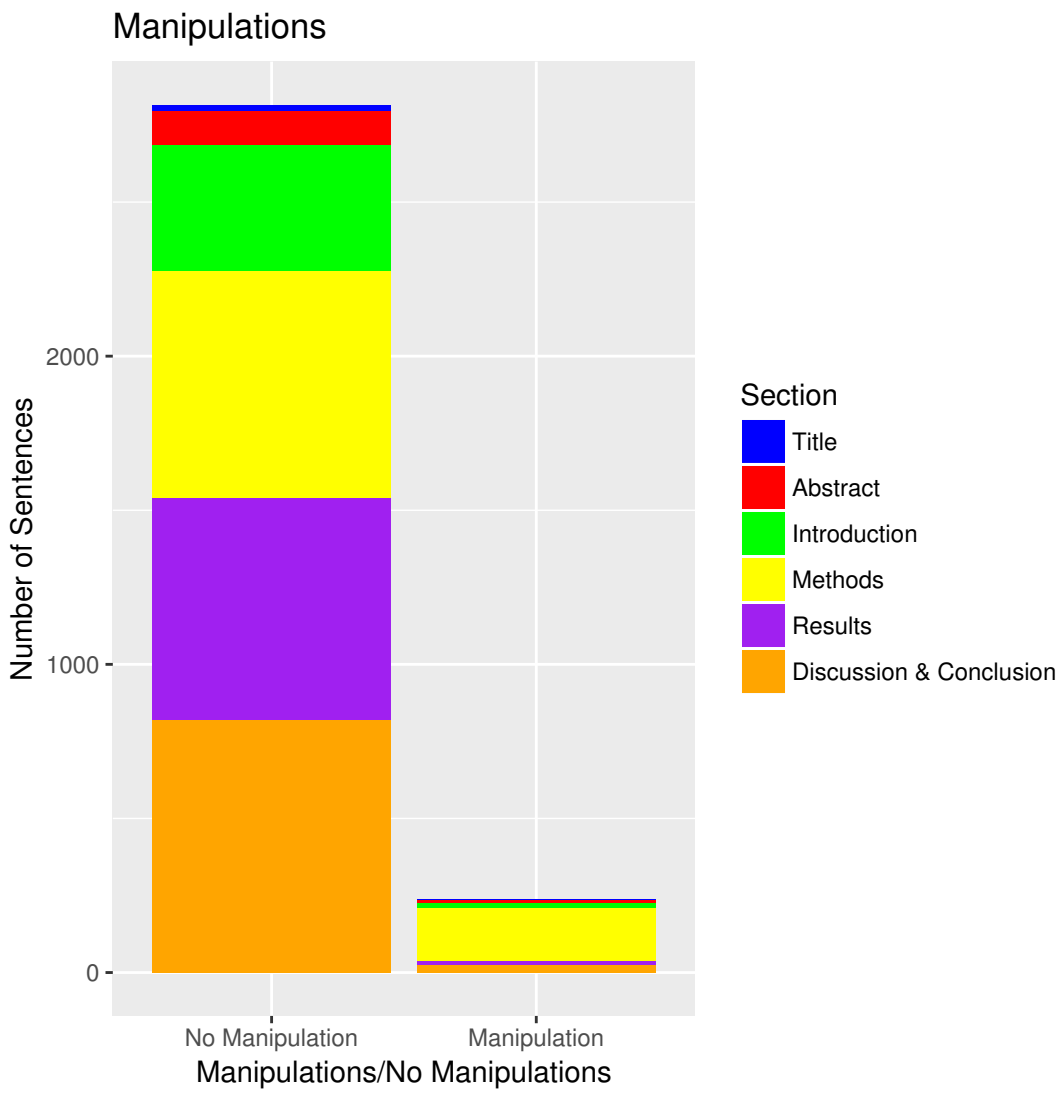

Figure 3.5: Where do authors describe their materials and other design aspects of their studies?

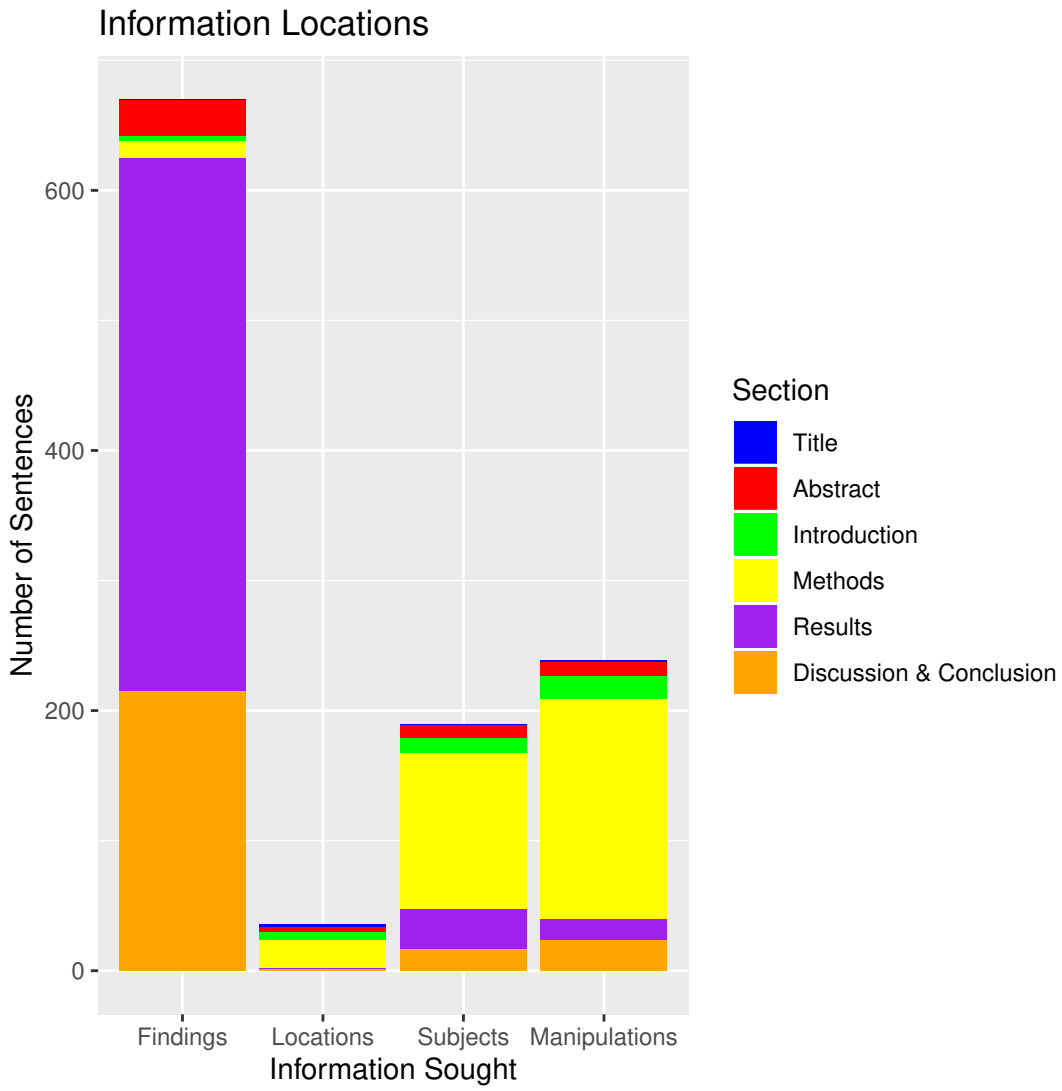

Figure 3.6: Where do authors present information that could help scientists determine systematic replications and whether findings replicate?

likely indicative of reporting the study's findings (e.g., *explained* and *revealed*). Furthermore, these terms appear to be general and not specific to the training set, as these terms could describe the findings of studies from many different fields, not just education.

Conversely, the terms used in the classifiers to describe the conditions of the study appear to be less generalizable across domains than the terms used to identify sentences containing results. Although these terms had high information gain scores for the training set, they are quite specific and likely not generalizable beyond this set of articles. The specific nature of some of the terms (e.g., *illinois* and *korea* for the location/setting classification task, or *accounting* and *taiwan* for the subjects task), suggests that the classifiers for these tasks might not perform well on articles not used to train the classifier. To determine the performance of the classifiers on unseen articles, I consider their performance on the test set.

3.5.3 Identifying sentences containing relevant information

So that scientists can determine whether studies are systematic replications of one another and whether the findings replicate, I built classifiers to detect sentences containing the conditions of and findings from studies. To select the best classifier, I used F_1 to equally balance between precision and recall.

Table 3.6 shows the results of the classification tasks to identify the sentences containing results and conditions of the studies. Within the table, we note the classifier that had the largest F_1 for each task, identify which feature set were used to build that classifier, how many words were used in the classifier, and the recall and precision of the classifier with the largest F_1.

When identifying sentences containing results, the best classifier based on maximizing F_1 is the SVM when using the χ^2 feature selection strategy. In addition to the top 200 words selected by χ^2 statistics, section headings were included, terms were not stemmed, and numerals were included. For this classifier, precision (0.905) was quite a bit higher than recall (0.598). While the performance for the results classifier appears reasonable, results for the classifiers that identify the conditions of the studies were mixed.

The classifiers used to identify the sentences containing the conditions of studies (the location, subjects, and manipulations) were strong in one case, when identifying sentences that discuss the location/setting, but were weak for classifiers that identify the subjects and manipulations of studies. The classifier for the location/setting had F_1, precision, and recall all equal to 0.8, which is quite high. On the other hand, F_1 was equal to 0.506 for the subjects classifier and 0.387 for the manipulations classifier. Recall was higher than precision for the subjects and manipulations classifiers. This suggests that more refinement is needed for these classifiers before they are put into large scale use.

In summary, I used various classifiers, feature selection strategies, and various transformations of the

feature set to maximize F_1. These experiments revealed that two classifiers, SVM and BOTR, two feature selection strategies, χ^2 and IG, and including locational details, maximzed performance of the classifiers. In addition, the classifiers selected generally performed best when there were more unigrams present, as we note that three of the four classifiers performed best when there were 200 words.

As expected, various text transformations and including other features besides unigrams often helped the classifiers discern between sentences that contain the information sought and those that did not. While authors describe the findings and conditions of their studies throughout the articles as is evident from Figure 3.6, authors generally report the findings of their studies in the results section and the conditions of their study in the methods section. With respect to unigram transformations, including numerals had mixed success with maximizing F_1. However, including numerals when classifying sentences containing findings might be useful because a numeral (01) was selected by χ^2 statistics (as seen in Table 3.5). Finally, using stemmed terms proved not to be as useful, likely because the test set had unstemmed terms.

3.6 Discussion

This paper tested the viability of using text classifiers to identify sentences about the findings and study conditions of articles in education. With this information, scientists could determine whether studies are systematic replications of one another, and further, whether the findings of studies replicate other findings. To determine whether studies are systematic replications, the conditions between two published studies should be quite similar (Schmidt, 2009; Brandt et al., 2014; Coyne et al., 2016). Currently, the literature lacks a metric to identify if two studies are similar enough to be considered replications; however, with information about the study conditions identified by the classifiers, scientists and subject matter experts could determine how "close" two studies are and whether they could be considered systematic replications. Likewise, the heterogeneity of irrelevancies can be used to determine which study conditions impact the studied relationship or effect (Shadish, 1995; Schmidt, 2009). Finally, a pattern of results with similar magnitudes and directions should be used to determine whether a finding is replicable. Beyond the implications for the usefulness of the information identified by the classifiers, the results of this paper shed light on how one could mine articles from the social sciences.

3.6.1 Implications for Text Mining Academic Literature

The results of this study revealed that the complete text of journal articles are needed to effectively identify the conditions and findings of studies. Prior work has focused mostly on the titles and abstracts of studies to

Table 3.6: Largest F_1 metrics for each classification task and associated classifier. "Sections?" corresponds to whether or not section headings were used as features in the classifiers; "Stemmed" refers to whether or not terms were stemmed when building the classifier; "Numerals" refers to whether numerals were removed from the collection. ✓denotes that section headings were included, or terms were stemmed, or numerals were included. X means the opposite of ✓.

Category	**Classifier**	**Sections?**	**Stemmed?**	**Numerals?**	**Feat. Selection**	**# words**	F_1	**Recall**	**Precision**
Results	SVM	✓	X	✓	χ^2	200	0.720	0.598	0.905
Condition									
location/setting	SVM	✓	X	X	IG	200	0.8	0.8	0.8
subjects	BOTR	✓	X	X	IG	50	0.506	0.69	0.4
maniuplations	BOTR	X	✓	✓	IG	200	0.387	0.561	0.295

screen for consideration of systematic reviews (O'Mara-Eves et al., 2015); however, only using the title and abstract for the particular use in this paper would be remiss because authors describe the conditions and findings of their studies throughout journal articles, as is evident in Figure 3.6. While authors often describe the conditions of their studies in the methods section of their articles and the findings of their studies in the results section, this is not always the case. This confirms the findings of Blake (2010) and Gabb et al. (2015), who mined biomedical texts, yet found the entire journal article is needed to effectively glean insight into the findings of published articles.

Beyond the finding that the full-text of articles is needed for this use case, this study also found that certain classifiers and feature selection strategies were superior with respect to maximizing F_1. Linear classifiers, such as SVMs, have proven to be successful for many text classification tasks (Forman, 2003; Aggarwal and Zhai, 2012; Groza et al., 2013; Gabb et al., 2015), and this study supports these findings because SVMs were the best performing for two of the classification tasks. Furthermore, ensemble techniques work particularly well when the data are significantly imbalanced (i.e., more training cases in one particular class versus another) (Guo et al., 2008), and the results of this study provides evidence in support of this conclusion. Particularly, the classifiers that identified sentences about the subjects and manipulations of studies performed best when the BOTR was used. Beyond the classifiers themselves, this study found that IG and χ^2 statistics were the best performing feature selection strategies for this particular use case. In general, experiments have shown that IG and χ^2 statistics work better than MI (Yang and Pedersen, 1997; Forman, 2003; Gabb et al., 2015), and the results of this study build evidence towards this conclusion. In addition to the unigrams, this study found that giving the classifiers more words (i.e., 200 vs 50, 100, or 150) resulted in better performance, which suggests that future iterations should consider a wider set of unigrams versus a smaller set.

Finally, the choice of performance metric for this study was not arbitrary. I chose to maximize F_1. Ideally, this would maximize both recall (i.e., identifying all relevant sentences that contain findings or conditions of sudies) and precision (i.e., only those sentences). However, maximizing F_1 results in a trade-off, as exhibited by the precision-recall curve (PRC) in Figure 3.7. Although this is the PRC for one of the classifiers built in this study, the others have a similar shape. From this curve, we note that if the classifier had perfect recall, then precision would be less than 0.25. This means the classifiers would identify many false positives, which goes against the main goal of this approach–to give scientists only succinct information about the conditions and findings of studies.

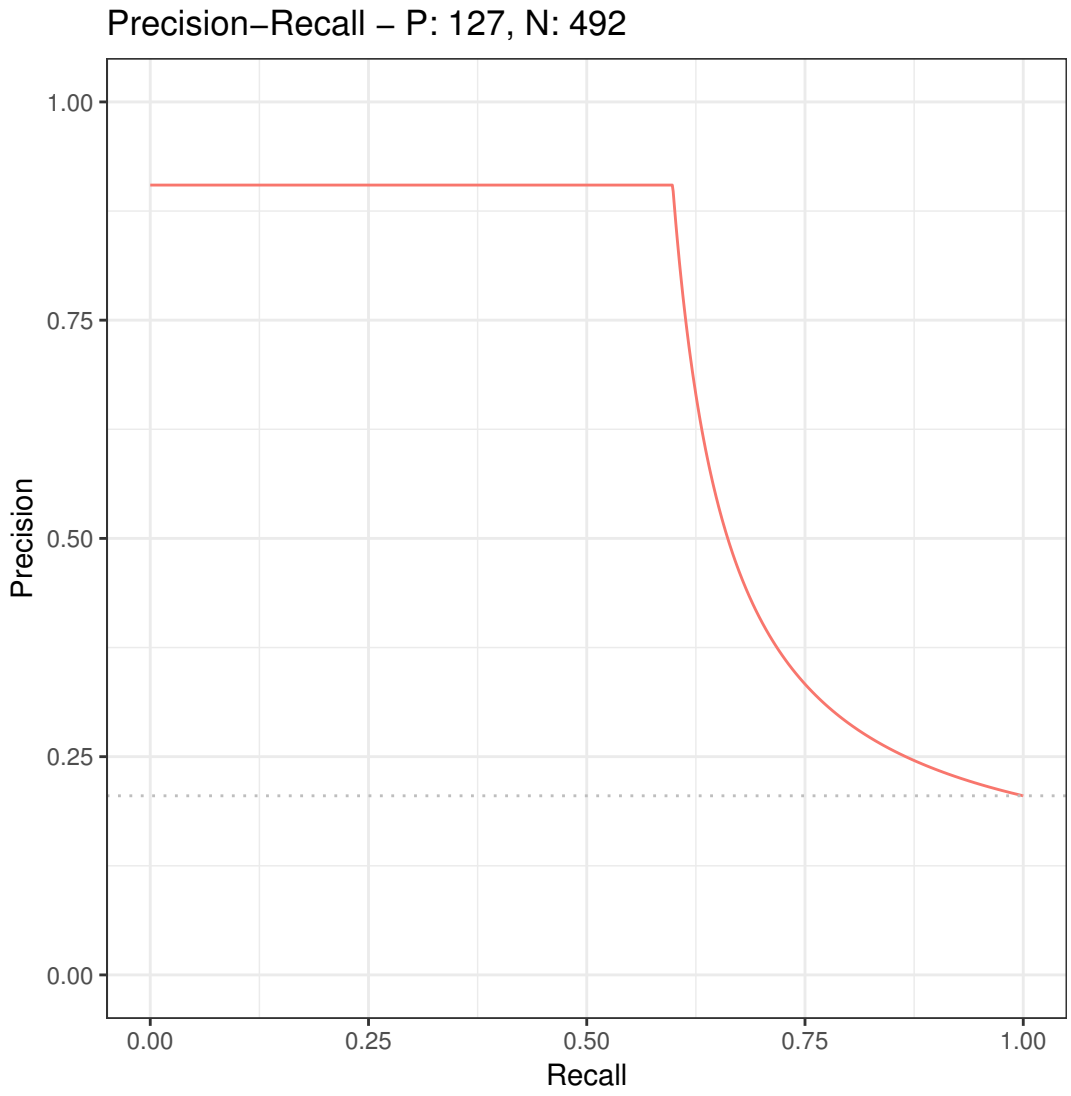

Figure 3.7: The precision recall curve for the classifier that identifies sentences containing the findings of studies.

3.6.2 Limitations

There are some limitations to the method proposed to automatically identify information to detect replications. First, the results of this approach demonstrate that the classifiers fail to identify all relevant sentences (i.e., perfect recall), which could mean that scientists miss some conditions or findings from studies. If the classifiers were further refined, perhaps recall and precision could increase, which would allow scientists greater access to the information in the articles.

Another limitation of this approach is the reliance on the fact that authors design effective studies, implement proper statistical analyses, and correctly report the findings of their studies. Many studies in the social science literature are underpowered (Maxwell, 2004) and deal with small sample sizes and imprecise measurements (Gelman and Carlin, 2014). Likewise, many papers have glaring statistical errors, resulting in findings that are implausible (Ioannidis, 2005; Nuijten et al., 2016), and different data analytic techniques can result in contradictory findings (Silberzahn et al., 2018). However, recent evidence suggests that some social science researchers have improved experimental and data analytic techniques, which could mitigate these problems (Nelson et al., 2018). This means that when older texts are used, researchers should be mindful of whether the study is completely explained and the findings do not follow from noise-mining.

Along with the increase in better experimental and statistical practices, the amount of scholarly text that is freely available has significantly increased. With funders of research (e.g., the National Science Foundation (National Science Foundation, 2015) and the Institute for Educational Sciences (Institute of Educational Sciences, 2016)), university systems (e.g., University of California System (University of California, 2015; Smith and Ventry, 2018)), and the scientific community increasingly encouraging sharing (Spellman et al., 2017; van der Zee and Reich, 2018), the ability to access machine readable text is still a barrier to scaling this method. The articles in this study were all PDFs, which required manual cleaning and conversion. This is feasible for a small set, but would likely be impractical for a larger corpus. To this end, all of the sentences in the study were coded by only me, as a previously labeled set of sentences was not available. Future work should incorporate multiple raters and estimates of inter-rater reliability.

3.6.3 Future Work

Besides obtaining annotations from multiple raters, there are several avenues of future research. The current approach requires scientists to determine whether studies are similar enough to be systematic replications and whether the findings of studies replicate. Automating this key step would allow for larger-scale studies of the literature, as with the current method, scientists must manually make these decisions. To automate this step, techniques would need to be developed that could model scientists decision making for whether two

studies are similar enough to one another to be deemed replications. Once this step is automated, then tools could identify whether particular findings are replicated in the literature and potentially identify contextual factors that are related to the findings.

Before the automated detection of replications could take place, the recall and the precision of the classifiers needs to be higher. The current study found acceptable performance for two of the classifiers but relatively poor performance for the other two. To improve the performance of the classifiers, more features besides unigrams and the location of sentences could be considered. For example, one set of features could represent the syntactic structure of sentences to identify sentences that contain the information sought by the classifiers. Ideally, new features would improve the generalizability of the classifiers that were built in this study. The unigrams used in the classifiers in this study were quite specific to the training set, so utilizing more general features could improve the performance of the classifiers on unseen articles.

3.7 Conclusion

This paper explored text classification strategies to identify the findings and conditions of published studies in full-text journal articles in the field of education. Specifically, classifiers were used to identify sentences containing results of studies and sentences containing descriptions about where the study took place, the subjects who participated in the study, and any manipulations that could influence inferences drawn from the studies. With this information, scientists can determine whether two published studies are similar enough to be considered replication studies and determine whether the findings of studies have been replicated in the literature.

To build the classifiers in this paper, I experimented with five supervised text classifiers and various features to optimize classifier performance on the basis of F_1. Two classifiers, SVMs and BOTR, performed the best when discerning between sentences describing the findings or conditions of published studies. Likewise, selecting unigrams by using IG and χ^2 statistics revealed the best performance. Beyond these contributions to the text classification literature, it was found that authors describe the conditions and findings of their studies throughout the text of a journal article. This suggests that using just the title and abstract of a journal article are insufficient, even though these portions of journal articles are exclusively used when mining academic texts.

The approach described in this paper is the first step toward automating the detection of replications in the scientific literature. The current approach can provide scientists with salient information about studies and requires them to manually determine whether studies are systematic replications and whether

the findings of studies are replicated. Automating these two steps could advance scientific discovery by removing barriers to scientists such as the inundation of scientific information and providing automation resolution to contradictions in the scientific literature.

Chapter 4

Reproducibility of Findings: The Details are in the Analyses

4.1 Introduction

The replicability crisis has called into question the reliability of social science research. This crisis has capitalized on the fact that the effects observed in one sample are not apparent in different samples within the same population, resulting in a lack of generalizability. Beyond the ability of a study's findings to generalize, another consideration is whether the findings of a study are *reproducible*. When replicating a study, a scientist seeks to investigate whether an effect exists across different samples or populations with varying contexts. On the other hand, reproducing a study seeks to implement the exact same procedures on collected data to arrive at the exact same effect estimates.

Because there are many problems with reported statistical analyses in the literature (e.g., between fifteen and fifty percent of statistical conclusions or calculations are incorrect (Bakker and Wicherts, 2011; Nuijten et al., 2016)), it is important that researchers besides the scientists who conduct a study are able to reproduce the reported results. The primary goal of this paper is to help scientists check the accuracy of findings and mitigate the use of Questionable Research Practices, or QRPs. These practices are estimated to be used by many social and biomedical scientists (John et al., 2012; Wang et al., 2018). To do so, this study utilizes text mining and natural language processing (NLP) approaches to extract the details needed to *reproduce* the statistical analyses reported in published papers, as it is often very difficult to reproduce the statistical and data analyses in published papers (Hardwicke et al., 2018). Specifically, the approach outlined in this paper proposes classifiers to extract the following statistical details from the text of journal articles:

1. the way that data were manipulated and structured after collection,
2. the method used to analyze the data,
3. the specific variables used in the analysis,
4. the implementation of this analytic technique,
5. and if any flexibilities or subjective data analytic decisions were disclosed.

To identify this information, various text classification strategies are used. To estimate the performance of these strategies, I used a set of 20 journal articles drawn from the field of education. The specific classification strategies proposed in this paper are regular expressions and supervised classifiers. After the classifiers are trained and tested, they are evaluated on the basis of accuracy and recall. Finally, I present ideas that could potentially improve these classifiers. Before I explain the classification strategy, I motivate this approach by discussing why this approach is presently viable and how others have extracted statistical details from published papers.

4.2 Related Work

Some proposed solutions to the replicability crisis have focused on making scientific practices more transparent, generally dubbed "open science" (Miguel et al., 2014; Spellman et al., 2017; van der Zee and Reich, 2018). With open science methods, scientists optimally make their original data, any statistical code or software, and research materials available for inspection and reuse by other scientists (Stodden, 2015). The ability of outsiders to easily inspect data and reanalyze it helps to ensure that QRPs were not conducted, such as the search for statistically significant findings (Simmons et al., 2011) and flexibility in data analysis (Gelman and Loken, 2014). While the ability to reproduce analyses will not eliminate problems such as the file-drawer (Rosenthal, 1979) or the lack-of-interest in replication studies (Madden et al., 1995; Makel and Plucker, 2014), it will help ensure that published findings are accurate.

To enhance participation in open science practices, initiatives have focused on awarding papers "badges" to indicate those papers that share research data and materials. After introducing badges for *Psychological Science*, data availability increased from 2.5% to 22.8% and materials availability increased from 12.7% to 30.3% (Kidwell et al., 2016). Unfortunately, sharing data and materials is not a panacea to improve the reproducibility of reported statistical analyses. While the journal *Cognition* has a mandatory data sharing policy, only 11 out of 35 simple statistical analyses were able to be reproduced by a team of scientists (Hardwicke et al., 2018). Therefore, we need to know what details authors need to reproduce the statistical analyses of published papers.

4.2.1 Reproducing Statistical Results

Problems and errors in data analysis of published papers are often functions of poor data collection, problems in handling and manipulating data, and errors in the choice of analytic technique or drawing appropriate conclusions from data collected (Brown et al., 2018). Therefore, we need information about these problematic

areas in order to reproduce the analyses in published papers.

First, we need to understand *how the data are structured and any manipulations the authors performed on the data*. Authors are often not specific about how they handled data, how or when certain participants are excluded, and whether missing data was problematic or how this was solved (Counsell and Harlow, 2017; Hardwicke et al., 2018). Thus, we need a clear understanding of how data are structured and were handled prior to analysis so that we may reproduce the findings. Once we have an understanding of the structure of the data and how it was manipulated, we need to know the specific *statistical method or data analytic technique* used to arrive at the findings. Authors are generally vague about the specific method(s) they used (Counsell and Harlow, 2017), for example, by stating they use a decision tree, but fail to state the specific algorithm they used to implement the model. In addition to the method used to analyze the data, authors should identify the *software used to implement any analyses*. Software might be in unstable releases (Epskamp, 2018) or assume certain default settings (Hardwicke et al., 2018), which could influence the reproducibility of findings (Stodden, 2015).

Once these preliminary details are known, we also need to know the *variables of interest*, and specifically, whether they were used as explantory or outcome variables. Surprisingly, authors do not always explicitly state which variables are used in their analyses as explanatory and response variables (Hardwicke et al., 2018). Additionally, we should be knowledgeable about any *subjective or flexible decisions* authors made. For example, in order to reproduce k-means clustering, we need to know the number of clusters, the distance metric used, whether multiple starting values were used, how many runs of the algorithm were considered, and how these choices were determined. By knowing this information, we can correctly re-implement the statistical analyses.

Beyond being able to reanalyze data using the same decisions as the original authors, we can identify flexibilities that might make findings non-replicable or non-reproducible. As Ioannidis (2005) notes, "the greater the flexibility in designs, definitions, outcomes, and analytic modes in a scientific field, the less likely the research findings are to be true" (p. e124). Knowledge of any subjective decisions or flexibilities in data analysis could help us detect whether authors found spurious relationships by going down the "garden of forking paths" or making data-dependent decisions (Gelman and Loken, 2014), excluded certain information to make findings appear more conclusive than the data support (Giner-Sorolla, 2012; John et al., 2012; Wang et al., 2018), or manipulating data and procedures to render statistically significant claims (Simmons et al., 2011). Beyond detecting these QRPs, understanding the choices authors made can help us understand their data analysis, as scientists might analyze the same data in many different ways and arrive at different conclusions (Silberzahn et al., 2018).

To detect the structure of data or manipulations to the data, the specific method or technique used to analyze the data, the software used for the analyses, which variables were considered, and any subjective or flexible choices the authors made (all five of which are referred to collectively as "components"), I develop a set of text mining and NLP tools. In the subsequent section, I identify ways that statistical details have been identified from published articles.

4.2.2 Text Mining and NLP for Statistical Details

Several text mining and NLP approaches have been developed to identify statistical details from scientific literature. Many of these approaches have explored biomedical literature, likely due to the standardized reporting of clinical trials, Consolidated Standards of Reporting Trials (Begg et al., 1996), or CONSORT, and the open-source availability of published texts (National Institutes of Health, 2008). Generally, tools for extracting study details have focused on certain types of information, such as the sample size (Hansen et al., 2008), descriptions of the population/problem, intervention, comparison, and outcome (PICO) elements (Boudin et al., 2010), or sentences describing randomized clinical trials (RCTs) (Kiritchenko et al., 2010). In contrast to identifying descriptions of clinical trials, Hsu et al. (2012) investigated the use of automated annotators, part-of-speech taggers, and regular expressions to identify statistical findings and details from biomedical articles.

Specifically, Hsu et al. (2012) built a system that extracted "p-values, confidence intervals, and statistical interpretation (e.g., no statistical difference)" (p. 354) from papers describing RCT studies about non-small cell lung cancer and chemotherapy treatments. Outside of biomedical sciences, others have built automated tools to investigate the reliability of p-values and methods to test the reliability of findings. The "statcheck" program (Nuijten et al., 2016) uses regular expressions to identify p-values reported using APA format, recomputes the p-values, and determines whether there is a discrepancy between the values reported in the paper and the recomputed value. The tools developed by Hsu et al. (2012) were used for small scale experiments on a small collection of annotated texts, while the tool proposed by Nuijten et al. (2016) were used in an unsupervised manner on a large set of texts. Some of the concepts needed to reproduce published statistical analyses follow a clear structure, and can use regular expressions to extract information. Others, however, are less structured, and I identify how these will be extracted from the text of journal articles next.

4.3 Methods

4.3.1 Statistical Details

As noted previously, scientists need five pieces of information from the text of journal articles to reproduce the statistical analyses in these articles. The text classifiers in this paper will identify sentences that contain information about the following five components:

- Data structure/processing: This concept captures sentences that describe specific structures of the data, the database schema, and table schema. In addition to identifying specifics on data structure, this concept captures any manipulations or transformations the authors performed on the data. For example, whether the authors excluded certain parts of the data (e.g., excluded participants who met some criteria) or made some transformations (e.g., created a composite score of several variables).

- Data analytic technique: This concept specifies the method used to analyze the data. These methods span parametric statistical techniques (for example, t-test or χ^2 test of independence) or other data analytic techniques (for example, machine learning algorithms such as naïve Bayes or C4.5).

- Software implementation: Because nearly all analyses in published papers are conducted using computer software, this concept captures which software is used to analyze data. Ideally, this will include the version or release date of the software.

- Variables of interest: For some of the analytic techniques, authors must specify the explanatory (or "independent") and response (or "dependent") variables that are used in the analyses. This concept captures when authors specifically identify whether and which variables are treated as explanatory or response variables in their analyses. Note that this component does not capture the variables when there does not exist an explanatory-response variable relationship.

- Subjective/flexible decisions: Researchers often make subjective choices about how to analyze data (e.g., might treat an ordinal variable as nominal, or discrete variables as continuous for simplicity). Additionally, they might be able to choose which variables are explanatory or response variables among many several measured outcomes. An example of this might be modeling success in a course, where potential dependent variables are the grade received (e.g., A, B, C, ...) or the proportion of points earned (e.g., 0 to 100%). This concept captures any subjective decisions authors communicate in their articles.

Before explaining how these components are identified by classifiers, I had to label whether or not sentences in parts of these articles contained one of these concepts. The specific coding guide used to annotate

sentences is provided in Appendix B. Before I specify the specific classification strategies used, I make note of preprocessing steps to allow for this study to be reproduced.

4.3.2 Selection and Processing of Studies

The journal articles used in this study are listed in Appendix A. In short, a collection of 65 articles written in English were considered for inclusion in this study. However, each article needed to be manually annotated using the coding guide, so I filtered the articles to a smaller set. Because the articles were all from the field of education, generally, and about what makes students successful in online classes, particularly, I only included studies about students or classes in the arts, education, business, or about multiple disciplines (i.e., the original study included students and courses for several different disciplines). Additionally, of the 65 articles considered, I omitted articles published prior to 2004 to make the set of articles more manageable. After these steps, the final set consisted of 20 articles. By including recent articles about diverse students and classes, this could enhance the potential examples the classifier might need to categorize outside of the training set.

Once the list of articles was finalized, I downloaded the full-text versions through either open-access websites or databases available through the University of Illinois at Urbana-Champaign library. All articles were downloaded in PDF format. Although PDFs are optimal for readers, they do not perform well when for text mining and NLP. Thus, all articles were manually converted (i.e., copy and pasted from PDF to raw text) and then analyzed. Once articles were in raw text format, I loaded them into R. During the conversion process, section headings were preserved to roughly follow the contexts specified in Purcell et al. (1997). Specifically, the following section headings were used: title, abstract, introduction, methods, result, discussion, and conclusion.

After being loaded into R (version 3.4.4), the articles were tokenized at the sentence level by the `tokenizers` package (Mullen and Selivanov, 2016). As such, sentences are used as the level of analysis throughout this study. Because articles were not annotated previously with the statistical concepts, I annotated each sentence using the coding guide in Appendix B. The articles were annotated such that a sentence was marked as containing a concept *only if* the concept was explicitly stated in the sentence. Further, a small-scale exploration revealed that many sentences describing statistical analyses and results are located within the methods and result section of articles. Therefore, only sentences that were in these two sections were considered. This resulted in $n = 1643$ total sentences to be considered for this study.

Once the sentences were annotated, various classification strategies were considered to identify sentences as containing the concepts of interest. Following from the success of regular expressions to identify elements

of clinical trials in Hsu et al. (2012), some classification tasks required only the raw text strings (i.e., the sentences with no pre-processing), while others required further processing. For the classification tasks that could not be accomplished via raw text strings, the following steps were taken to prepare the text for analysis.

Texts were prepared in two stages. Because part-of-speech (POS) tags were found to aid in classifying scientific artifacts in biomedical literature (Groza et al., 2013), the words in each sentence were tagged using the R package `cleanNLP` (Arnold, 2017), using the `spaCy` backend. After the sentences were tagged, I summed the number of times each POS appeared in the sentences. After POS tagging was complete, further text processing was completed. All characters were made to be lower case, English stop words were removed, and URLs were removed from the text using the R package `quanteda` (Benoit, 2018).

After the text were processed, text were converted to a bag-of-words representation. Specifically, words were represented in a matrix where the i^{th} column represents word i and the j^{th} row represents sentence j. Cell ij in this matrix is the frequency of word i in sentence j. In addition to the frequencies of the words, the bag-of-words matrix was augmented with the POS tags by sentence.

4.3.3 Methods of Classification and Features Used

Various techniques were used to extract the statistical details in the journal articles, as shown in Table 4.1. Regular expressions were used to identify sentences containing three of the concepts because authors often use familiar names for the methods they used to analyze their data (Counsell and Harlow, 2017) and primarily use certain software (Muenchen, 2018). Further, we are interested in sentences that identify the variables of interest by explicitly noting which explanatory or response variables were used, so I used regular expressions that identify sentences containing these words or words with similar connotations/meanings for these two ideas. For the other two concepts–data structure/manipulations and flexibilities/subjective choices in analyses–the words used by authors are likely less exact than those for the three concepts listed previously. Therefore, I used statistical classifiers to identify sentences containing these concepts.

To identify sentences that contain the methods used to analyze the data, the software used, and the variables of interest, regular expressions were used. Specifically, the algorithms developed loop through each sentence and determine whether or not the sentence contains one of the pre-determined words or phrases. If the sentence contains the word or phrase, it is marked as such by the algorithm. There is minimal text pre-processing required when identifying sentences containing these concepts. Because the algorithms require a specific set of words and phrases to classify sentences, I discuss how these were developed in the next section.

To detect sentences containing information about the data structure/processing of data and subjective/flexible decisions made by authors, statistical classifiers were used. Following from the findings in

Table 4.1: Text classification methods to identify the concepts needed to reproduce statistical analyses. Note that SVM=support vector machine; BOTR=boosted decision tree; BATR=bagged decision tree; POS=part-of-speech.

Concept	Classification Method	Text Representation
Data structure/processing	SVM, BOTR, BATR	bag-of-words and POS tags
Data analytic technique	regular expressions	raw string
Variables of interest	regular expressions	raw string
Software implementation	regular expressions	raw string
Subjective/flexible decisions	SVM, BOTR, BATR	bag-of-words and POS tags

Chapter 3, I explored the performance of three classifiers: support vector machines (SVM), boosted decision trees using AdaBoost (BOTR), and bagged decision trees (BATR). These were implemented by using `e1071` (Meyer et al., 2018), `ipred` (Peters and Hothorn, 2018), and `fastAdaboost` (Chatterjee, 2016), respectively.

To train the classifiers for these concepts, I used the bag-of-words generated when pre-processing the texts. Because many words within the corpus of sentences are not indicative of how the data were structured or subjective choices made by authors, I used three supervised feature selection strategies, χ^2 statistics, mutual information, and information gain, as defined in Equations 3.2, 3.4, and 3.3. For each feature selection strategy, I selected terms that had the highest 50, 100, and 200 scores from these metrics. In addition to the words selected by the feature selection strategies and appearing in each sentence, I included the POS frequencies generated previously. Next, I outline the specific steps used to develop the regular expression classifiers.

4.3.4 Developing Regular Expressions

Because regular expression classifiers were used for three of the concepts, we need to identify the exact phrases the classifier will search for in the raw text strings. I drew on published literature to identify sets of reasonable terms for these classifiers.

To detect sentences that contain information about the specific data analytic method, I drew on two sources that identified analyses that are ubiquitous in current research. Counsell and Harlow (2017) surveyed 126 articles published in several psychology journals and identified 19 statistical techniques[1]. The techniques identified include ubiquitous methods such as analysis of variance (ANOVA) and z or t tests for means, while also identifying specialized methods such as structural equation models and robust canonical correlation. Additionally, I included the techniques listed in Wu et al. (2008) to identify popular data mining approaches. Table 4.2 lists the specific methods identified from Counsell and Harlow (2017) and Wu et al. (2008) that were used to detect sentences for this component.

[1] A list of these methods is located in Table 1 of Counsell and Harlow (2017).

Table 4.2: Methods used in regular expressions to identify statistical and data analytic techniques. Note that descriptive statistics, due to vagueness in this technique, was omitted from the list of methods in Counsell and Harlow (2017).

Source	**Methods**
Counsell and Harlow (2017), Table 1, p. 142	ANOVA, z or t tests on means, multiple regression, correlation, χ^2, structural equation models, logistic regression, factor analysis, principal component analysis, ANCOVA, multilevel/mixed-effects models, generalized linear models, MANOVA, Mann-Whitney U test, z test on dependent correlations, meta-analysis, discriminant function analysis, (robust) canonical correlation analysis, MANCOVA
Wu et al. (2008)	Decision trees, C4.5, k-means, clustering, support vector machines, Apriori, association rules, mixture models, EM algorithm, PageRank, ensemble learning, AdaBoost, kNN/k-nearest neighbor, naïve Bayes, CART/classification and regression trees

Table 4.3: The software tools the algorithms are designed to search for within a text collection. MPlus was not included in Muenchen (2018), however, given that structural equation modeling and multilevel models were identified by Counsell and Harlow (2017), this software was included in this algorithm.

Source	**Software**
Muenchen (2018)	R, Stata, SAS, SPSS, Matlab, Weka, MPlus*, Python, RapidMiner, JMP, MiniTab, GraphPad Prism, Apache Hadoop, Statistica, C/C++/C#, Fortran, Apache Spark, Caffe, Apache Mahout, Tensorflow, IBM Watson, KNIME

To detect sentences describing the software used to analyze the data, I used the results from Muenchen (2018). Specifically, I used the software listed in Figure 2a of Muenchen (2018), which identifies software tools used by at least 750 articles following a search of papers catalogued by Google Scholar. The specific software that the regular expression classifier searches for are in Table 4.3.

Finally, to detect sentences that specifically identify the independent and dependent variables in their analyses, I used a list of terms that could be used to describe when a variable is treated as an explanatory or response variable [2]. These terms are listed in Table 4.4.

Now that the terms have been specified for the regular expression algorithms, we need to consider how to evaluate the classifiers that identify the five concepts of interest.

[2]This list was generated from `https://en.wikipedia.org/wiki/Dependent_and_independent_variables` on 6 November 2018.

Table 4.4: The terms used to search for sentences that contain independent and dependent variables.

Variables	**Search Terms**
Explanatory Variables	independent variable, predictor variable, regressor, covariate, control variable, manipulated variable, explanatory variable, exposure variable, risk factor, input variable
Response Variables	dependent variable, response variable, regressand, predicted variable, measured variable, explained variable, experimental variable, responding variable, outcome variable, label

4.3.5 Evaluating Classifier Performance

There are several information retrieval metrics that can be used to evaluate the performance of these classifiers. Each of the classification tasks are binary and the results of the classifiers can be displayed and evaluated using a confusion matrix, like the one in Table 3.4.

The most straightforward method of determining the performance of a classifier is to measure its accuracy (Manning et al., 2009), which is computed as

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.1}$$

Accuracy can give us evidence of the performance of these algorithms; however, it might be problematic because there are very few sentences that contain the information sought by the classifiers. Therefore, accuracy might be high, but the classifier might not effectively identify sentences containing the relevant components. Therefore, we would also like to ensure that all relevant information is identified by the classifiers so that the analyses can be reproduced given the identified sentences. To measure this, I will use recall to evaluate the classifiers. Following from the confusion matrix quantities in Table 3.4, recall is defined as in Equation 3.7. Specifically, recall is the proportion of sentences identified by the classifier as containing the relevant components out of all of the possible relevant sentences. When recall equals one, the classifiers identify every sentence that contains one of the components, while a recall of zero means the classifier does not identify any sentence that contains the relevant information.

Beyond the magnitudes of accuracy and recall, it is also helpful to consider the performance of the classifiers article-by-article. For example, if authors mention the software used throughout their manuscript, we only *need* to capture this information once in order to have the information needed to reproduce the findings of the study. Therefore, I will consider accuracy, recall, and within article performance.

Finally, the classifiers are evaluated using different test sets. For the three classifiers built using regular expressions, no training set was used because the regular expressions were not developed from the collection of annotated sentences. When using the statistical classifiers, the data were split into training and testing sets, where 16 articles were used as training examples and four articles were used as testing examples. This resulted in 1293 sentences to be used for training and 350 sentences to be used for testing.

Table 4.5: Most frequent terms in the collection of sentences.

Word	Frequency
students	991
learning	807
online	654
course	553
study	448
data	368
student	301
courses	300
time	199
results	191

4.4 Results

Before presenting the results of the classifiers, I present descriptive statistics about the prevalence of the component in the 20 articles. These are discussed next, followed by descriptions of the classifier performance.

4.4.1 Data Description

The 20 articles used to train and test these methods varied considerably in length with respect to the number of sentences in the methods and results section. The minimum number of sentences an article had in these two sections was 33, while the maximum number of sentences in these two sections was 136. The articles had a median of 85.5 sentences and a mean of 82.15 sentences in the methods and results section. Table 4.5 shows the most frequent terms in this collection of texts. It is not surprising that most terms are about online courses and students given the content of these articles.

Often, the sentences in the methods section described the procedures of the study–how the data were collected and constructs were measured–along with the way the data were analyzed. The results section generally included particulars about how the data were analyzed but not sentences about the procedures of the studies.

Table 4.6 shows the number of sentences in the corpus that communicate the five components of interest. Some of the components are more frequently discussed in the methods section (i.e., data structure/processing, software implementation), while others appear to be more frequently discussed in the results section (i.e., data analytic technique and variables of interest). However, each of these components are relatively rare in the set of sentences considered in this study, as noted by the proportion in the corpus. The concept that is most prevalent, subjective/flexible decisions, appears in around eight percent of sentences. Some components are, unsurprisingly, very rare. For example, authors describe the software they use in less than one percent of sentences.

Table 4.6: The proportion of sentences in each section and in the collection that discuss the concepts of interest. The total number of sentences in the corpus is $n = 1643$.

Concept	**Method**	**Results**	**Number of Sentences**	**Proportion in Corpus**
Data structure/processing	113	11	124	0.0755
Data analytic technique	35	58	93	0.0566
Variables of interest	21	80	101	0.0615
Software implementation	10	3	13	0.0079
Subjective/flexible decisions	74	60	134	0.0816

Table 4.7: Summary statistics for the number of sentences in each article that describe each concept.

Concept	Min	Median	Mean	Max	Variance
Data structure/processing	1	5	6.2	13	15.5368
Data analytic technique	0	4	4.65	10	5.9237
Variables of interest	0	2.5	5.05	22	31.2079
Software implementation	0	0	0.65	2	0.6605
Subjective/flexible decisions	0	5	6.7	20	35.5895

Finally, we consider how frequent the components are discussed in each article. In Table 4.7, we see that the components of interest are often represented by only a few sentences in each article. With the exception of the component data structure/processing, at least one article has no sentences that include that component.

4.4.2 Classifier Performance

Five classifiers were built, one for each of the five components needed to reproduce the statistical analyses in published papers. Due to the difference in test sets, I first present the results of the three classifiers using regular expressions and then I present the results of the statistical classifiers.

Regular Expression: Data Analytic Technique

First, we consider the regular expression classifier that detects whether a sentence discusses the statistical or data analytic method used to analyze the data. This classifier had an accuracy of 0.8941 and its recall was 0.7527. Because not all sentences containing a statistical method or data analytic technique were identified by the classifier, we should examine the 23 sentences that were false negatives to determine which methods or techniques were missed.

A closer examination of the false negatives revealed that the classifier misses some methods. For example, the studies in the test set use techniques such as "cumulative odds model," "ordered probit analysis," "Kruskal-Wallis test," "hierarchical regression analysis," and "hierarchical cluster analysis," and these methods were not included in the classifier. Some other sentences that were false negatives note that the authors conducted "regression analysis" or "cluster analysis." While these techniques are used to analyze the data, they are vague and not as specific as those listed in Table 4.2. When considering within article performance, this classifier was able to identify at least one analytic technique used in 18 of the 20 articles. Thus, the classifier was able to detect at least one sentence delineating a statistical method or data analytic technique in 90% of this collection.

Regular Expression: Software Implementation

Second, we consider the regular expression classifier that identifies whether or not a sentence identifies the software implementation used by authors. This classifier has an accuracy of 0.9939 and a recall of 0.9231. There are 13 sentences in this collection that identify the software used by authors, and 12 of these were correctly identified by the classifier. The false-negative sentence mentions "MS Excel," which was not identified as a popular software for data analysis in Muenchen (2018); however, this sentence was in an article that also used SPSS, and was correctly identified by the classifier.

Regular Expression: Variables of Interest

Next, we consider the regular expression classifier that identifies whether a sentence explicitly identifies the explanatory and response variables used in the statistical or data analyses. This classifier had an accuracy of 0.9276 and a recall of 0.3168. Out of the 101 sentences that explicitly define the explanatory and response variables, 32 sentences were identified by the classifier as doing so. An analysis of the false negatives revealed that many of the sentences describe the results of studies (e.g., by explaining that a certain explanatory variable is not significantly related to some response variable, without explicitly stating these variables as explanatory or response variables). When examining whether the regular expression classifier works well for each article, the classifier was able to identify sentences containing the variables of interest in 9 of the 20 articles.

Statistical Classifier: Data Structure/Processing

For the two classification tasks that used statistical classifiers, I experimented with three supervised classifiers (SVM, BOTR, BATR) and three feature selection strategies (χ^2, MI, IG) to select words to be used for these classifiers. Additionally, I considered whether 50, 100, or 200 of the top-ranked features from the selection strategies would perform the best.

Based solely on accuracy, the best classifier is the SVM with 50 words that were selected by IG, where accuracy was 0.9286 and recall was 0.2581. On the basis of recall, the best classifier was the SVM with 200 words selected by χ^2 statistics, and the accuracy was 0.9200 and recall was 0.2903. Because we consider recall as somewhat more important than accuracy, we will further consider the SVM with 200 words selected by the χ^2 statistic.

Of the 31 sentences in the test set that contain information about the structure of the data or an manipulations, 9 were identified by the SVM as containing this information. To consider within article performance, we examine Table 4.8, which considers the performance of the classifier for the four articles

Table 4.8: Within article performance of the SVM built with 200 words selected by χ^2 statistics to identify sentences describing the structure of data and how the authors manipulated data.

Article	**# of sentences with concept**	**# identified by classifier**	**recall**
One	12	1	0.0833
Two	5	4	0.8
Three	4	0	0
Four	10	4	0.4

Table 4.9: Top 20 words selected by the χ^2 statistic used in the SVM to detect sentences describing the structure of data and anything the authors did to manipulate the data.

Words Selected	database, data, removed, records, log, user, stored, excluded, student, within, moodle, file, information, lms, ranging, microsoft, spreadsheet, feature, 1, original

in the test set. We note that the classifier had high recall (0.8) for article two in the test set; however, the classifier was only able to identify one sentence that describes this concept in article one and none in article three. Because of this finding, I decided to take a closer look at the words selected by the χ^2 statistic and used in the classifier. These words are in Table 4.9.

From the list of words in Table 4.9, we see that there are terms that might describe how data are structured, for example *database*, *records*, *log*, *stored*, *moodle*, *spreadsheet*. Others, for example are likely indicative of preprocessing the data, such as *removed* and *excluded*. When examining the sentences that were correctly classified as containing details about the data structure or manipulations, six of the sentences were more focused on how the data were structured, while three were more focused on data manipulation.

Statistical Classifier: Subjectivity/Flexible Decisions

Finally, we consider the statistical classifier to detect sentences that identify any subjective decisions or flexibilities the authors presented in their data analyses. As before, I tested SVMs, BOTRs, and BATRs with the three feature selection strategies. Two of the classifiers I considered maximized accuracy at 0.8829. These classifiers were the SVM using 50 words selected by MI and the BOTR using 100 words selected by IG. For these two classifiers, the SVM had a recall of 0 and the BOTR had a recall of 0.0244. The classifier that maximized recall is the BATR with 100 words selected by IG where recall was 0.0732 and accuracy was 0.8543. This BATR model has slightly lower accuracy than the BOTR and SVM, but because recall is modestly higher, we will further consider this classifier.

In the test set, there were 41 sentences that communicated a subjective or flexible decision the authors made when analyzing their data. The BATR that maximized recall was able to correctly identify three sentences that communicated a subjectivity or flexibility. Table 4.10 shows the within article performance for this classifier. Generally, the classifier was unable to identify sentences that describe this component, but

Table 4.10: Within article performance of the BATR built with 100 words selected by IG to identify sentences describing the subjective or flexible decisions made by the study authors.

Article	**# of sentences with concept**	**# identified by classifier**	**recall**
One	17	2	0.1176
Two	0	0	0
Three	12	0	0
Four	11	1	0.0909

Table 4.11: Top 20 words selected by information gain used in the BATR to detect sentences describing the subjective decisions or flexibilities in data analysis that the authors communicated.

Words Selected	8, calculated, activity, time, points, 15, score, cumulative, experts, scores, two, weeks, differences, clusters, weighted, performed, four, 0, agreement, similarity

had some success in test articles one and four. When examining the text of the correctly identified sentences, there is no clear pattern with respect to the words in each sentence. The top 20 words selected when using information gain are in Table 4.11.

It is not apparent that the terms in Table 4.11 would communicate subjectivity or flexibility in data analysis. Perhaps some of these words might represent how authors scored responses, with the words *calculated cumulative scores*, or varying how they parse their data to focus on measures at *weeks* versus a different time length. The word *weighted* might be indicative of different techniques to make part of the data more important than another part. In addition, some unsupervised techniques allow authors to make many decisions, so it is not surprising to see *clusters* or *similarity* in this list of terms. The appearance of numerals and numbers likely suggests that the classifier overfit the training data, as these specific numbers probably do not indicate subjective of flexible decisions made by authors. Furthermore, these word identified in Table 4.11 could also be used for many other uses beyond the suggested examples.

4.5 Discussion

This study considered three regular expression and two statistical classifiers that sought to identify sentences containing five concepts needed to reproduce the statistical analyses in published papers. The three regular expression classifiers were developed by identifying various statistical methods and data analytic techniques, popular statistical software used in published papers, and whether the variables used were treated as explanatory or response variables. The statistical classifiers used the frequency of the parts-of-speech in each sentence coupled with words selected by supervised feature selection strategies to identify sentences that describe how data were structured or manipulated, and whether the authors made any subjective decisions or exhibited flexibilities in their analyses. When examining the annotated corpus of sentences used to test

these classifiers, it appears that these concepts are quite rare. One of the concepts (software implementation) appears in less than one percent of sentences, while authors communicated flexibilities and subjective choices in their data analysis in about eight percent of sentences in the annotated corpus.

Given the rarity of these concepts, it is expected that the accuracy of the classifiers is close to one. The classifier with the worst accuracy and recall was the classifier that identified sentences containing descriptions of subjective decisions made by authors, where the BATR with 100 words had a recall of 0.07317 and accuracy of 0.8543. The poor performance of this classifier is not very surprising, as sentences that communicate this concept likely use a variety of words that might not be clearly indicative of this concept. In contrast, authors clearly identify this information with some of the other concepts. For example, authors are likely to state the specific name of statistical software or the particular method they used with similar words as other authors, but might communicate the specific eccentricities of their data analysis less uniformly.

Overall, the results from these classifiers demonstrate that some of these approaches are viable to detect the statistical details needed to reproduce the analyses in published papers. The classifier to extract the statistical method had a recall of 0.7527, which is similar to the recall in Hsu et al. (2012), where recall was reported as 0.76. However, Hsu et al. (2012) used biomedical articles, and we used articles from education, which are likely less structured than biomedical articles (Sándor and Vorndran, 2009). Likewise, the classifier used to extract the statistical software used in published papers was also quite effective, with nearly perfect accuracy.

The classifier to identify sentences that explicate the variables of interest performed well on the basis of accuracy, but recall was somewhat low. Authors might not clearly identify their variables as explanatory or response variables; however, the relationship between variables might be implicit. Thus, from the results of this classifier, we note that some authors do explicitly state the role of each variable in their analyses, but others do not.

Finally, the statistical classifiers had relatively poor performance. The classifier to detect sentences describing the structure of the data or any manipulations performs quite well with respect to accuracy, at 0.9200, and fair recall, at 0.2903. The classifier to identify sentences describing subjective and flexible decisions had very low recall and fairly low accuracy. The poor performance is not very surprising, as authors likely describe the structure of the data, how they manipulate data, and any subjective or flexible decisions in their analyses with many different words.

In addition to individual performance for each classifier, we also observed that the classifiers perform quite well for some articles and need improvement when identifying this information from other articles. For example, the statistical classifier to detect the concept "data structure/processing" was able to identify 80

percent of the articles containing this concept in one article, but none of the sentences in another article. This finding suggests that future work should investigate how to improve this classifier to achieve high performance for all articles.

Beyond proposing and considering the performance of the proposed classifiers, this study used a set of 20 annotated journal articles to train and test these classifiers. Because these were annotated by one human, we have complete information about whether the concepts needed for reproducibility are included in this set of journal articles. While this is a small set of articles from several journals, this might give us insight into the availability of the information needed to reproduce analyses in published manuscripts. We noted that some concepts are never discussed in some articles, but others do have several sentences that discuss the concepts studied in this paper. However, as noted by Hardwicke et al. (2018) when reproducing statistical analyses from *Cognition*, "a common problem we encountered was unclear, incomplete, or incorrect specification of the data analysis pipeline in the original article" (p. 14). This appears to be true in this collection, as many articles have very few sentences that describe how the data were analyzed.

On the whole, this paper provides several contributions to the extant literature. First, it provides a descriptive summary about the number of sentences that discuss the concepts needed to reproduce the statistical analyses in a set of published articles in the field of education. The 20 articles considered in this study predate wide-reaching conversations about the replicability and reproducibility of social science research, but they do provide some evidence that authors often provide scant details about how their data were analyzed. Further, this work contributes to on-going development of tools to extract information from the text of journal articles. Whereas several investigations have identified and evaluated methods to extract statistical details, and particularly findings of studies, the work in this paper targets details needed to reproduce statistical findings. Some of the details needed to reproduce statistical analyses were extracted with much success in this paper, while others need refinement before putting these classifiers in production.

4.5.1 Future Work

While all of the classifiers performed well with respect to accuracy, recall was quite low for some of the classifiers. For example, the classifier to detect the explanatory and response variables used in a study had fairly low recall. Perhaps this classifier could be improved by looking at sentences that contain specific syntactic structures. Work in mining biomedical literature has investigated the use of dependency parses that identify relationships between related concepts in the sentences of journal articles (for example, Blake (2010)). When looking for independent and dependent variables, authors presumably identify the relationship between variables, and looking for sentences with certain types of relationships could improve the performance of

this classifier. Additional work should include specifications for identifying the variables that are considered when there is not an explanatory-response variable relationship present, such as unsupervised techniques.

For the statistical classifiers, other methods beyond those considered in this study could be considered to improve performance. For example, some methods have been proposed to improve the performance of SVMs, and these methods could be considered to refine those used in this investigation. To this end, parameter tuning could be considered, while minding the potential for overfitting to the training data. Besides exploring different models that could enhance the performance of these classifiers, other features besides words and POS tags could be considered. For example, words could be represented using embeddings, such as the Word2Vec model (Mikolov et al., 2013). This could potentially improve performance over the bag-of-words model assumed in this paper. Ideally, representing words in this manner could enhance the generalizability of these models.

With respect to generalizability, the sentences used to train and test these classifiers were from articles in education, generally, and about what makes college students successful in online courses, specifically. With articles from different disciplines, the feature set for the statistical classifiers should be larger and have a wider set of words to indicate sentences describing the structure of data or any subjective choices authors make when analyzing data. Once classifiers are less domain specific, and performance is improved by experimenting with the aforementioned methods, the classifiers could be productionalized or released as open-source software to help reviewers screen submissions using statistical methods or data analytic techniques for reproducibility.

4.5.2 Limitations

There are certain limitations that should be considered for this method. At the present, the development of classifiers is limited due to the lack of training examples. That is, there was not a previously annotated corpus to identify these five components in journal articles. Because of time-constraints, this could limit the development of classifiers for these five components. Likewise, the articles in the corpus analyzed in this paper were annotated by only the author. Further refinement of the training and testing examples should be accompanied by estimates of inter-rater reliability with additional coders.

Beyond the limitations due to the lack of available data, another is the applicability of the method. While open science practices have encouraged greater data availability, the sharing of data collected from studies is not mainstream and has many logistical challenges (i.e., storage, preservation, useability) (van der Zee and Reich, 2018). To this end, the ability to reproduce published analyses from the description and instructions in journal articles has proven to be difficult. The method proposed in this paper implicitly

assumes the data analysis procedure is sufficiently described in the manuscript so others may reproduce the analyses. However, given the scarcity of details authors provide, it may not be possible to reproduce the analyses given just the text of analyses.

Finally, while this approach identified in this paper has implications for the reproducibility of analyses, much of the conversation in the social sciences has focused on the non-replicability of findings. Unfortunately, reproducibility does not imply replicability (Leek and Peng, 2015). That is, even if the analyses are able to be reproduced, and the same parameter estimates are obtained, this does not mean that the effect is replicable beyond that instance. Thus, further work should investigate the relationship between reproducibility and replicability.

4.6 Conclusion

This paper proposed classifiers to extract the statistical details needed to reproduce the analyses in published papers. To do so, we identified five concepts that those who seek to reproduce statistical analyses would need to reanalyze data. Specifically, these concepts are the structure of the data and any manipulations authors performed, the specific statistical or data analytic technique used, the software implementation of the analysis, the specific independent and dependent variables used in the analysis, and any subjective or flexible decisions the authors made when conducting their analyses. To extract these details from the text of journal articles, we used a set of 20 published papers from educational research journals that examined the success of college students in online classes.

The specific classifiers developed employed regular expressions and text classifiers such as SVMs, and ensemble techniques of boosting and bagging, where the base classifiers were decision trees. We found that many of the classifiers were able to achieve fairly high accuracy, although some had very low recall. We suspect that authors likely use many different words when discussing these concepts, and this might influence the ability of the classifiers to effectively identify these concepts from the set of articles considered in this study. We hypothesized about various features and techniques that could be used to enhance the performance of these classifiers and make them more robust.

While being able to reproduce the analyses of a study does not imply that the findings of a study are observable beyond the confines of that paper it demonstrates that the analytic technique was well documented. Further, if the data and this technique are correct for the data, it renders credibility to the findings of the study. With changing attitudes towards open science and ensuring findings are robust, the techniques proposed in this paper could help reviewers and authors check their work to make sure they are

complete and transparent when explaining how they analyzed their data.

Chapter 5

Conclusion

As noted throughout this thesis, the inability of findings to be replicated or reproduced can lead to the detriment of the reliability and credibility of educational research. Replicating and reproducing research is vital to the scientific enterprise as it allows us to check and ensure the robustness, reliability, and accuracy of scientific claims made by researchers. To replicate a study in education, we need to design a study that examines the same hypothesized relationship as another study while being mindful of how the context of a study can influence its findings. When reproducing the statistical analyses of a study, we must be aware of the choices researchers made throughout the data collection and analysis phase to arrive at the exact same conclusions. This thesis explored ways that we can identify the contextual factors, the findings, and data analytic steps that researchers specify in journal articles.

In chapter three, we developed a text classification strategy that identifies the details needed to determine whether published studies are systematic replications of one another. Drawing on the findings in chapter two, we needed information about the hypothesized relationship that was tested in the paper and the contexts in which it was tested. Specifically, we built text classifiers that identified sentences from the text of journal articles about the findings of studies, where we can infer the hypothesized relationship of interest, and details about where a study took place, who was studied, and any manipulations the authors considered. The results demonstrated that identifying the findings and locations of studies is relatively easy for the classifiers, but identifying who was studied and any manipulations the authors considered was difficult. After building hundreds of classifiers, we found that support vector machines and boosted decision trees, while using information gain and χ^2 statistics to select the features for these models, yielded the best performance. Future work should focus on improvement of classifier performance by considering other features and classification techniques and should develop more automated approaches to detecting systematic replications in the extant literature.

Given the amount of published findings, automatically detecting whether one study is a replication of another is a key next step from the work in this thesis. Before this is possible, the difference between a replication study and a non-replication study should be made more explicit. Specifically, a quantitative

measure of how "different" studies are on the basis of contextual factors and the studied relationships of interest could help us determine whether studies are replications of one another. To date, no such metric has been developed, and this is needed to introduce more automation into systems to identify systematic replications. If two studies are considered to be replications based on this difference metric, then we can use the information from the text classifiers in chapter 3 to automatically accumulate which findings have replicated and under what contexts. This would remove the current step of requiring humans to determine whether a study is sufficiently similar to another to deem it a replication and accumulate the findings relevant to these studies.

In addition to proposing a text classification strategy to detect the findings and contexts of published studies, we also proposed text classifiers that extract the details needed to reproduce statistical and data analyses in published papers. The classifiers identified sentences that describe how authors analyzed data (i.e., the method and software used), which variables were collected and analyzed, how their data were structured, and any subjective or flexible decisions they made. We found that some of these details are readily detectable by regular expressions, such as the method used to analyze data and which software the authors used, but others are not as easily identified. Particularly, subjective decisions made when researchers analyze data were very difficult for the classifiers to identify from the set of sentences examined. Future work should further improve these classifiers so that they can be used to help researchers and reviewers gauge how transparent authors are when describing their statistical and data analyses.

As noted previously, the approach developed in chapter three can help scientists determine whether studies are systematic replications of one another and accumulate findings in support for or against hypothesized relationships, but a key next step would be to automate human decision making about whether studies are replications of one another. To further integrate the findings and approaches in this thesis, the work in chapter four could be utilized when automating the discovery of published replication studies. A measure of whether the statistical or data analyses are reproducible could filter-out studies that are not reproducible. Likewise, further work could identify papers that contain statistical errors or problematic practices, and not include these potentially troublesome findings when automatically detecting systematic replications.

Beyond improving the classifiers used in this study, future work could consider casting this problem as an information retrieval task. Instead of classifying sentences as containing the needed information, an information retrieval approach would allow sentences to be ranked that are likely to contain the needed information. On the other hand, a different approach could be to use an information extraction approach, which could help identify and summarize information across different studies for users. To this end, further work in developing systems to help scientists should consider which method would yield the most utility for

users. These designs should consider the distillation of the most pertinent sources of information to help scientists identify which scientific claims have been replicated.

In sum, this thesis contributes to the on-going conversation about the replicability and reproducibility of social science research and the development of tools to process and realize insights from scholarly texts. Many prior recommendations have identified general approaches to mitigating the chances of non-replicability or non-reproducibility, but stop short of identifying specific recommendations to enhance the replicability and reproducibility of research. The framework in chapter two speaks to this gap by explicating how a systematic replication study can be designed and the statistical details needed to reproduce a study, which augments recent developments such as the open science movement. The approaches in chapters three and four take this framework and develop strategies to glean information from the text of journal articles. These chapters showed that targeted information can be identified from social science articles with some success. While many of the developments on mining scholarly literature have mined biomedical texts, this thesis demonstrates that there is promise in mining social science articles. More broadly, the work in this thesis is a step towards leveraging text mining and natural language processing to understand the replicability and reproducibility of scientific claims.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., and Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1(3):357–366.

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 163–222. Springer Science+Business Media.

Allen, E. and Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *Journal of the American Medical Association*, 282:634–635.

Anderson, S. F. and Maxwell, S. E. (2016). Theres more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12.

Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., and Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 71–78. ACM.

Andres, J. M. L., Baker, R. S., Siemens, G., Gašević, D., and Spann, C. A. (2017). Replicating 21 findings on student success in online learning. *Technology, Instruction, Cognition, and Learning*, 10(4):313–333.

Arnold, T. (2017). A tidy data model for natural language processing using cleannlp. *The R Journal*, 9(2):1–20.

Aytug, Z. G., Rothstein, H. R., Zhou, W., and Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, 15(1):103–133.

Bakker, M. and Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3):666–678.

Bauernfeind, R. H. (1968). The need for replication in educational research. *The Phi Delta Kappan*, 50(2):126–128.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., and Simel, D. (1996). Improving the quality of reporting of randomized controlled trials: The consort statement. *Journal of the American Medical Association*, 276(8):637–639.

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data*. R package version 1.2.0.

Berliner, D. C. and Glass, G. V. (2015). Trust but verify. *Educational Leadership*, 72(5):10–14.

Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times Magazine*.

Black, E. W., Dawson, K., and Priem, J. (2008). Data for free: Using LMS activity logs to measure community in online courses. *The Internet and Higher Education*, 11(2):65–70.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

Blake, C. and Pratt, W. (2006). Collaborative information synthesis ii: Recommendations for information systems to support synthesis activities. *Journal of the American Society for Information Science and Technology*, 57(14):1888–1895.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Boudin, F., Nie, J.-Y., Bartlett, J. C., Grad, R., Pluye, P., and Dawes, M. (2010). Combining classifiers for robust PICO element detection. *BMC medical informatics and decision making*, 10(1):29.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., and Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50:217–224.

Brown, A. W., Kaiser, K. A., and Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, 115(11):2563–2570.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., and Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, 96:42–54.

Chatterjee, S. (2016). *fastAdaboost: a Fast Implementation of Adaboost*. R package version 1.0.0.

Chhin, C. S., Taylor, K. A., and Wei, W. S. (2018). Supporting a culture of replication: An examination of education and special education research grants funded by the institute of education sciences. *Educational Researcher*, 47(9):594–605.

Cook, B. G., Collins, L. W., Cook, S. C., and Cook, L. (2016). A replication by any other name: A systematic review of replicative intervention studies. *Remedial and Special Education*, 37(4):223–234.

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., and Therrien, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children*, 85(1):104–118.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

Counsell, A. and Harlow, L. L. (2017). Reporting practices and use of quantitative methods in canadian journal articles in psychology. *Canadian Psychology/psychologie canadienne*, 58(2):140–147.

Coyne, M. D., Cook, B. G., and Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education*, 37(4):244–253.

Crandall, C. S. and Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66:93–99.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1):7–29.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, pages 71–78.

Epskamp, S. (2018). Reproducibility and replicability in a fast-paced methodological world.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? an empirical support from us states data. *PloS one*, 5(4):e10271.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Flake, J. K., Pek, J., and Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4):370–378.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Gabb, H. A., Lucic, A., and Blake, C. (2015). A method to automatically identify the results from journal articles. *iConference 2015 Proceedings*.

Gehlbach, H. and Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, 11(2):296–315.

Gelman, A. (2015). Working through some issues. *Significance*, 12(3):33–35.

Gelman, A. (2018a). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*, 41.

Gelman, A. (2018b). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, 44(1):16–23.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6):460 – 465.

Gelman, A., Mattson, G., and Simpson, D. (2018). Gaydar and the fallacy of decontextualized measurement. *Sociological Science*, 5(12):270–280.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7(6):562–571.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12.

Greenfield, P. M. (2017). Cultural change over time: Why replicability should not be the gold standard in psychological science. *Perspectives on Psychological Science*, 12(5):762–771.

Groza, T., Hassanzadeh, H., and Hunter, J. (2013). Recognizing scientific artifacts in biomedical literature. *Biomedical Informatics Insights*, 6:15.

Guo, X., Yin, Y., Dong, C., Yang, G., and Zhou, G. (2008). On the class imbalance problem. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 4, pages 192–201. IEEE.

Hansen, M. J., Rasmussen, N. Ø., and Chung, G. (2008). A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., and Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal cognition. *Royal Society Open Science*, 5(8).

Hoekstra, R., Kiers, H. A., and Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3.

Holder, B. (2007). An investigation of hope, academics, environment, and motivation as predictors of persistence in higher education online programs. *The Internet and Higher Education*, 10(4):245–260.

Howe, K. R. and Moses, M. S. (1999). Ethics in educational research. *Review of Research in Education*, 24(1):21–59.

Hsu, W., Speier, W., and Taira, R. K. (2012). Automated extraction of reported statistical analyses: Towards a logical representation of clinical trial literature. In *AMIA Annual Symposium Proceedings*, volume 2012, page 350. American Medical Informatics Association.

Hunt, K. (1975). Do we really need more replications? *Psychological Reports*, 36(2):587–593.

Institute of Educational Sciences (2016). IES policy regarding public access to research.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Jamil, F. M. (2018). A reflection on the evolution of a replication study. *Journal for Research in Mathematics Education*, 49(1):111–115.

Jiang, J. (2012). Information extraction from text. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 11–42. Springer Science+Business Media.

Joel, S., Eastwick, P. W., and Finkel, E. J. (2018). Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science*, 1(1):86–94.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532.

Kauffman, H. (2015). A review of predictive factors of student success in and satisfaction with online learning. *Research in Learning Technology*, 23.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217.

Kidwell, M. C., Lazarevi, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., and Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14(5):1–15.

Kim, R., Olfman, L., Ryan, T., and Eryilmaz, E. (2014). Leveraging a personalized system to improve self-directed learning in online educational environments. *Computers & Education*, 70:150–160.

Kiritchenko, S., de Bruijn, B., Carini, S., Martin, J., and Sim, I. (2010). Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):56.

Klingsieck, K. B., Fries, S., Horz, C., and Hofer, M. (2012). Procrastination in a distance university setting. *Distance Education*, 33(3):295–310.

Lang, T. A. and Altman, D. G. (2013). Basic statistical reporting for articles published in biomedical journals: the "statistical analyses and methods in the published literature" or the sampl guidelines. *Handbook, European Association of Science Editors*, 256:256.

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., and Goodman, S. N. (2017). Five ways to fix statistics. *Nature*, 551(7682):557–559.

Leek, J. T. and Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. R. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

Lindsay, R. M. and Ehrenberg, A. S. (1993). The design of replicated studies. *The American Statistician*, 47(3):217–228.

Madden, C. S., Easley, R. W., and Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising*, 24(4):77–87.

Makel, M. C. and Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6):304–316.

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7(6):537–542.

Malmivaara, A. (2019). Generalizability of findings from randomized controlled trials is limited in the leading general medical journals. *Journal of Clinical Epidemiology*, 107:36–41.

Malouf, D. B. and Taymans, J. M. (2016). Anatomy of an evidence base. *Educational Researcher*, 45(8):454–459.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to information retrieval*. Cambridge University Press.

Martin, G. N. and Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, 8:523.

Martone, M. E., Garcia-Castro, A., and VandenBos, G. R. (2018). Data sharing in psychology. *American Psychologist*, 73(2):111–125.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., and O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9(2):147–163.

McBee, M. T., Makel, M. C., Peters, S. J., and Matthews, M. S. (2018). A call for open science in giftedness research. *Gifted Child Quarterly*, 62(4):374–388.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-0.

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, 1(1):131–144.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., and Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343(6166):30–31.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., and Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS medicine*, 4(3).

Muenchen, R. A. (2018). The popularity of data science software.

Mullen, L. and Selivanov, D. (2016). tokenizers: A consistent interface to tokenize natural language text.

National Institutes of Health (2008). Revised policy on enhancing public access to archived publications resulting from nih-funded research.

National Science Foundation (2015). NSF's public access plan: Today's data, tomorrow's discoveries (nsf 15-22).

Nelson, L. D., Simmons, J., and Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69:511–534.

Nosek, B. A. and Lakens, D. (2014). Registered reports. *Social Psychology*, 45(3):137–141.

Notice from Department of Health and Human Services (2015). Findings of research misconduct. pages 69230–69231.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., and Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1):5.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–1–aac4716–8.

Papamitsiou, Z. and Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64.

Pashler, H. and Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Patil, P., Peng, R. D., and Leek, J. (2016a). A statistical definition for reproducibility and replicability. *bioRxiv*, page 066803.

Patil, P., Peng, R. D., and Leek, J. T. (2016b). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–544.

Peters, A. and Hothorn, T. (2018). *ipred: Improved Predictors*. R package version 0.9-8.

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., and Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, 42(8):424–432.

Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:76.

Purcell, G. P., Rennels, G. D., and Shortliffe, E. H. (1997). Development and evaluation of a context-based document representation for searching the medical literature. *International Journal on Digital Libraries*, 1(3):288–296.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638–641.

Sándor, Á. and Vorndran, A. (2009). Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 36–44. Association for Computational Linguistics.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2):90–100.

Schulz, K. F., Altman, D. G., and Moher, D. (2010). Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1):18.

Shadish, W. R. (1995). The logic of generalization: Five principles common to experiments and ethnographies. *American Journal of Community Psychology*, 23(3):419–428.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Š. Bahnk, Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T. J., Hederos, K., Heene, M., Mohr, A. J. H., Hgden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D. M., Lei, R., Lindsay, T. A., Liverani, S., Madan, C. R., Molden, D., Molleman, E., Morey, R. D., Mulder, L. B., Nijstad, B. R., Pope, N. G., Pope, B., Prenoveau, J. M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlter, E., Schönbrodt, F. D., Sherman, M. F., Sommer, S. A., Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., and Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356.

Simmons, J., Nelson, L., and Simonsohn, U. (2012). A 21 word solution. *Dialogue*, 26:4–7.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1):76–80.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569.

Smith, M. and Ventry, D. J. (2018). Potential changes to ucs relationship with elsevier in january 2019. date accessed 4 December 2018.

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, 25(10):970–975.

Smith, V. C., Lange, A., and Huston, D. R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks*, 16(3):51–61.

Spellman, B., Gilbert, E. A., and Corker, K. S. (2017). Open science: What, why, and how. Accessed on 6 Feb 2018 from psyarxiv.com/ak6jr.

Star, J. R. (2018). When and why replication studies should be published: Guidelines for mathematics education journals. *Journal for Research in Mathematics Education*, 49(1):98–103.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in psychology*, 8:862.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2:1–19.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.

Tackett, J. L. and McShane, B. B. (2018). Conceptualizing and evaluating replication across domains of behavioral research. *arXiv preprint arXiv:1801.05049*.

Therneau, T. and Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

University of California (2015). UC Presidential Open Access Policy.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., and Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23):6454–6459.

van der Zee, T. and Reich, J. (2018). Open education science. *AERA Open*, 4(3).

Wang, M., Yan, A., and Katz, R. (2018). Researcher requests for inappropriate analysis and reporting: A u.s. survey of consulting biostatisticians. *Annals of Internal Medicine*, 169(8):554–558.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70:129–133.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., and Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1832.

Wilkinson, L. and The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8):594–604.

Willging, P. A. and Johnson, S. D. (2004). Factors that influence students' decision to dropout of online courses. *Journal of Asynchronous Learning Networks*, 13(3):115–127.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

Zimbardo, P. G. (1973). On the ethics of intervention in human psychological research: With special reference to the stanford prison experiment. *Cognition*, 2(2):243 – 256.

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41:e120.

Appendix A

List of Articles

Table A.1: 20 articles used to build and test the classifiers in Chapters 3 and 4

Authors	**Year**	**Title**	**Journal**	**Volume**	**Issue**	**Pages**
Abdous, M'hammed; He, Wu; Yen, Cherng-Jyh	2012	Using data mining for predicting relationships between online question theme and final grade	*Journal of Educational Technology & Society*	15	3	77-88
Agudo-Peregrina, Ángel F.; Iglesias-Pradas, Santiago; Conde-González, Miguel Ángel; Hernández-García, Ángel	2014	Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning	*Computers in Human Behavior*	31	n/a	542-550
Aragon, Steven R.; Johnson, Elaine S.	2008	Factors influencing completion and noncompletion of community college online courses	*American Journal of Distance Education*	22	3	146-158

Table A.1: Continued

Authors	**Year**	**Title**	**Journal**	**Volume**	**Issue**	**Pages**
Beqiri, Mirjeta S.; Chase, Nancy M.; Bishka, Atena	2009	Online course delivery: An empirical investigation of factors affecting student satisfaction	*Journal of Education for Business*	85	2	95-100
Black, Eric W.; Dawson, Kara; Priem, Jason	2008	Data for free: Using LMS activity logs to measure community in online courses	*Internet and Higher Education*	11	2	65-70
Davies, Jo; Graff, Marin	2005	Performance in e-learning: Online participation and student grades	*British Journal of Educational Technology*	36	4	657-663
Holder, Bruce	2007	An investigation of hope, academics, environment, and motivation as predictors of persistence in higher education online programs	*Internet and Higher Education*	10	4	245-260
Hu, Ya-Han; Lo, Chia-Lun; Shih, Sheng-Pao	2014	Developing early warning systems to predict students' online learning performance	*Computers in Human Behavior*	36	n/a	469-478
Hung, Jui-Long; Rice, Kerry; Saba, Anthony	2012	An educational data mining model for online teaching and learning	*Journal of Educational Technology Development and Exchange*	5	2	77-94

Table A.1: Continued

Authors	**Year**	**Title**	**Journal**	**Volume**	**Issue**	**Pages**
Iglesias-Pradas, Santiago; Ruiz-de-Azcárate, Carmen; Agudo-Peregrina, Ángel F.	2015	Assessing the suitability of student interactions from moodle data logs as predictors of cross-curricular competencies	*Computers in Human Behavior*	47	n/a	81-89
Kim, Jungjoo; Kwon, Yangyi; Cho, Daeyeon	2011	Investigating factors that influence social presence and learning outcomes in distance higher education	*Computers & Education*	57	2	1512-1520
Kim, Rosemary; Olfman, Lorne; Ryan, Terry; Eryilmaz, Evren	2014	Leveraging a personalized system to improve self-directed learning in online educational environments	*Computers & Education*	70	n/a	150-160
Klingsieck, Katrin B.; Fries, Stefan; Horz, Claudia; Hofer, Manfred	2012	Procrastination in a distance university setting	*Distance Education*	33	3	295-310
McElroy, Barbara Woods; Lubich, Bruce H.	2013	Predictors of course outcomes: Early indicators of delay in online classrooms	*Distance Education*	34	1	84-96
Nistor, Nicolae; Neubauer, Katrin	2010	From participation to dropout: Quantitative participation patterns in online university courses	*Computers & Education*	55	2	663-672

Table A.1: Continued

Authors	**Year**	**Title**	**Journal**	**Volume**	**Issue**	**Pages**
Shelton, Brett E.; Hung, Jui-Long; Lowenthal, Patrick R.	2017	Predicting student success by modeling student interaction in asynchronous online courses	*Distance Education*	38	1	59-69
Smith, Vernon C.; Lange, Adam; Huston, Daniel R.	2012	Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses	*Journal of Asynchronous Learning Networks*	16	3	51-61
Willging, Pedro A.; Johnson, Scott D.	2004	Factors that influence students' decision to dropout of online courses	*Journal of Asynchronous Learning Networks*	8	4	105-118
You, Ji Won	2015	Examining the effect of academic procrastination on achievement using LMS data in e-learning	*Journal of Educational Technology & Society*	18	3	64-74
You, Ji Won	2016	Identifying significant indicators using LMS data to predict course achievement in online learning	*Internet and Higher Education*	29	n/a	23-30

Appendix B

Coding Guide for Reproducing Statistical Results

Each sentence in the corpus were marked as either having or not having the information listed below. Sentences may contain more than one bit of information.

1. Raw data transformations
    (a) Does the sentence describe the structure of the data?
    (b) Does the sentence contain where the data are stored for public access?
2. Data transformations and processing
    (a) Does the sentence describe any data cleaning procedures? (e.g., exclusions of data, transformations)
    (b) Does the sentence denote where a traceable script or program to re-implement any data cleaning?
    (c) Does the sentence contain the 21-word solution (Simmons et al., 2012) or anything that appears to capture the spirit of this statement?
    (d) Are there any other relevant statistical details that would help someone outside of the project draw meaning from the data?
3. Data analysis plans
    (a) Does the sentence name the exact statistical method or analytic technique?
    (b) Is there acknowledgement that the analytic plan was derived before data collection?
    (c) Does the sentence acknowledge the assumption of the method?
    (d) Does the sentence describe how the authors test the methods of the assumption?
    (e) Does the sentence state the exact hypotheses tested?
    (f) Does the sentence describe any sequential testing or simplifications to the initially proposed test or model? (For example, how authors simplified a regression model or evaluated their model/approach)

(g) Does the sentence acknowledge the software or other tools used to implement the data analysis plan (not pre-processing or storage)?

(h) Does the sentence describe any settings used in the software?

(i) Does the sentence acknowledge any subjective choices they made? (For example, determining the number of clusters in k-means clustering)

(j) Does the sentence explicitly identify (if any) the independent/explanatory variables used in the analysis?

(k) Does the sentence explicitly identify (if any) the dependent/response variables used in the analysis?

4. Difficulties in analysis

   (a) Does the sentence make mention of errors in data?

   (b) Does the sentence note whether missing data were problematic and how this was resolved?

   (c) Does the sentence communicate any other problems encountered with respect to data analysis?

5. Flexibilities in collection and analysis

   (a) Does this sentence acknowledge flexibility in data collection?

   (b) Does this sentence acknowledge flexibility in data analysis?

6. Overview and rationale

   (a) Does the sentence provide an overview of the analytic technique specific to the particular study (e.g., how data were specifically used in a technique)?

   (b) Does the sentence provide a rationale for a particular technique?

Notes

- Sentences that reference figures were not coded as containing the information as the approach relies on the text of sentences
- Sentences describing scale validation, when the primary goal of the study is to use the scales as a measurement/data collection tool, are not coded as containing this information

Appendix C

Regular Expressions Classifiers

There were three regular expressions developed in chapter 4. The following is the R (version 3.4.4) of the three regular expressions that were developed.

Regular Expression Classifier to Identify Statistical Methods or Data Analytic Techniques

This regular expression takes a string of text (ideally, a sentence) and loops through a list of strings to identify if and which statistical method or data analytic technique were mentioned in each sentence. The vector `ana.meth` saves which statistical method or data analytic technique is mentioned first within each string for each item in the list.

```
ana.meth<-character(length(sent))
for(i in 1:length(sent)){
if(grepl("*linear*.regression|multiple*.regression",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"linreg"
} else if(grepl("anova|analysis*.variance*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"anova"
} else if(grepl("z.test|t.test",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"zttest"
} else if(grepl("correlat*",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"correlation"
} else if(grepl("chi*.square|x2",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"chisquare"
} else if(grepl("structur*.equation*.model*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"SEM"
} else if(grepl("logistic*.regression",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"logreg"
} else if(grepl("factor*.analys*.",sent[i],ignore.case=TRUE)==TRUE){
```

```
ana.meth[i]<-"factanal"
} else if(grepl("principal*.component*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"pca"
} else if(grepl("ancova|analysis*.covariance*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"ancova"
} else if(grepl("multilevel*.model*.|mixed*.effect*.model*.|hierarchic*.linear*.model*.
|nest*.model*.|mixe*.model*.| random*.coefficien*.model*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"mlm"
} else if(grepl("manova|multivariate*.analysis*.variance*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"manova"
} else if(grepl("mann*.whitney",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"mannwhit"
} else if(grepl("meta*.analysi*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"metaanal"
} else if(grepl("discriminant*.function*.|discriminant*.analysis",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"dfa"
} else if(grepl("canonical*.correlation*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"cannoncorrel"
} else if(grepl("mancova|multivariate*.analysis*.covariance*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"mancova"
} else if(grepl("z*.test*.dependent*.correlation",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"ztestdepcorr"
} else if(grepl("decision*.tree|c4.5",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"dectree"
} else if(grepl("k*.means|clustering",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"kmeansclust"
} else if(grepl("support*.vector*.machine*.|svm",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"svm"
} else if(grepl("apriori*.|association*.rule*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"aprioriassocrule"
} else if(grepl("em*algorith*|mixture*.model*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"emalgomixmodel"
```

```
} else if(grepl("page*.rank",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"pagerank"
} else if(grepl("adaboost|ensemble*.method*.|ensemble*.learn*.",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"adaboostensemble"
} else if(grepl("kNN|k*.neares*.neighbor",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"knn"
} else if(grepl("naive*.bayes|nave*.bayes",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"naibay"
} else if(grepl("cart|classifi*.regress*.tree",sent[i],ignore.case=TRUE)==TRUE){
ana.meth[i]<-"cart"
} else{
ana.meth[i]<-0
} }
```

Regular Expression to Identify Variables of Interest

This classifier takes a list of strings and identifies whether the sentence mentions a variable that is explicitly treated as an explanatory or response variable.

```
idv<-character(length(sent))
for(i in 1:length(sent)){
if(grepl("independent+.variable+.|predictor+.variable+.|regressor+.|
covariate+.|control+.\\bvariable+.|manipulated variable+.|explanatory variable+.|
exposure variable+.|risk factor+.|input variable+.", sent[i],ignore.case=TRUE)==TRUE{
idv[i]<-"ind"
} else if(grepl("dependent+.variable+.|response+.variable+.|regressand|predicted variable+.|
measured variable+.|explained variable+.|experimental variable+.|responding variable+.|
outcome variable+.|label", sent[i],ignore.case=TRUE)==TRUE){
idv[i]<-"dep"
} else{ idv[i]<-"0"
} }
```

Regular Expression to Identify Software Implementations

This classifier takes a list of strings and identifies if and which some of the most popular statistical and data analytic softwares are identified in the sentence. The output vector `soft` identifies the specific software mentioned.

```
soft<-character(length(sent))
for(i in 1:length(sent)){
if(grepl("^R$",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"R"
} else if(grepl("stata",sent[i],ignore.case=TRUE)==TRUE{
soft[i]<-"Stata"
} else if(grepl("SAS",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"SAS"
} else if(grepl("SPSS",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"SPSS"
} else if(grepl("Matlab",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"Matlab"
} else if(grepl("Weka",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"Weka"
} else if(grepl("Mplus",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"mplus"
} else if(grepl("Python",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"python"
} else if(grepl("Rapid*.miner",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"rapidmine"
} else if(grepl("JMP",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"JMP"
} else if(grepl("Minitab",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"minitab"
} else if(grepl("GraphPad*.Prism",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"graphpad"
} else if(grepl("Apache Hadoop|Hadoop",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"hadoop"
```

```
} else if(grepl("Statistica ",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"statistica"
} else if(grepl("Java",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"java"
} else if(grepl("systat",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"systat"
} else if(grepl("statgraphics",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"statgraphics"
} else if(grepl("^C$|^C\\+\\+$|^C#",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"CorC++orC"
} else if(grepl("Fortran",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"Fortran"
} else if(grepl("Apache Spark",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"ApacheSpark"
} else if(grepl("Caffe",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"Caffe"
} else if(grepl("Apache Mahout",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"ApacheMahout"
} else if(grepl("Tensorflow",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"tensorflow"
} else if(grepl("IBM Watson",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"IBMWatson"
} else if(grepl("KNIME",sent[i],ignore.case=TRUE)==TRUE){
soft[i]<-"KNIME"
} else{
soft[i]<-"0" } }
```