Responsible Integration of Behavioral Science in Computer Science Research and Development

Elizabeth M. Niedbala[1] and Kimberly J. Ferguson-Walter[2,*] and Dana S. LaFon[3]
[1]Department of Defense
[2]Laboratory for Advanced Cybersecurity Research
[3]National Security Agency
[*]Corresponding author: kimberly.j.ferguson-walter.civ@mail.mil

Abstract

Cross disciplinary research is essential for technological innovation. For decades, computer science (Comp Sci) has leveraged behavior science (Behav Sci) research to create innovative products and improve end user experience. Despite the natural challenges that come with cross disciplinary work, there are no published manuscripts outlining how to responsibly integrate Behav Sci into Comp Sci research and development. This publication fills this critical gap by discussing important differences between Behav Sci and Comp Sci, particularly with regard to how each field fits under the umbrella of science and how each field conceptualizes data. We then discuss the consequences of misusing Behav Sci and provide examples of technology efforts that drew inappropriate or unethical conclusions about their behavioral data. We discuss in detail common errors to avoid at each stage of the research process, which we condensed into a useful checklist to use as a tool for teams integrating Behav Sci in their work. Finally, we include examples of good applications of Behav Sci into Comp Sci research, the design of which can inform and strengthen digital government, e-commerce, defense, and many other areas of information technology.

1. Introduction

As technology has advanced it has become more common to blend behavioral science (Behav Sci) with computer science (Comp Sci) to create innovative products, improve end user experience, and solve real world issues that arise between humans and technology. This is particularly evident within digital government, e-commerce, and defense, as they seek to advance research and development in areas such as artificial intelligence and cyber security. Crucial to this merging is the comprehensive understanding of each field by the other to ensure that neither is misconstrued, misinterpreted, or misguided in their conclusions. This paper discusses how to responsibly use Behav Sci at each stage of research, the pitfalls that may occur during cross disciplinary research, and how to avoid them by addressing the following research questions: 1) What are key differences between Behav Sci and Comp Sci that can help researchers needing to integrate them? 2) What are the consequences of misusing Behav Sci? 3) What does it mean to responsibly integrate Behav Sci and Comp Sci? 4) What are common errors made when integrating Behav Sci into Comp Sci research and practice? 5) What checklist can technologists responsibly apply Behav Sci to their work?

When discussing responsible research practice, it is important to outline what is meant by *responsible.* The focus of our discussion is not on research misconduct (e.g., fabrication, falsification, or plagiarism), but instead on avoiding honest errors, biases, and pitfalls that can occur when integrating a different field into one's own expertise area. In essence, *responsible* research is thoughtful, rigorous, and high quality research.

1.1. Related Work

It is common for researchers to publish guidance on how to conduct high quality empirical research within their respective fields. Manuscripts outlining standards for research practice and common biases to avoid exist across many fields including the behavioral sciences [1, 2], public health [3], biomedical sciences [4], consumer research [5], and marketing [6]. However, it is less common to find research guides explaining how to responsibly integrate two scientific fields. Searching for research guidelines on integrating Behav Sci and Comp Sci is difficult. Keyword searches yield papers describing technology and information systems as a part of human culture [7], humans being the weakest link in cyber security [8], and using Behav Sci to enhance IT usage in healthcare settings [9]. Even though for decades researchers have been conducting multidisciplinary research that merges Behav Sci with Comp Sci, we failed to find a comprehensive guide on how to responsibly do so. There are, however, some papers that come close to, and provide key pieces of, the type of research guidelines needed.

For instance, there is a thorough guide for conducting empirical research in software engineering [10] by researchers who have a long history of

URI: https://hdl.handle.net/10125/79611
978-0-9981331-5-7

advocating good research practice in technological fields. The authors outline standards to follow at each stage of the research process and include practical examples of good and bad research practice. Their recommendations are a valuable resource for any team that does not have extensive experience conducting or reporting empirical research. However, it is not sufficient for a team that wants to apply Behav Sci principles to their work. Other researchers advocate leveraging Behav Sci to improve cyber security and describe how psychological principles can make technology more effective [11]. While they detail the specific benefits of applying Behav Sci to Comp Sci, responsible integration is not a focus. Finally, there are papers that describe the risks, benefits, and institutional review board (IRB) process for computer security research involving human subjects research (HSR) [12, 13]. They discuss specific characteristics of computer security research that can complicate the ethical treatment of human subjects, and how to navigate these challenges. We encourage scientists to review essential ethics-focused papers like these, since a thorough review on the ethical treatment of research subjects is beyond the scope of this paper.

Instead, our paper will discuss common principles and errors that computer science researchers and developers should be aware of when applying Behav Sci to their work. We encourage scientists from each field to speak one another's language and to be conscious of the unique considerations that must be made when integrating two fields. Multidisciplinary research is challenging, but the challenge should not get in the way of high quality research.

2. What is Behavioral Science?

Behav Sci is a scientific discipline in which the actions and reactions of humans and animals are studied through observational and experimental methods (American Heritage Dictionary, 2018). In short, it is the scientific study of human behavior. It employs research methods, that is, systematic procedures for gathering information and assessing research questions or hypotheses. Behav Sci practitioners use research results to make decisions, such as treatment modalities (e.g., clinical psychology), design systems to be compatible with human function (e.g., human factors), and recommend policy changes to improve functioning in work settings (e.g., industrial-organizational psychology).

2.1. Behavioral Science is Science: Part of the 'S' in STEM

Science, by definition, is empirical [14]. Thus, answers are obtained by using systematic, structured observation (i.e., data collection) using empirical validation, that is, verified by observation as opposed to logic or theory [15]. An idea may be obvious, intuitive, or broadly accepted, but it requires empirical validation to be scientifically accepted [14]. The methodology for determining this validation, which is core to all sciences, is the scientific method [16]. As such, the entire scientific community shares the principles of the scientific method.

The core of what makes Behav Sci a science is this application of the scientific method providing measurable characteristics. Behav Sci applies the scientific method which requires a rigorous scientific research design to collect data in a manner that lead to accurate conclusions, repeatability, and generalizable results. As in other fields, peer-review validates academic work to ensure poor methods and incorrect conclusions are not proliferated in the field. It is critical to discern peer-reviewed from non-peer reviewed publications. Peer-review enables the field to self-regulate, maintaining a level of standard to ensure validity and reliability of results and accountability for good scientific practices. Some responsibility for publication of problematic research falls on those who take part in the review process; this critical task is not always given the effort required to avoid unsatisfactory outcomes. In summary, Behav Sci is empirically based, employs the scientific method, involves systematic collection of data, produces valid and reliable studies using representative samples, analyzes empirical datasets for statistical significance, sets quality standards within the field, and uses its science as a basis for practitioner solutions.

Myth of *Soft* versus *Hard* Science. In layman's terms, "hard" sciences are notionally described as fields in which constructs are measured with high confidence and precision (e.g., measuring weight), whereas "soft" sciences are thought of as fields where there is difficulty in establishing measurable criteria (e.g., measuring social influence). Regardless of the type of construct being measured, *any* field that employs the scientific method is a science. Likewise, all sciences, including so-called "hard" sciences, are vulnerable to confounding variables, bias, threats to internal validity, threats to external validity, and other flaws.

Unfortunately, the colloquial distinction is in the assumption that "hard" sciences are inherently more difficult, use true scientific principles [17], and are judged to be more significant, meaningful, or important. Conversely, the label of "soft" implies that the behavioral sciences are easier and are neither as rigorous nor as accurate as other sciences. Those outside of Behav Sci often refer to the social sciences as "soft", provoking a perception that rigor and objectivity are absent. Perhaps being unaware of the strong dependency on the scientific method within the field or the reliance on empirical work foundational to psychological practice might explain this inaccurate perception. What is essential to understand is that what Behav Sci produces is not a final fact that is indisputable and without caveats (neither is the case for so-called “hard” sciences). They produce scientifically drawn conclusions that support or refute falsifiable hypotheses and contain a certain parameter of error. In other words, science[1].

3. Comp Sci Research and Development

There are many established definitions of computer science [18] and we will not attempt to redefine it here. What is key to this discussion, is that within the sciences, Comp Sci is less mature due to the invention dates of the technologies, and debates still occur questioning whether Comp Sci is science or engineering. “Is Computer Science Science?” argues that Comp Sci “meets every criterion for being a science, but it has a self-inflicted credibility problem [19].” The author suggests a fundamental issue arises from the term “computer science” being all-encompassing for a field that is inherently multidisciplinary. It includes engineering (e.g., programming), art (e.g., gaming), technology (e.g., system administration), and other aspects of “computing” that may not always need the scientific process. Likewise, there are sub-areas of computing for which integration with Behav Sci is rare and unnecessary, such as hardware architectures, formal methods, or those focused on the theory of computation. Many subfields of Comp Sci, such as, artificial intelligence (AI), cybersecurity, and human computer interaction (HCI), do have experimental components that can often benefit from integration with Behav Sci. AI researchers focus on using automated reasoning, knowledge representation, and learning to fully automate tasks or create human-machine teams able to outperform what was previously available. Cybersecurity researchers attempt to understand possible threats to networks and computer systems, solutions to keep them secure, and the humans that defend/attack them. HCI researchers examine ways humans interact with technology, striving to improve the interaction, and working towards improved efficiency or other benefits. Comp Sci research focuses on answering innovative questions, but the actual development of the technology is a next necessary step. When applying Behav Sci research or techniques, both research and development must understand how to do so responsibly.

An empirical evaluation of peer-reviewed Association for Computing Machinery (ACM) publications in 1993 noted that 40% of those with claims that needed empirical support had none at all, compared to 15% in a non computer science journals [20]. There continues to be efforts across the Comp Sci community to collect artifacts (e.g., data and code) to reduce unvalidated research claims. And there are clearly research teams within Comp Sci that do not fall into the traps discussed in this paper. However, while many conferences are publishing guidelines on how to perform (and document) Comp Sci research, there is still some confusion and debate. For example, in April 2021, the University of Minnesota was banned from making further contributions to the open-source Linux kernel project after one of their professors carried out research retroactively approved by their IRB as exempt [21]. This exemption did not consider the potential negative effects on the Linux maintainers or contributors (who were not asked to opt-in as participants). The explicit goal of the now controversial experiment, as the researchers have since emphasized[2], was to improve the security of the Linux kernel by demonstrating to developers how malicious code could be introduced into the repository. The paper was accepted to a top computer security conference, but later withdrawn by the authors as a part of making amends to the community.

Comp Sci researchers are not the only ones to fall into the mistakes identified in this paper, such as lack of well-documented rigorous analysis, and reproducible studies. Many fields would benefit from more careful application of the scientific method. Peer-reviewed research papers have stated that most published research findings are wrong [22]. Fields adept at experimental design and HSR such as Behav Sci and medicine publish papers with biased research studies, a lack of power or generalizability of results, and that are not always reproducible [23]. For fields (such as Comp Sci) that are reluctant to publish negative results and that do not reward replication studies, even more issues can be expected.

4. Consequences of misusing Behav Sci research

Cross disciplinary researchers and developers should recognize that the word *data* can be interpreted differently between fields. Data in Behav Sci is generally the result of systematic collection of observations scientifically designed to be a representative sample of a larger population. Thus, analytical conclusions about that dataset can be generalized to that larger population with a statistical degree of accuracy. Data in Comp Sci is broader and can describe any collection of information that was not necessarily gathered in a systematic or scientific way (e.g., an acquired dataset of search results). How researchers understand the term “data” can determine what conclusions they attempt to draw from that information, many of which may be

[1] While it may be a common misconception that Behav Sci in not part of STEM, the Department of Defense STEM scholarship includes “Cognitive, Neural, and Behavioral Sciences.” However in 2021, only 2.2% of applicants (with 1% awarded) while “Computer and Computational Science and Computer Engineering” made up 20.8% (with 26.7% awarded).

[2] https://www-users.cse.umn.edu/~kjlu/papers/clarifications-hc.pd

erroneous. For example, conclusions about an acquired dataset cannot be generalized to a larger population because it is not a representative sample.

This paper attempts to highlight common misapplications of Behav Sci to Comp Sci research and development and provide guidance to avoid them. For instance, in Section 4.1, Example 1 illustrates how a small yet significant effect can be pitched as a promising real-world intervention, despite high quality scientific evidence against it. Inaccuracies in interpretation of or representation of these data also have direct impact on the soundness, usability, and benefit of the technology. Care must be taken when algorithmically representing Behav Sci study results, at the risk of ultimately creating inaccurate models and predictions. In Section 4.1, Example 2 discusses the discovery of biased and prejudiced algorithmic results caused by training models and algorithms on biased human data. Decisions based on these algorithmic results can affect the lives of those associated with those results (e.g., finding terrorists in big data, triaging medical or mental health patients, prioritizing loan applicants, making employment selections, vetting security risks, etc.). In other cases, damage to the field of Behav Sci can be significant in that, if the algorithm fails when Behav Sci is incorporated incorrectly, the assumption is that the behavioral or social science is inaccurate or faulty (i.e. "soft science" is to blame).

4.1. Misapplication of Behav Sci to Comp Sci Research/Endeavors

Example 1: Using Pseudoscientific Theories. Pseudoscience exists in every scientific field, including the behavioral sciences. There are many misconceptions about human psychology that are largely unsupported by high quality data, yet have become popularized through the general public, the media, and bad practitioners. One example of a popular misconception is that individuals have unique learning styles. In learning styles theory, the infamous meshing hypothesis predicts that an individual performs better when instructional information is presented in the mode that matches his or her learning style. The problem with this hypothesis is that there is virtually no evidence that matching instruction to one's supposed learning style is effective (see [24] for a review). Pashler et al. [24] define the specific statistical evidence needed to claim support for the meshing hypothesis, which virtually no learning styles studies have found. While it is true that the majority of the population prefer learning in certain ways [25], the experimental evidence that matching causes better learning is tremendously weak, and many reputable cognitive and behavioral scientists consider learning styles a myth [24, 26, 27].

One group of cyber security researchers decided to use the meshing hypothesis as a foundation for their research on improving cyber security training. They tested if matched learning styles instruction was correlated with better information security awareness (ISA) for employees [28]. They surveyed 1,048 adults about their preferred style of learning cyber-security material (i.e., preferred method), about the types of information security training they had received at work in the past (i.e., received method), and about their knowledge, attitudes, and behaviors regarding information security (i.e., security awareness). The extent to which subjects preferred method and received method matched was hypothesized to predict scores on the ISA questionnaire. They found a small, positive correlation between training match and ISA (although gender and age were each stronger predictors of performance than matching) and the authors claimed that the meshing hypothesis was therefore supported (even though the results did not meet standard statistical criteria [24]). Based on the weak correlational findings, they concluded that companies could save time and money by tailoring their ISA training to groups of employees' preferred learning styles. On the contrary, based on the majority of rigorous experiments on learning styles, it is possible that significant time and resources would be wasted through such an endeavor. The danger of misusing behavioral science in this example is that a real-world intervention was proposed even though it was not based on sound theoretical grounding, meaning that customers could waste significant time and resources implementing new policies that are unlikely to have an effect. Ideally the research team would have based their experiment on a stronger theoretical foundation and discussed the practical significance of their results (i.e., effect size) so as to manage expectations for those seeking to use their intervention.

Example 2: Bias in Algorithms. Researchers have identified many issues with fairness in artificial intelligence (AI) models and algorithms [29, 30, 31, 32]. Gender, racial and other biases, are being propagated, usually unwittingly, through an irresponsible use of Behav Sci or statistical data. For example, ProPublica analyzed the risk assessment scores produced by a recidivism algorithm used to help determine sentencing and determined the formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. When isolating the effect of race from criminal history, age, and gender, black defendants were still 77% more likely to be pegged as higher risk of committing a future violent crime and 45% more likely to be predicted to commit a future crime of any kind [33]. White defendants were also mislabeled as low risk more often than black defendants. The for-profit company that created the product disputes the analysis, but will not disclose the proprietary details of the formula. Proponents of these kinds of tools note it is difficult to "construct a score that doesn't include items that can be correlated with race—such as poverty, joblessness and social marginalization. If those are omitted from your risk assessment, accuracy goes down," admitting that the model may be biased, but not taking responsibility for their model's lack of representation of the real-world.

The AI model and the people who created it may not be biased; bias can simply be built into the data on which the model was trained. This issues is blatant in the recidivism example, but there are more misapplications of data in other AI efforts, such as the training data provided to Uber's self-driving car to teach it how to react to pedestrians. After a 20-month investigation following the death of a woman in Arizona who was hit and killed by a self-driving Uber car, the National Transportation Safety Board (NTSB) released a report finding that the automated car lacked the capability to classify an object as a pedestrian unless that object was near a crosswalk, likely due to incomplete training data (`https://www.ntsb.gov/news/events/Documents/2019-HWY18MH010-BMG-abstract.pdf`).

Misapplying Behav Sci can be dangerous because when the data feeding the algorithm is biased, misunderstood, or misapplied, disastrous results can occur. Much AI research uses publicly available data for training, whether it be image, text, or historical event data. It is the responsibility of each researcher to adequately vet the data and its planned usage, and clearly lay out any limitations and caveats. Researchers have begun performing due diligence and discovering and announcing limitations and biases in common data sets, such as: "unequal representation and gender stereotypes in image search results for occupations" [34]; "Man is to computer programmer as woman is to homemaker: debiasing word embeddings" [31].

Example 3: Mistaking Correlation for Causality. Data science is an interdisciplinary field that overlaps with Comp Sci and that can fall into some, if not more, pitfalls. One data scientist claimed that *"Google searches are the most important dataset ever collected on the human psyche"* (Steven-Davidowitz, 2017)[3]. He argued that the data he analyzed from Google Trends can provide more accurate findings because people are incentivized to be honest when they do Google searches—as opposed to surveys or other social media sites [35]. The author attempts to answer questions such as: *"How many American men are gay?"* and *"Which state is the most racist?"* Google's data likely has high ecological validity, given that Google collects data from people not otherwise willing to provide information for research. However, making claims about human psychology or human behavior solely based on trends of internet searches is precarious—no matter how *big* the data available is. As stated above, this kind of data is not a statistically accurate representation of the population. Unless caution is used, other data scientists may fall into several traps as seen in this work.

For instance, the analysis presented has the tendency to frame correlational observations as causal phenomena. However, correlation does not imply causality. In non-randomized non-controlled data collection like Google searches, there are many co-variates that are not accounted for and the methodology does not provide real explanatory power. Ignored are potential differences in human intentions behind each search. For example, some people's Google searches may be representative of a morbid curiosity, perhaps tempted by an outrageous phrase suggested by the auto-complete, rather than actual ground truth representation of self. Bots and trolls, stirring up controversy, can also skew such a data source. Moreover, no statistical analysis was performed to determine if the differences detected were meaningful. Results were provided in percents and calculated based on raw number of searches—not accounting for multiple searches per person. Simply examining counts or percentages can be misleading.

For data scientists that do perform statistical analysis, p-hacking can be a concern since exploratory data analysis (EDA) often leads to an interesting finding in the data which is then interpreted and reported [36]. But EDA on *big* data can take you into the realm of *the law of multiple comparisons*, where statistically significant results that are meaningless will be discovered if enough comparisons are performed [37]. One solution is to have documented research hypotheses up front, such that you always have a justification to examine/compare particular data [38].

5. Responsible Integration of Behav Sci in Comp Sci Research/Practice

As challenging as it may be, Comp Sci researchers and developers should strive to avoid the common pitfalls and ethical issues that come with using Behav Sci. The research process occurs in stages, and each stage is uniquely vulnerable to bad practice. Entire textbooks, courses, and degree programs are dedicated to teaching the research methods and statistical analyses applicable to one's own field, and we encourage scientists to seek out the requisite education before conducting empirical research. However, even the most experienced scientists have blind spots, particularly when integrating theories, methods, and analysis techniques from another area. The best way to avoid pitfalls in cross disciplinary research is to work with a multidisciplinary team of researchers that possess the relevant expertise for each area of interest. However, it

[3] www.facebook.com/decepticon2017/videos/1967571623516839

is also important for each team member to be conscious of the errors that can arise during cross disciplinary research. For this reason, we highlight some specific blind spots to which computer scientists and developers may be vulnerable when conducting research involving Behav Sci.

5.1. Countering Errors at Each Stage of the Research Process

Stage 1: Make an Observation/Ask a Question/Identify a Problem. The first stage of the research process is to make an observation about the world, find a question to answer, or identify a problem to solve. This process occurs rather intuitively and does not require expertise in itself. Humans are curious beings and are naturally good question-askers. And researchers should not be limited in the questions they can ask—a computer scientist should feel free to wonder how a human interacting with technology thinks, feels, learns, and behaves under different conditions. What is crucial to this process, though, is that the assembled research team is expertly equipped to answer the question. If it is related to human behavior (the term "behavior" including emotion, cognition, performance, etc.)–whether group or individual—a behavioral scientist should be involved.

Stage 2: Conduct Background Research. The second stage of the research process is to conduct a literature search of the relevant scientific body of research. It is not sufficient to move directly from "ask a question" to "generate a hypothesis" without first researching the underlying theories that explain why humans might behave in a certain way.

One pitfall to avoid during literature searches is blindly accepting pseudoscientific theories. Unfortunately, there are many misconceptions about human psychology that are largely unsupported by high quality data, yet persevere because they seem attractive, surprising, or sexy to the general public (for a review, see [39]). Psychological theories that are largely considered unsupported include the existence of learning styles, the fact that opposites attract in romantic relationships, and the idea that some people are left-brained while others are right-brained [39]. One way to avoid pseudoscience is by combating confirmation bias in literature searches. Specifically search for evidence for and against the theory and thoughtfully consider the evidence from both sides. Also be conscious of the journal quality and reputation from which the evidence originates. Only rely on research from high quality, peer-reviewed journals. Behavioral scientists are trained to evaluate the quality of human subjects research and are able to spot predatory journals in their field[4], so they will be invaluable during this step of the research process.

Stage 3: Generate a Hypothesis. If theory explains *why* you believe an effect will occur, the hypothesis states exactly *what* you expect will occur. A strong hypothesis defines the experimental units clearly (even for subjective or difficult to observe human experiences such as cognition, emotion, and performance) and indicates the direction of the predicted effect relative to a comparison (if no direction is predicted, explain the reasoning). There should be a specific hypothesis for every measure, and each hypothesis must be falsifiable. It has been observed that computer science theorists rarely generate falsifiable theories, even though falsifiability is a crucial part of the scientific method [20].

Statistics is its own field, and the way various fields use it requires a deeper understanding and interpretation. For example, p-values do not explain what is the probability that the null hypothesis is true given the data, but rather what is the probability of observing these or more extreme data given the null hypothesis is true. Collaborating with someone who holds a thorough understanding of falsifiability, null hypotheses, alternative hypotheses, alpha, beta, power, Type I and II error, critical values, and p-values is essential.

Stage 4: Collect/Assemble the Data. Before data collection or assembly begins, the ethics of collecting and using human data must be considered. Because Comp Sci research is often primarily concerned with systems and processes, researchers may not be aware they should follow existing ethical guidelines for HSR [13]. We believe that all researchers and developers are ethically responsible for the outcomes of their research. Therefore, teams should put protections in place to minimize the risk of harm to human subjects during data collection, and to use any findings derived from human data in an ethical manner.

After ethics have been considered and before data collection begins, researchers should use a systematic approach to determine the sample size needed to detect the effect. The basics of power analysis and sample size are reviewed in introductory statistics material and assist researchers in designing well-powered studies. Some psychological effects are notoriously small and difficult to detect. The typical effect size differs substantially between subfields of psychology, likely due to differences in instrument reliability [40]. Be aware of the typical effect size for the methods that will be used, and calculate the sample size accordingly. The consequence of conducting an under-powered study is that significant time, money, and resources are wasted running a study that has a low probability of finding an effect if one actually exists. Essentially, conducting an under-powered study sets the team up for failure before even beginning.

Data collection and assembly should follow a well thought out plan and the details of sampling

[4] Predatory journals take fees from authors, provide no peer review, and do not follow appropriate academic publication standards (Research Medical Library, 2020).

and data collection should be thoroughly reported in the manuscript. Include an expert in research methodology or experimental design on the team to ensure that the study is well designed and conducted in a controlled manner. For instance, experimental studies should follow appropriate sampling methods, employ randomization, include thoughtful comparison groups, and control for threats to validity and reliability (e.g., selection bias, regression to the mean, attrition, testing effects, etc.). The study procedure, sampling method, instrumentation choice, randomization process, and any data collection anomalies should be reported with enough detail that another scientist can evaluate and reasonably replicate the study methods (behavioral scientists adhere to the APA Style Journal Article Reporting Standards [JARS] for this reason).

Stage 5: Analyze the Data. A common issue in computer science research and development is the lack of detail in reporting data analysis methods, which is likely due to the strict page or word limits on conference and journal submissions. In the data analysis stage, the research team should include a specialist in statistical methods, and the team should thoroughly explain the reasoning behind all data processing and analysis decisions. When analyzing Behav Sci data, it is standard to report the descriptive statistics for the sample, the number of subject drop outs and data point exclusions, and effect sizes for all main analyses (among other details; see the APA JARS for suggested guidelines on reporting Behav Sci data). Do not commit fundamental errors of data interpretation, such as using causal language to describe correlational relationships.

Stage 6: Draw Conclusions, Refine Hypothesis. By the last stage of the research process, the data has been analyzed and conclusions can now be drawn about several things. Discuss whether the data provide evidence for or against the theory the hypotheses were grounded in. Highlight any data anomalies or research practices that may have limited the ability to cleanly test research questions (e.g., be aware of threats to internal validity), and discuss the study limitations thoroughly. It is important to remember that scientific findings are not indisputable facts; each study describes a probabilistic outcome of one particular sample of observations under one set of circumstances within one time frame. Many behavioral scientists view each study as a single data point that either supports or fails to support a larger theoretical idea.

Generalizability of the findings to a population (i.e., external validity) should also be discussed. Be conscious of the sample tested and understand that the data based on a sample should not be generalized to populations that were not represented in the sample. Similarly, do not commit the ecological fallacy, which is the erroneous belief that simple conclusions can be drawn about an individual based on group data (for discussions on the ecological fallacy see [41, 42]).

Finally, it is beneficial to discuss if the findings have practical implications for the real world (i.e., ecological validity). If the research was conducted in a laboratory or another tightly controlled setting, the findings may not easily generalize to applied settings. If the effects were significant but small, do due diligence and highlight this for the reader. Understanding ecological validity is particularly important for teams that make recommendations for real-world interventions, training programs, or policy changes. Too many researchers (including behavioral scientists) exaggerate the real-world impact the interventions will have without first conducting replications, investigating important boundary conditions, or testing them in a field setting.

5.2. Good Examples

Example 1: Good Experimental Methods. The Tularosa Study was a cybersecurity effort that used human experts to test the impact a specific cyber defense has on cyber attacks [43]. It was one of the largest experiments of its kind. It took over a year to run the 138 participants and over 1611 GB of data were collected. The first publication focused on the experimental design and methodology of the HSR and a follow-up paper presented the statistical data analysis performed to address the hypotheses [44]. It is important to note, that in order to include the necessary details in the methodology paper and stay within the 10-page limit for the conference, an online appendix [45] was also created which provided additional information including: A) Individual Measures: Exact wording of Questions and questionnaires provided for participants to answer; B) Task Briefing: Exact wording of scenario, instructions, and rules provides to participants; C) Schedule: Hourly schedule of the 2-day experiment. Combining these all into one paper unfortunately would have meant omitting important details. Comp Sci differs from Behav Sci research in that, conference publications are a meaningful metric; top conferences have a lengthy peer-review process and paper acceptance rates in the teens. Journal publications, while more lenient on page length, are often too slow for the fast pace of Comp Sci. These publications indicate good experimental methods were used for this research, and crucially, were explicitly described in detail. Some noteworthy examples include:

- **Hypotheses:** Research hypothesis were developed in advance and experimental design was based off of the research conditions.
- **Interdisciplinary:** Behavioral scientists were included in the research from day one.
- **Experimental Conditions:** Each participant is in a separate condition for the *between participants* comparisons, which avoids any learning bias.
- **Control Condition:** A separate control condition was executed. While this is sometimes mistakenly

viewed as a “waste” of participants, this baseline is needed to measure what difference the experimental manipulation actually makes.

- **Construct Validity:** Included a survey questions to confirm that the experimental manipulation had been successful.
- **Participants:** Details about participant expertise and the recruitment process was provided; this helps justify the generalizability of results.

Details regarding the IRB approval of the HSR and steps taken to protect the participants and their anonymity were described. Many top Comp Sci conferences require these details for publication. The authors followed up with a lessons learned paper providing detailed information on the internal versus external validity trade-offs made in the experimental design, and the limitations of the study, in hope of aiding future designs [46]. Examples are summarized below:

Internal validity: Same proctors, reading from standardized script; Time on task monitored, cataloged, and regulated, including breaks and lunch; Participants were proctored and not allowed to discuss the task with each other; All participants per condition were presented identical copies of cyber range; Cross conditions copied as closely as possible; Required all participants use provided laptop with standard set of cyber attack tools to ensure equality across participants.

External validity: Required use of standard tools could have hampered performance for those not used to that toolset; Having participants work alone, rather than on teams, could have hampered performance; Broad participant population was a random sample of U.S. professional red teamers (then randomly assigned to conditions); schedule was tightly controlled and duration was shorter than real engagements; Red teamers differ from the unethical hackers we need to protect against; specific differences are not well-documented.

The data analysis paper provided details on which hypothesis each analysis addressed, the statistical method used, and which data were used and why [44]. Quantitative findings were bolstered with observational data from participant self-reports. Limitations were discussed, followed by caveats that further experimentation is needed and justification for the generalizability of the study results.

Example 2: Good Reporting Practice. Another example of successfully integrating Behav Sci into computer science research is an experiment testing the effects of affective framing and control on privacy behavior [47]. The authors identified a clear real-world issue during the sign-up process for new online services: Users often have negative associations with privacy notices, feel a lack of control over their privacy, and may not fully comprehend the privacy notice before consenting. In their study, they investigated if the visual design of privacy notices and the level of control given by the notices influenced the user’s affect, privacy comprehension, and privacy disclosure. Their paper thoroughly reported their reasoning behind every research decision they made and meticulously documented all of their research methods and analyses. Moreover, they thoughtfully discussed the practical and theoretical implications of their findings while highlighting limitations within the study. We view this paper as a strong example for how research teams should document and report each stage of the research process. Some noteworthy examples from each stage include:

- Conduct background research: The authors relied on established models to shape their specific research questions, many of which were reputable theories widely accepted by behavioral scientists. For any theories that were associated with contradicting evidence, they discussed findings from both sides.
- Collect/assemble the data: In their methods section, the authors clearly defined their research design, sampling methods, data collection procedures, and decisions behind every measure included. For instance, before analysis, they conducted principle component analyses on their dependent measures to verify that the items loaded correctly onto the expected factors. For any scales that did not, they either revised the measures or excluded them from further analysis.
- Analyze the data: All parametric assumption checks, descriptive statistics, main analyses, and effect sizes were reported. They explained why they chose each statistical technique; for example explaining criteria for using univariate or multivariate tests on groups of variables.
- Draw conclusions: In their discussion section, the authors drew practical conclusions for the real world without overreaching, and thoughtfully discussed the theoretical implications of their data. They included limitations to their study, and responsibly highlighted the presence of small effects despite statistical significance.

These are only two of many successful applications of Behav Sci into Comp Sci research. We hope our paper and attached checklist provide additional research teams with the knowledge necessary to responsibly integrate Behav Sci into their work, and the insights needed to boost new cross-disciplinary collaboration.

6. Summary

The integration of behavioral science into computer science is a natural and important step in advancing technology, particularly within the digital government and defense industry. Because the government refines public policy, oversees complex government operations, promotes security and privacy, and delivers public goods and services, a commitment to effective multidisciplinary research practice must be upheld. The government is responsible for the people it serves, and thus should invest in high quality cross-disciplinary practice between fields. This paper serves as a guide for those projects that involve integrating behavioral and computer science. We highlight the key differences between Behav Sci and Comp Sci, identify pitfalls of integrating Behav Sci into computer science at each stage of research, and offer suggestions on how to avoid those pitfalls to ensure responsible integration. These avoidance strategies have been summarized into a checklist (see Appendix A) for the reader's convenience. Core to integrating these fields is assembling a multidisciplinary team that understands the challenges of each involved discipline, correctly interprets the research results, and draws accurate conclusions about the data. By taking the steps outlined in this paper, both behavioral scientists and computer scientists stand to advance technology in a responsible, accurate, and meaningful way.

References

[1] P. Podsakoff, S. MacKenzie, J. Lee, and N. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, vol. 88, pp. 879–903, Oct. 2003.

[2] R. MacCoun, "Biases in the interpretation and use of research results ," *Annual Review of Psychology*, vol. 49, pp. 259–287, 1998.

[3] J. Ioannidis, S. Greenland, M. Hlatky, M. Khoury, M. Macleod, D. Moher, K. Schulz, and R. Tibshirani, "Increasing value and reducing waste in research design, conduct, and analysis," *Lancet*, vol. 383, no. 9912, 2014.

[4] D. L. Sackett, "Bias in analytic research," *Journal of Chronic Diseases*, vol. 32, no. 1, pp. 51–63, 1979.

[5] B. J. Calder, L. W. Phillips, and A. M. Tybout, "Designing Research for Application," *Journal of Consumer Research*, vol. 8, pp. 197–207, 09 1981.

[6] R. Petty and J. Cacioppo, "Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science?," *Journal of Marketing Research*, vol. 33, pp. 51–63, 1996.

[7] D. E. Leidner and T. Kayworth, "Review: A review of culture in information systems research: Toward a theory of information technology culture conflict," *MIS Q.*, vol. 30, pp. 357–399, June 2006.

[8] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link' ? a human/computer interaction approach to usable and effective security," *BT Technology Journal*, vol. 19, pp. 122–131, July 2001.

[9] R. Kukafka, S. Johnson, A. Linfante, and J. Allegrante, "Grounding a new information technology implementation framework in behavioral science: a systematic analysis of the literature on it use," *Journal of Biomedical Informatics*, vol. 36, no. 3, 2003.

[10] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, 2002.

[11] S. L. Pfleeger and D. D. Caputo, "Leveraging behavioral science to mitigate cyber security risk," *Computers & Security*, vol. 31, no. 4, pp. 597–611, 2012.

[12] M. Johnson, S. Bellovin, and A. Keromytis, "Computer security research with human subjects: Risks, benefits and informed consent," in *International Conference on Financial Cryptography and Data Security*, pp. 131–137, Springer, 2011.

[13] E. Buchanan, J. Aycock, S. Dexter, D. Dittrich, and E. Hvizdak, "Computer science security research and human subjects: Emerging considerations for research ethics boards," *Journal of Empirical Research on Human Research Ethics*, vol. 6, pp. 71 – 83, 2011.

[14] F. Gravetter and L. Forzano, *Research Methods for the Behavioral Sciences.* Cengage Learning., 6th ed., 2019.

[15] J. Simpson and E. Weiner, *The Oxford English Dictionary.* Oxford: Clarendon Press, 1989.

[16] H. Gauch, *Scientific Method in Brief.* Cambridge University Press, 2012.

[17] K. Reece, "She's going soft: A commentary on hard and soft sciences," 2014.

[18] G. Dodig-Crnkovic, "Scientific methods in computer science," 2002.

[19] P. Denning, "Is Computer Science Science?," *ACM*, 2005.

[20] W. Tichy, "Should computer scientists experiment more?," *Computer*, vol. 31, no. 5, pp. 32–40, 1998.

[21] M. Clark, "University of minnesota banned from contributing to linux kernel," *The Verge*, Apr 2021.

[22] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, August 2005.

[23] C. Camerer, A. Dreber, F. Holzmeister, and et al., "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015.," *Nat Hum Behav*, vol. 2, pp. 637–644, 2018.

[24] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork, "Learning styles: Concepts and evidence," *Psychological Science in the Public Interest*, vol. 9, pp. 105–119, 2008.

[25] L. J. Massa and R. E. Mayer, "Testing the ati hypothesis: Should multimedia instruction accommodate verbalizer-visualizer cognitive style?," *Learning and Individual Differences*, vol. 16, no. 4, pp. 321–335, 2006.

[26] C. R. Riener and D. Willingham, "The myth of learning styles," *Change the Magazine of Higher Learning*, 2010.

[27] P. A. Kirschner, "Stop propagating the learning styles myth," *Computers & Education*, vol. 106, 2017.

[28] M. Pattinson, M. Butavicius, B. Ciccarello, M. Lillie, K. Parsons, D. Calic, and A. McCormac, "Adapting cyber-security training to your employees," in *HAISA*, 2018.

[29] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, 2012.

[30] A. Guo, E. Kamar, J. Wortman Vaughan, H. Wallach, and M. Ringel Morris, "Toward fairness in AI for people with disabilities: A research roadmap," in *ACM ASSETS Workshop on AI Fairness for People with Disabilities*, ArXiv, 2019.

[31] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, pp. 4349–4357, 2016.

[32] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

[33] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.," *ProPublica*, May 2016.

[34] M. Kay, C. Matuszek, and S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations," in *CHI Conference on Human Factors in Computing Systems*, ACM, 2015.

[35] S. Steven-Davidowitz, *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are.* Dey Street Book, May 2017.

[36] S. B. Bruns and J. Ioannidis, "P-curve and p-hacking in observational research," *PLoS one*, vol. 11, no. 2, 2016.

[37] A. Gelman and E. Loken, "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time," 2013.

[38] J. Wicherts, C. Veldkamp, H. Augusteijn, M. Bakker, R. van Aert, and M. van Assen, "Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking," *Frontiers in Psychology*, vol. 7, 2016.

[39] S. Lilienfeld, S. Lynn, J. Ruscio, and B. Beyerstein, "Busting big myths in popular psychology," *Scientific American Mind*, vol. 21, no. 1, pp. 42–49, 2010.

[40] T. Schäfer and M. A. Schwarz, "The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases," *Frontiers in Psychology*, vol. 10, 2019.

[41] S. Piantadosi, D. Byar, and S. Green, "The ecological fallacy," *American Journal of Epidemiology*, vol. 127, no. 5, pp. 893–904, 1988.

[42] S. Schwartz, "The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences,," *Am J Public Health*, vol. 84, pp. 819–24, May 1994.

[43] K. Ferguson-Walter, T. Shade, A. Rogers, E. Niedbala, M. Trumbo, K. Nauer, K. Divis, A. Jones, A. Combs, and R. Abbott, "The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception," in *Hawaii International Conference on System Sciences (HICSS)*, 2019.

[44] K. J. Ferguson-Walter, M. M. Major, C. K. Johnson, and H. Muhleman, Daniel, "Examining the efficacy of decoy-based and psychological cyber deception," in *USENIX Security Symposium*, Apr. 2021.

[45] K. Ferguson-Walter, T. Shade, A. Rogers, M. Trumbo, K. Nauer, K. Divis, A. Jones, A. Combs, and R. Abbott, "Appendix to The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception," 2019.

[46] K. Ferguson-Walter, M. Major, D. Van Bruggen, S. Fugate, and R. Gutzwiller, "The world of CTF is not enough data: Lessons learning from a cyber deception experiment," *In Proceedings of First IEEE Workshop on Human Aspects of Cyber Security (HACS)*, 2019.

[47] A. Kitkowska, M. Warner, Y. Shulman, E. Wästlund, and L. A. Martucci, "Enhancing privacy through the visual design of privacy notices: Exploring the interplay of curiosity, control and affect," in *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[48] J. W. Kotrlik and H. A. Williams, "The incorporation of effect size in information technology, learning, and performance research," *Information Technology, Learning, and Performance Journal*, 2003.

A. Checklist for responsible behavioral science integration into computer science research and development

Integrating behavioral science (Behav Sci) research into computer science (Comp Sci) is valuable for the advancement of technology. Below is a checklist intended to assist those interested in such integration as to when and how to do so responsibly for higher quality scientific outcomes.

When to integrate a behavioral scientist:

- Does your research relate to human behavior, cognition, emotion, or performance, whether group or individual?
- Does your research involve or require a review of any Behav Sci literature?
- Does your data draw conclusions about human behavior or use humans for data collection?

How to integrate Behavioral Science responsibly:

- Collaborate with a qualified behavioral scientist. Find the right specialists to consult so your experiments and conclusions are solid.
- If you draw conclusions about human behavior, include a specialist in statistical methods for Behav Sci.
- Do not assume because you are human, you are a human scientist. Acquiring some background in Behav Sci is useful, however, without significant degreed study, this will not be enough to replace a Behav Sci expert. It is important to note that behavioral scientists should also seek out collaboration with computer scientists when Behav Sci studies lean into technological subjects.
- If unsure whether your research is integrating Behav Sci, consult an expert in the subspecialty area of Behav Sci in which you are interested. Creating and sustaining a professional relationship will not only assist in ensuring responsible integration but will encourage both disciplines to be open to and spark scientific curiosity.

Behavioral Scientists should assist the multi-disciplinary team in considering the following:

- Are effect sizes being properly interpreted and reported (see [48])?
- Understanding and applying the strength of statistical correlations accurately to avoid misguided conclusions.
- Is human subject research being conducted? If so, have standard ethical guidelines been considered?
- Comprehensive Behav Sci literature reviews are conducted on the Behav Sci construct being integrated into Comp Sci. This also avoids using a limited corpus of literature (i.e., only using one or two studies that support your hypothesis).
- Avoiding the confirmation bias by reviewing Behav Sci research both for and against the research hypothesis. This assists in avoiding biased inclusion, that is, selecting Behav Sci results that support your goal.
- Avoiding the ecological fallacy, that is, ensuring that aggregate data or the results of a study are not applied to an individual.
- Assist in determining if you have established validity and reliability for the product, tool, or algorithm with scientific rigor as well as articulating that rigor in academic publications.
- Avoiding the assumption that scientific conclusions are facts. They are a collection of statistical correlations and evidence based hypotheses.