Students Stumble With Inconclusive Results: An Exploratory Analysis on How Students Interpret Unexpected Results

Joss Ives, Aaron M. Kraft, James Day, and D. A. Bonn
Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, BC, V6T 1Z1

Success in inquiry labs often requires students to grapple with results that contradict their expectations. Previous work has shown that students who see the goal of the lab in terms of model confirmation rather than model testing struggle to engage in inquiry. Our study set out to extend this previous work by looking at the impact of asking students to hypothesize in a first year lab activity designed to produce unexpected results. Our exploratory analysis shows that hypothesizing does not play a major role in students interpreting their results correctly; instead, inconclusive results are the most significant factor in explaining correct student interpretations. We will show that these misinterpretations are not the result of model confirmation bias, but rather, misunderstanding the statistical nature of the results.

2023 PERC Proceedings edited by Jones, Ryan, and Pawl; Peer-reviewed, doi.org/10.1119/perc.2023.pr.Ives

I. INTRODUCTION

The first year lab sequence at our institution teaches students experimental skills through a scaffolded scientific inquiry process [1]. By iteratively making measurements, quantitatively comparing these measurements and then reflecting on this comparison, these labs have been shown to teach critical thinking skills [2]. In the second week of this ten week lab sequence, students are introduced to the t-score, a continuous scale allowing for quantitative comparisons between two measured values [3]. The t-score is framed as a metric quantifying how distinguishable two measured values with uncertainties are. Students are given instructions shown in Fig. 1 for information on how to interpret the t-score which is formally defined as:

$$t' = \frac{|A - B|}{\sqrt{(\delta A)^2 + (\delta B)^2}}. \quad (1)$$

Reflecting on this t-score is key to the iterative inquiry in these labs.

In the third lab of this sequence, the students use this t-score to compare the period of a pendulum at 10° and 20°. Many students come into this lab unaware of the small angle approximation used to derive $T = 2\pi\sqrt{L/g}$. The results of this lab often surprise students because a high precision measurement will uncover a breakdown in this model. These labs with unexpected results are thought to be useful for developing inquiry and modelling skills, because they force students to grapple with outcomes that contradict their own internal models [4, 5]. Forming scientific questions in response to these confusing results is a key aspect to scientific inquiry known as problematizing [6]. Therefore providing students with the opportunity to grapple with these results and supporting their success in this process is vital for the development of their scientific inquiry skills.

Previous work has shown that students struggle to problematize in these labs because they see the goal of the lab in terms of verifying rather than testing models [7–11]. This framework leads students to think that there is a specific outcome of the lab that they must achieve, leading them to ignore results that contradict that outcome. But in order to effectively grapple with contradiction, students must first interpret comparisons. In this example, interpreting their t-score correctly is necessary for problematizing. We set out to investigate how these model confirmation frameworks affect students' t-score interpretations, focusing on the role of hypothesizing. Our main research questions were:

1. How does asking a student to hypothesize impact their ability to interpret their t-score correctly?
2. How does a student's self reported surprise at the result impact their ability to interpret their t-score correctly?
3. What other factors impact their ability to interpret their t-score correctly?

While our research questions and motivations were focused initially on model confirmation bias, we found that this did not play a major role in explaining correct t-score interpretations. Instead, we found that inconclusive results lead students to misinterpret their results because of an underlying misunderstanding of the statistical nature of the t-score.

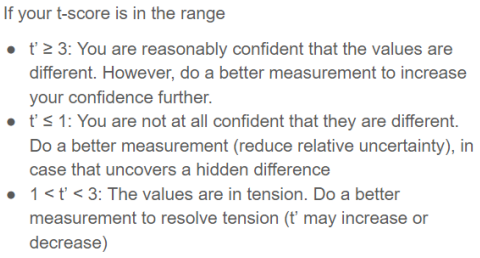

FIG. 1. Instructions given to students on how to interpret a t-score.

II. EXPERIMENTAL CONTEXT AND METHODS

This study is an exploratory analysis of results from a first year physics standalone lab course at a large R1 institution in Canada. Our participants were students enrolled in the fall semester of this course. These students fall primarily into two groups, those taking enriched first year physics, and those enrolled in a first year science cohort program. Data were collected from the third lab in the sequence, the pendulum lab. In order to test the effects of asking students about their experimental expectation, the six sections of the course were split into three "Hypothesizing" sections and three control sections. At the beginning of the lab, each section was asked to complete a short "Start of Lab Survey" for completion credit. The control sections were asked conceptual questions on the lab while the experimental sections where asked to record their expectation. At the end of the lab all sections were asked to complete two reflection questions:

1. Based on your results in this lab, by how much did your periods at 10 degrees and 20 degrees disagree with each other? (Strong disagreement, slight disagreement, values are in tension, we can't see much difference, or values are identical)
2. Is this the result you expected or did it surprise you?

In order to answer our three research questions, we coded the reflection questions for students' experimental conclusion (Agreement, Disagreement, Tension), whether they used their t-score to justify this conclusion, whether they reported being surprised by the result of the experiment, and their hypothesis. The hypothesis was coded because many students provided it when discussing why they were or were not surprised, but it was not explicitly asked for in the control group. Because the information was not volunteered by all students, it was coded with three categories, Agreement, Disagreement and Unsure. Their t-scores were also collected from their lab notebooks. Their stated experimental conclusion (Agreement, Disagreement, Tension) was compared to their final t-score to come up with a binary correct interpretation variable

that became the focus of the analysis. Table I shows the frequency at which students draw different specific experimental conclusions based on their final t-scores.

In addition to data specific to the pendulum lab, we collected responses to the Physics Lab Inventory of Critical Thinking (PLIC) [12] at the beginning and end of the semester. Incentivized with .5% extra credit each and with response rates of approximately 70%, these pre- and post-surveys assessed how well students used data and evidence to make decisions in an experimental physics context. The PLIC post semester survey also included demographic questions. For our analysis, demographic variables class standing and gender were transformed from categorical to binary variables: first year vs non-first year and under-represented vs over-represented, respectively. These two variables, as well as the PLIC pre- and post- scores, were multiply imputed using default predictive means matching in the MICE library in R [13].

The binary outcome variable for this analysis was drawn from Table I and described if their experimental conclusion matched the correct conclusion that should be drawn from their t-score. In that table, entries on the diagonal are considered correct because the stated experimental outcome matches the t-score, entries on the off diagonals are incorrect.

The students in our lab worked in groups of two or three so their t-scores and reflection question responses are correlated. Typically such correlations are accounted for using a mixed-effect logistic regression [14], however, small group sizes can lead to issues with bias and separation in these models [15–17]. Therefore, we chose to use Generalized Estimating Equations (GEE) with a logistic link function. While GEE provides similar information to the more standard logistic regressions, the coefficients are not exactly the same as they are averaged over the population. GEE also allows you to specify a correlation structure which correlates coefficients for the students within each group.

GEE is a population average modelling technique that accounts for correlated responses like those with group clustering or longitudinal studies [17–19]. It works by specifying an estimating equation for the coefficients associated with each predictor in our model. As with a logistic regression, our model is given by a logistic function where the probability of a correct/incorrect $(0 \text{ or } 1)$ interpretation is a function of the predictors $(x_1 \cdots x_m)$ and their associated coefficients $(\beta_1 \cdots b_m)$

$$P(y = 1 \text{ or } y = 0) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}.$$

Unlike logistic regression, which uses a likelihood minimizing technique to estimate the coefficients, GEE uses an algorithm known as a sandwich estimator to determine the coefficients [17, 18]. It does this in a way that accounts for correlations within groups in the data.

Ten variables were used as potential covariates for the modelling. Model selection was performed using the Quasi Information Criteria (QICu) for GEE. These metrics calculate the quality of fit and add a penalty for the number of covariates included. Covariates are added to a base model in a stepwise manner and only included in the final model if they lower the metric by 2.0. [18]. Our ten covariates were: final t-score*, categorical; hypothesizing*, binary; surprise*, binary; experimental expectation, categorical; use of t-score in reflection question answer, binary; gender, binary; class standing, binary; PLIC pre-score*, continuous; PLIC post-score, continuous; first round t-score; categorical.

Covariates indicated with (*) were used as the base model. Final t-score, Hypothesizing and Surprise were chosen for the base model because of their relationship to the three research questions with the inclusion of Hypothesizing answering question 1 and the inclusion of surprise answering question 2. The other covariates were included to answer our third research question by looking at what other factors impact students. PLIC Pre-score was chosen because it helps control for students ability to make decisions based on data in a physics context. Dummy coding was used for the final t-score and first round t-score with $t > 3$ used as the baseline, weighted effect coding was used for the experimental expectation [20]. Because these labs stress the iterative process of experimental physics, all students are required to complete at least two rounds of experimentation.

III. RESULTS

Our Results, shown in Table I show that nearly 80% of students did interpret their t-score correctly, but other trends are visible. The majority of the results showed that the periods of the pendulums are distinguishable, as expected, however, 10% of those students concluded that the periods agreed with each other. More striking, however, is that 45% of students who should have reported inconclusive results $(1 < t' < 3)$ instead characterized their results either as agreement or disagreement. Student reflection questions and histograms of the t-scores provided no evidence that students with t-scores close to the boundary took this into account in their analysis. A McNemar's Chi Squared test was performed to investigate the asymmetry in this table. The test shows that this asymmetry is significant, $\chi^2(3, N = 215) = 25.4, p = 1.2 \times 10^{-5}$. This indicates that the 20% of students who misinterpreted their t-scores are not randomly distributed and there may be difference in student interpretation based on t-score. To investigate this further, we used GEE as described above.

TABLE I. Table showing the stated experimental conclusion vs the t-score from the experiment.

	t'<1 (N=38)	1<t'<3 (N=56)	t'>3 (N=121)
Agreement (N=54)	32	9	13
Tension (N=36)	4	31	1
Disagreement (N=125)	2	16	107

TABLE II. Table showing the results from Generalized Estimating Equations with effect sizes. Effect size was calculated by converting the log odds ratio to a Cohen's d equivalent. ** indicates $p < .05$.

Covariate	log(Odds Ratio)	Effect Size
Final t' : Tension	-2.20 ± .45**	-1.21 ± .25
Final t': Agreement	.06 ± .64	.06 ± .06
Hypothesizing	.75 ± .49	.41 ± .27
Surprise	.60 ± .36	.33 ± .20
Use of t' in reflection question	2.70 ± .58**	1.49 ± .32
PLIC Pre-score	.33 ± .23	.18 ± .13

Results from GEE are shown in Table II where an exchangeable correlation structure was used. The exchangeable correlation structure was chosen based on best practices for clustered data as well as QIC comparisons between independent and unstructured correlation structures [18]. Cohen's d is also shown as a familiar representation of effect size. Of the ten potential covariates, only one was included in the model selection process using the QICu. That only one of the six potential covariates were added in the model selection process is, in itself, interesting. Gender and Class Standing, the two demographic variables, were not included in the model. Previous work [21], suggests that gender can play a major role in how students experience first year labs where they have to work in groups, but that is not the case with the cohort in this lab. The subjects are mostly first year students, so the class standing variable may not provide much meaningful information. In future analysis on the next semester's data, this may change because the course has a higher proportion of upper year students. The QICu similarly did not select the experimental expectation. This variable is not taken directly from the students' hypothesis so it is not directly related to a research question. The design of our study means that only half of the students have an actual hypothesis or experimental conclusion recorded, however, we were still able to determine some of this information from the reflection questions. When asked whether their results surprised them, many students volunteered a hypothesis that was then used to code the experimental expectation variable. We believe this variable serves as a reasonable proxy for an actual hypothesis. When looking at students in the experimental hypothesizing group, 20 of 24 students who were coded with "Agreement" for their expectation recorded the same hypothesis and 23 of 27 who where were coded with "Disagreement" recorded the same hypothesis. Though a proxy for the hypothesis, students' omission of this information from the reflection question explains why it was not selected in the QICu process.

Table II shows the log odds ratios (coefficients from GEE) for the covariates included in the modelling. The log odds ratios allow us to answer our three research questions. The large coefficients for final t-score in Tension and use of t-score in reflection questions shows that these are the most significant factors explaining why students misinterpret their t-scores, answering research question three. Cohen's d shows that both of these factors have an effect size that would be considered very large. While hypothesizing does have a positive effect, Cohen's d shows that it's close to a medium effect but still on the small side and less important than the t-score, answering our first research question. Our second research question is answered by looking at the coefficient for Surprise, which is even smaller than that of Hypothesizing, indicating a small effect. The most significant pedagogical factor is instead the impact of inconclusive results. Including the code that indicates whether students used their t-score to justify their reflection question allows us to distinguish between students who are misinterpreting their t-score and students who don't even know to use a t-score. With the inclusion of this variable we can interpret the final t-score variable knowing that it's providing information for students who actually misinterpreted rather than ignored their t-score.

Our results clearly show that inconclusive data provide a serious stumbling block for students. The odds ratio from our GEE analysis indicates that for an average student who obtained a t-score in tension ($1 < t' < 3$), the odds of interpreting that t-score correctly are 9 times lower than an average student who obtained a t-score showing distinguishability. This is a pedagogically significant difference as confirmed by the very large Cohen's d of 1.21 ± .25. The coding for this categorical variable directly compares students who obtained a t-score in tension to those who obtained $t' > 3$, this was done because having a baseline with the largest population is a good practice and obtaining $t' > 3$; is the outcome we hope our students achieve. When changing the baseline to $t' < 1$ the log odds ratio for tension becomes -2.26 so the conclusions of this analysis remain the same.

We also compared GEE with an independent correlation structure to the logistic regression and confirmed that they gave the same results as expected. The independent correlation GEE model and the logistic regression model had the same covariates and same coefficients after running the model selection process.

A. Discussion

To better understand why these inconclusive results present such a challenge to students, we ran a focus group asking questions related to how they think about their t-scores. This focus group was run in Spring of 2023 after the 4th lab of the semester. There were three participants in the focus group, two of whom intended on majoring in physics. When asked how they felt about getting a t-score in tension, or any inconclusive result, a student responded: "that actually makes us uncomfortable because I may be going on a stretch, but the way we design physics experiments is to answer the question in terms of either, yes, this happens or no, it doesn't happen. We don't have a possibility for 'I don't know', or 'maybe' due to these questions.... Yes or no is the result. It's either yes or no. Like those are results. Doesn't matter if it's yes or no,

but if it doesn't come to that point, if it's inconclusive, then that's simply not a result of it."

This thinking represents a misunderstanding of the statistical nature of the t-score and is instead indicative of the point paradigm approach to measurement and uncertainty [22]. The student believes that there is a a true result of the lab and that result is either a binary yes or no. The values agree or disagree, an effect exists or does not exist. In this framework, a t-score in between one and three is not a result that can be reported. This may explain why students who end the lab with inconclusive results are much more likely to misinterpret their results. It's important to note that not everyone in the focus group shared this understanding as another student suggested that they thought inconclusive results were still valid as they were just on the spectrum between being confident in a result and not confident in a result. While the expert sees the t-score as a continuous spectrum with a grey area in the tension region, a novice may see it as three distinct categories. Agreement and Disagreement exist on the opposites ends of the same spectrum, but tension exists somewhere else entirely as it's not a viable result of the lab.

This thinking is further explained by another student in the group who, when asked if they would interpret a t-score of 2.9 differently from a t-score of 3.1, said that those results were completely different because "The way we've been instructed is just to think of it that way, like it's bigger than three than... we're confident that it's different. That is smaller than three, regardless of how smaller, it's uncertain." This student is suggesting that they were told to follow the cut-offs in t-score as hard and fast rules and not think of them as a spectrum. While this instruction is inconsistent with how t-score is presented in our labs, this student felt that the instruction suggested distinct cut-offs because of the accidental bias in the slide presenting the t-score shown in Fig. 1. This kind of thinking may result in the misinterpretations we have observed.

These results provide a clear path forward to better support students nuanced approach to statistical thinking. We can draw two main lessons from this study, first, students need to end our labs with conclusive results, and second, we need to adjust our instructions for the t-score to support expert-like statistical thinking. The first point is relatively easy to address. In the case of our pendulum labs, students are given the tools to perform high quality conclusive measurements, but they sometimes lack enough time. This study shows that these time crunches are not an insignificant factor in student success. With this knowledge in mind, we are now piloting a two week pendulum lab to ensure that students have enough time to complete a high quality conclusive experiment. Broadly speaking, this also points to the importance of providing students with adequate tools in inquiry labs. We must be mindful of the measurement quality needed to achieve learning goals in these labs and ensure the tools provided can achieve that quality. For example, students need to measure the period in our lab with a relative uncertainty of .1% to observe the small angle approximation. We are careful to design the experiment such that they can achieve this precision with just a stopwatch and repeated measurements.

The t-score was initially introduced to support expert-like statistical thinking and overcome the point paradigm [3]. And while it has been extremely successful in building scientific critical thinking skills, more work is needed. The third student in the focus group shows us exactly the challenge we face in our current t-score framing when they says that they have "just got to trust the statistics and the studies of other people that this [the categorical t-score framing] is like good absolute values that you have to follow." In other words, our current instruction is inadvertently biasing students towards this framework and does not support students when their results are inconclusive. To overcome this, we are reworking the t-score instructions to emphasize the continuous nature of the scale and the validity of inconclusive results. The next step of this study will present the new instructions and assess their impact.

IV. CONCLUSION

We investigated what factors contribute to students misinterpreting their results in a first year lab with unexpected outcomes at a large R1 institution in Canada. Motivated by previous work on the role of model confirmation frames in these types of labs, we designed a study to look at the impact of asking students to hypothesize. Our exploratory analysis shows that hypothesizing had a very small positive impact but was likely not pedagogically significant. Instead, our results clearly show students struggle the most when obtaining inconclusive results. Our GEE analysis shows that for students who obtain an inconclusive result, the odds of characterizing their t-score correctly are 9.06 times lower than those who obtain a conclusive result, with a Cohen's d of -1.21 indicating a large effect. Focus group interviews show that these students may struggle because they do not understand that an inconclusive result is still a result. These focus groups also uncovered further issues with how students interpret their t-scores. Finally we've proposed changes to how t-scores are introduced in our lab and will investigate this intervention in a follow up study.

ACKNOWLEDGMENTS

A.M.K. designed the study, collected the data, analysed the data, and wrote the manuscript with supervision from J.D. J.I. and D.A.B.

[1] N. G. Holmes, Ph.D. thesis, University of British Columbia (2014).

[2] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Proceedings of the National Academy of Sciences **112**, 11199 (2015), ISSN

0027-8424, 1091-6490, URL https://pnas.org/doi/full/10.1073/pnas.1505329112.

[3] N. G. Holmes and D. A. Bonn, The Physics Teacher **53**, 352 (2015), ISSN 0031-921X, URL http://aapt.scitation.org/doi/10.1119/1.4928350.

[4] E. Brewe, American Journal of Physics **76**, 1155 (2008), ISSN 0002-9505, 1943-2909, URL http://aapt.scitation.org/doi/10.1119/1.2983148.

[5] P. A. Bartlett and K. Dunnett, *Secret objectives: promoting inquiry and tackling preconceptions in teaching laboratories* (2019), arXiv:1905.07267 [physics], URL http://arxiv.org/abs/1905.07267.

[6] A. M. Phillips, J. Watkins, and D. Hammer, Physical Review Physics Education Research **13**, 020107 (2017), ISSN 2469-9896, URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.13.020107.

[7] A. M. Phillips, M. Sundstrom, D. G. Wu, and N. Holmes, Physical Review Physics Education Research **17**, 020112 (2021), ISSN 2469-9896, URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.17.020112.

[8] E. M. Smith, M. M. Stein, and N. Holmes, Physical Review Physics Education Research **16**, 010113 (2020), ISSN 2469-9896, URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.16.010113.

[9] E. M. Smith and N. G. Holmes, Nature Physics **17**, 662 (2021), ISSN 1745-2473, 1745-2481, URL http://www.nature.com/articles/s41567-021-01256-6.

[10] M. M. Stein, E. M. Smith, and N. G. Holmes, in *2018 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, Washington, DC, 2019), URL https://www.compadre.org/per/items/detail.cfm?ID=14856.

[11] D. Hu and B. M. Zwickl, in *2017 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, Cincinnati, OH, 2018), pp. 11–14, URL https://www.compadre.org/per/items/detail.cfm?ID=14680.

[12] C. Walsh, K. N. Quinn, C. Wieman, and N. Holmes, Physical Review Physics Education Research **15**, 010135 (2019), ISSN 2469-9896, URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.15.010135.

[13] S. van Buuren and K. Groothuis-Oudshoorn, Journal of Statistical Software **45**, 1 (2011).

[14] B. Van Dusen and J. Nissen, Physical Review Physics Education Research **15**, 020108 (2019), ISSN 2469-9896, URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.15.020108.

[15] M. A. Mansournia, A. Geroldinger, S. Greenland, and G. Heinze, American Journal of Epidemiology **187**, 864 (2018), ISSN 0002-9262, 1476-6256, URL https://academic.oup.com/aje/article/187/4/864/4084405.

[16] E. Laszkiewicz, Metody Ilosciowe w Badaniach Ekonomicznych **XIV**, 19 (2013).

[17] A. E. Hubbard, J. Ahern, N. L. Fleischer, M. V. d. Laan, S. A. Lippman, N. Jewell, T. Bruckner, and W. A. Satariano, Epidemiology **21**, 467 (2010), ISSN 1044-3983, URL https://journals.lww.com/00001648-201007000-00007.

[18] J. W. Hardin and J. M. Hilbe, *Generalized estimating equations* (CRC Press, 2013), 2nd ed.

[19] U. Halekoh, S. HÃžjsgaard, and J. Yan, Journal of Statistical Software **15** (2006), ISSN 1548-7660, URL http://www.jstatsoft.org/v15/i02/.

[20] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the Behavioral Sciences* (Routledge, 2015).

[21] D. Doucette, R. Clark, and C. Singh, European Journal of Physics **41**, 035702 (2020), ISSN 0143-0807, 1361-6404, URL https://iopscience.iop.org/article/10.1088/1361-6404/ab7831.

[22] A. Buffler, S. Allie, and F. Lubben, International Journal of Science Education **23**, 1137 (2001), ISSN 0950-0693, 1464-5289, URL http://www.tandfonline.com/doi/abs/10.1080/09500690110039567.