Finding Significant p in Coffee or Tea: Mildly Distasteful

Sami Sarsa
Aalto University
Espoo, Finland
sami.sarsa@aalto.fi

Arto Hellas
Aalto University
Espoo, Finland
arto.hellas@aalto.fi

Juho Leinonen
Aalto University
Espoo, Finland
juho.2.leinonen@aalto.fi

ABSTRACT

Students' preferences have an impact on their behavior, and behaviors can in turn affect student performance. Earlier work has found that students who tend to work earlier in the course or curse more in their source code tend to perform better. But could other types of preferences also affect student performance? In this work, we examine the relationship between student preferences such as preferring coffee over tea, and students' performance in the course. Our results suggest that certain preferences are related to better overall performance in the course, but only for certain cohorts of students. Indeed, this work provides an example of how easy it is to find statistically significant correlations in educational settings.

ACM Reference Format:
Sami Sarsa, Arto Hellas, and Juho Leinonen. 2022. Finding Significant p in Coffee or Tea: Mildly Distasteful. In *Koli Calling '22: 22nd Koli Calling International Conference on Computing Education Research (Koli 2022), November 17–20, 2022, Koli, Finland*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3564721.3565953

1 INTRODUCTION

Predicting academic performance is a popular research area within computing education research [10]. While much of the research on predicting performance has focused on factors that are directly related to studying, such as time management behavior [6, 12, 14] and log data gathered from learning management systems [3, 11, 26], prior work has also found that student demographics can be used to predict performance [5, 19, 21]. Even blood types [16] and cursing in code [15] have been found to correlate with performance.

One unexplored area is preferences not directly related to studying. For example, the preference of dogs and cats has been found to correlate with personality traits [8]. No prior work, however, has studied whether this preference correlates with programming performance. Similarly, anecdotal stereotypes suggest that programmers drink plenty of coffee [17, 24], which makes us wonder whether a preference of tea over coffee could perhaps be used to identify poorly performing students? Similarly, while it is common knowledge that you should never compare apples and oranges, we bravely go against this conventional wisdom and inquire about preference of apples over oranges. In this work, we answer the question "*How do the preferences of apples and oranges, cats and dogs, and coffee and tea correlate with course performance for different student demographic cohorts?*"

Koli 2022, November17–20, 2022, Koli, Finland

ACM ISBN 978-1-4503-9616-5/22/11.
https://doi.org/10.1145/3564721.3565953

2 RESULTS AND DISCUSSION

The data was collected from an online platform that hosts computer science courses offered by Aalto University. The data consists of the numbers of exercises completed (i.e., course performance), background information (age, gender, courses taken and self-estimated experience) and preferences (*apples* vs *oranges*, *cats* vs *dogs*, and *coffee* vs *tea*). We compute Mann-Whitney U tests for the different preference comparison groups separately for each background cohort. We compute the rank-biserial correlation as the effect size, and we denote it as r. We choose $\alpha = 0.05$ as our p-value threshold. To reduce false discovery rate within the multiple comparisons, we apply the conservative Bonferroni correction to the p-values. A bolded text denotes a p-value below our chosen $\alpha = 0.05$ and an asterisk denotes a p-value small enough to reject the null hypothesis of the test after the Bonferroni correction.

Our results (see Table 1) suggest that the studied preferences are generally not correlated with performance. However, we found statistically significant results for two cohorts even after multiple comparisons correction. The two significant findings were that preferring oranges over apples for 56-65-year-old learners led to better performance, and a preference of tea over coffee led to better performance for those with considerable self-estimated programming experience. Both results are surprising, as anecdotal evidence has suggested opposite results: programming professionals have a reputation of heavy coffee consumption [17, 24], and who has heard of well-performing students gifting *oranges* to their teacher? We note however, that this study was conducted in Finland, a country with high overall coffee consumption rate and we are unaware whether the coffee consumption of programmers deviates from the country average.

Regardless of going against anecdotal evidence, one could draw implications from such results. Students between the ages of 56 and 65 who prefer apples over oranges could need additional support, as could those with considerable programming experience and a preference towards coffee over tea. A concrete practice teachers could employ would be to provide fruit and hot drinks to students on lectures, observe students' preferences, and then provide additional support to those with detrimental preferences who – based on our results – are at risk of poor performance.

In all seriousness, this work highlights the ease of finding spurious statistical significance in an educational research setting. A worrisome phenomenon known as p-hacking, i.e., fishing for significance [1, 22], has become widely known in the scientific community, and also acknowledged in the computing education research community [9]. Combined with the enduring problem of publication bias [23, 25], it is a major issue when interpreting results, especially for meta-analyses [7, 18] as distortions in peer-reviewed evidence accumulate. The term p-hacking may sound as if it refers to malicious intent, but it is prone to emerge also accidentally in

Table 1: Apples vs Oranges, Cats vs Dogs, Coffee vs Tea effect on course completion: Mann-Whitney U test for different cohorts

	Apples vs Oranges			Cats vs Dogs			Coffee vs Tea		
Cohort	U_1	p	r	U_1	p	r	U_1	p	r
Age: 18-25	24192.0	0.2769	-0.0588	23269.0	0.9328	0.0048	22022.5	0.8156	0.0136
Age: 26-35	47463.0	0.4123	0.0382	35515.5	0.6023	-0.0271	25271.5	**0.0167**	-0.1385
Age: 36-45	26056.5	0.2462	0.0628	18710.0	0.8218	-0.0138	15190.5	0.5112	0.0458
Age: 46-55	9084.0	**0.0223**	0.1639	6087.5	0.5542	-0.0464	4467.0	0.2809	-0.0957
Age: 56-65	236.5	**0.0010***	-0.4995	318.0	0.0811	-0.2639	240.0	0.4299	-0.1489
Age: 65-	97.0	0.9102	-0.0300	92.5	0.1363	0.4015	29.0	0.0856	-0.4957
Courses taken: 0-1	58356.0	0.2539	0.0506	48908.5	0.8243	0.0105	39118.5	0.9394	0.0040
Courses taken: 2-4	21900.5	0.4279	-0.0441	17474.0	0.2929	-0.0647	17334.5	0.3134	-0.0620
Courses taken: 5-10	11830.0	0.4025	0.0546	7959.0	0.1794	-0.0967	6986.0	0.328	-0.0759
Gender: female	47656.5	0.0811	-0.0791	43627.5	0.1624	-0.0666	36650.5	0.1853	-0.0688
Gender: male	137588.5	**0.0104**	0.0921	102139.0	0.9471	-0.0026	77614.0	0.084	-0.0751
Gender: other	58.0	0.9491	-0.0252	89.0	0.2765	0.2714	63.5	1.0	-0.0078
Self-estimated experience: 1-2	44449.5	0.6481	0.0215	39290.5	0.506	0.0334	31752.5	0.4878	0.0388
Self-estimated experience: 3-4	23536.5	0.9458	-0.0037	20381.0	0.8244	-0.0130	18790.5	0.4385	0.0486
Self-estimated experience: 5-9	17315.5	0.8598	0.0105	10525.0	**0.0325**	-0.1464	9525.0	**0.00038***	-0.2410

exploratory research. This highlights the need to report every step of the research [4, 27], or even preregister the analysis [2], and for education on the use and interpretation of p-values [13]. Issues in reporting inferential statistics, such as not reporting exact p-values, not applying corrections for multiple tests, and not reporting effect sizes are unfortunately common in computing education literature [20].

REFERENCES

[1] Anne-Laure Boulesteix. 2010. Over-optimism in bioinformatics research. *Bioinformatics* 26, 3 (2010), 437–439.

[2] Neil CC Brown, Eva Marinus, and Aleata Hubbard Cheuoua. 2022. Launching Registered Report Replications in Computer Science Education Research. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*. 309–322.

[3] Adam S Carter, Christopher D Hundhausen, and Olusola Adesope. 2015. The normalized programming state model: Predicting student performance in computing courses based on programming behavior. In *Proc. of the Eleventh Annual Int. Conf. on Int. Computing Education Research*. 141–150.

[4] Luca Chiodini and Matthias Hauswirth. 2021. Wrong Answers for Wrong Reasons: The Risks of Ad Hoc Instruments. In *21st Koli Calling Int. Conf. on Computing Education Research*. 1–11.

[5] Ying Cui, Fu Chen, Ali Shiri, and Yaqin Fan. 2019. Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences* (2019).

[6] Stephen H Edwards, Jason Snyder, Manuel A Pérez-Quiñones, Anthony Allevato, Dongkwan Kim, and Betsy Tretola. 2009. Comparing effective and ineffective behaviors of student programmers. In *Proc. of the fifth int. workshop on Computing education research workshop*. 3–14.

[7] Malte Friese and Julius Frankenbach. 2020. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods* 25, 4 (2020), 456.

[8] Samuel D Gosling, Carson J Sandy, and Jeff Potter. 2010. Personalities of self-identified "dog people" and "cat people". *Anthrozoös* 23, 3 (2010), 213–222.

[9] Patricia Haden. 2019. *Inferential statistics*. Cambridge University Press, Chapter 6, 133–172.

[10] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 2018. Predicting academic performance: a systematic literature review. In *Proc. companion of the 23rd annual ACM conf. on innovation and technology in computer science education*. 175–199.

[11] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. *Proc. of the 2015 ITiCSE on Working Group Reports* (2015), 41–63.

[12] Ayaan M Kazerouni, Stephen H Edwards, and Clifford A Shaffer. 2017. Quantifying incremental development practices and their relationship to procrastination. In *Proc. of the 2017 ACM Conf. on Int. Computing Education Research*. 191–199.

[13] Daniël Lakens. 2021. The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science* 16, 3 (2021), 639–648. https://doi.org/10.1177/1745691620958012 arXiv:https://doi.org/10.1177/1745691620958012 PMID: 33560174.

[14] Juho Leinonen, Francisco Enrique Vicente Castro, and Arto Hellas. 2021. Does the early bird catch the worm? Earliness of students' work and its relationship with course outcomes. In *Proc. of the 26th ACM Conf. on Innovation and Technology in Computer Science Education V. 1*. 373–379.

[15] Juho Leinonen and Arto Hellas. 2017. Thought crimes and profanities whilst programming. In *Proc. of the 17th Koli Calling Int. Conf. on Computing Education Research*. 148–152.

[16] Kartika Maharani, Teguh Bharata Adji, Noor Akhmad Setiawan, and Indriana Hidayah. 2015. Comparison analysis of data mining methodology and student performance improvement influence factors in small data set. In *2015 Int. Conf. on Science in Information Technology (ICSITech)*. IEEE, 169–174.

[17] Danila Petrova. 2019. *The Relationship Between Coffee and Developers*. https://simpleprogrammer.com/coffee-and-developers/

[18] Hannah R Rothstein, Alexander J Sutton, and Michael Borenstein. 2005. Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments* (2005), 1–7.

[19] Isabel Ruthotto, Quintin Kreth, Jillian Stevens, Clare Trively, and Julia Melkers. 2020. Lurking and participation in the virtual classroom: The effects of gender, race, and age among graduate students in computer science. *Computers & Education* 151 (2020), 103854.

[20] Kate Sanders, Judy Sheard, Brett A Becker, Anna Eckerdal, Sally Hamouda, and Simon. 2019. Inferential statistics in computing education research: A methodological review. In *Proc. of the 2019 ACM conf. on int. computing education research*. 177–185.

[21] Sami Sarsa, Arto Hellas, and Juho Leinonen. 2022. Who Continues in a Series of Lifelong Learning Courses?. In *Proc. of the 27th ACM Conf. on on Innovation and Technology in Computer Science Education Vol. 1*. 47–53.

[22] Joseph P Simmons and RA LeBoeuf. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[23] Theodore D Sterling. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. of the American statistical association* 54, 285 (1959), 30–34.

[24] Lisa Sudlow. 2015. *Dev Tidbits #001: How much caffeine do developers consume?* https://developermedia.com/how-much-caffeine-do-developers-consume/

[25] Alex J Sutton. 2009. Publication bias. *The handbook of research synthesis and meta-analysis* 2 (2009), 435–452.

[26] Christopher Watson, Frederick WB Li, and Jamie L Godwin. 2013. Predicting performance in an introductory programming course by logging and analyzing student programming behavior. In *2013 IEEE 13th int. conf. on advanced learning technologies*. IEEE, 319–323.

[27] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* (2016), 1832.