DOI:10.1145/3372161

Evidence That Computer Science Grades Are Not Bimodal

By Elizabeth Patitsas, Jesse Berlin, Michelle Craig, and Steve Easterbrook

Abstract

Although it has never been rigorously demonstrated, there is a common belief that grades in computer science courses are bimodal. We statistically analyzed 778 distributions of final course grades from a large research university and found that only 5.8% of the distributions passed tests of multimodality. We then devised a psychology experiment to understand why CS educators believe their grades to be bimodal. We showed 53 CS professors a series of histograms displaying ambiguous distributions that we asked them to categorize. A random half of participants were primed to think about the fact that CS grades are commonly thought to be bimodal; these participants were more likely to label ambiguous distributions as "bimodal." Participants were also more likely to label distributions as bimodal if they believed that some students are innately predisposed to do better at CS. These results suggest that bimodal grades are instructional folklore in CS, caused by confirmation bias and instructor beliefs about their students.

1. INTRODUCTION

It is a prevailing belief in the computer science education community that CS grades are bimodal, and much time has been spent speculating and exploring why that could be (For a review, see Ahadi and Lister[1].) These discussions generally do not include statistical testing of whether the CS grades are bimodal in the first place. From what we have seen, people take a quick visual look at their grade distribution, and if they see two peaks, they conclude that it is bimodal. But eyeballing a distribution is unreliable; for example, if you expect the data to have a certain distribution, you are more likely to see it.

Anecdotally, we have seen new instructors and TAs (and students) who have shown histograms of grades and told the grades were "bimodal." The bimodality perception hence becomes an organizational belief, and those who enter the community of practice of CS educators are taught this belief.

1.1. Explanations for bimodal grades

A number of explanations have been presented for why CS grades are bimodal, all of which begin with the assumption that this is the case.

Prior experience. A bimodal distribution generally indicates that two distinct populations have been sampled together.[5] One explanation for bimodal grades is that CS1 classes have two populations of students: those with experience, and those without.[1]

In many places, high school CS is not common or standardized, and so students enter university CS with differing amounts of prior experience. However, this explanation fits students into only two bins. Prior experience is not as simple as "have it" vs. not-there is a wide range of how much prior programming experience students may have, and practice with nonprogramming languages such as HTML/CSS could also be beneficial.[18]

Learning edge momentum, stumbling points, and threshold concepts. One family of explanations posits that some CS concepts are more difficult for students to learn, and if they miss these concepts, they fall behind, whereas their peers advance ahead of them. As it is typically taught, CS1 builds on itself heavily. So once a student falls behind, they continue to fall further and further behind.[1] This may be exacerbated by the fact that some concepts may be key to understanding ("threshold concepts"). One might think of this explanation as a variant of the prior-experience explanation, where the students who have better study skills succeed, and those with weaker skills fall behind.

The Geek Gene Hypothesis. Some would instead argue that the two populations in CS1 classes are those who have some "natural talent," giftedness, or predisposition to succeed at computing. Guzdial has referred to this belief as the "Geek Gene Hypothesis".[6] This belief appears to be quite prevalent. In a survey of CS faculty, Lewis found that 77% of them strongly disagree with the statement "Nearly everyone is capable of succeeding in the computer science curriculum if they work at it."[14] However, there seems to be little evidence that there is indeed a "Geek Gene", and plenty of evidence that effective pedagogy allows for all students to succeed.[8]

Coarse assessment. Another line of explanation implicates instructors' assessment tools as the source of bimodally distributed grades.[28, 20] A common trend on CS exams is to ask a series of long-answer coding questions. Zingaro et al. found that these questions offer only coarse assessment information to instructors: students either put all the pieces together, or fail to. Instructors do not adequately identify when a student has partial understanding nor quantify how much understanding a student has of a concept.

As an alternative, Zingaro et al. experimentally compared using short-answer questions that build upon each other to have one isomorphic long-answer question. When the different conceptual parts of the question were broken up, the resulting grades were normally distributed. The all-or-nothing nature of long-answer questions could lead to grades more likely to be (or appear) bimodal.[28]

Or perhaps CS grades are not bimodal? A competing view

The original version of this paper was published in the *Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER).*

of CS grades put forth by Lister is that the grades are not, in fact, bimodal.[15] Lister observed that CS grade distributions are generally noisy, and in line with what statisticians would accept as normally distributed. Lister argued that the perception of bimodal grades results from instructors' beliefs in the Geek Gene Hypothesis, and hence, instructors see bimodality where there is none.[15] Lister's argument was theoretical, and based on statistical theory; in this paper, we test his argument by statistically analyzing actual grade distributions.

2. WHAT IS A BIMODAL DISTRIBUTION?

To properly tackle the question of "are CS grades bimodal?", we should first clearly establish what bimodality means. For a comprehensive discussion of this, we suggest the reader consult[25]; we summarize some major points of that article in this section.

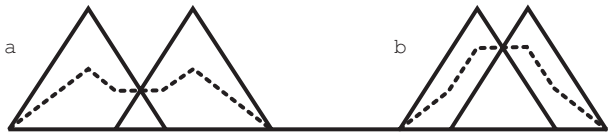
Most standard continuous probability distributions have a mean, a median, a mode, and some measure of the distribution's width (variance). Standard distributions include the normal (Gaussian), Pareto, Poisson, Cauchy, Student's t, and logistic distributions. When we plot them (or likely, a sample thereof) with a histogram, we see their probability density. All of these distributions have a single mode, and have a probability density that can be modeled with a function that has a single term. For example, the normal distribution's PDF is

$$f(x) = ae^{\frac{(x-b)^2}{2c^2}}$$

In this function, a represents the height of the curve's peak, b is the position of the center of the peak, and c represents the width of the curve.[27]

In contrast, a bimodal distribution has two *distinct* modes. A 'multimodal' distribution is any distribution with multiple distinct modes (two or more). For example, consider these examples from.[25] Both are created by the equal mixture of two triangular distributions (solid lines). The sums are shown with dashed lines:

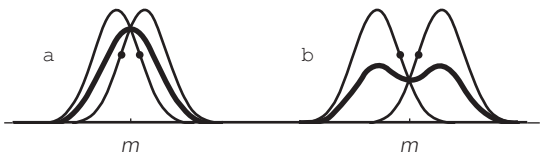
As we can see, when the two subdistributions are far away



(example **a**), we get a distribution with two peaks. But when the two subdistributions are close together (example **b**), they add together to form a plateau, with a single peak. Example a is considered bimodal; example **b** is not.

The same is true for normal distributions (also from Schilling et al.[25]):

For a distribution to be bimodal, the subdistributions



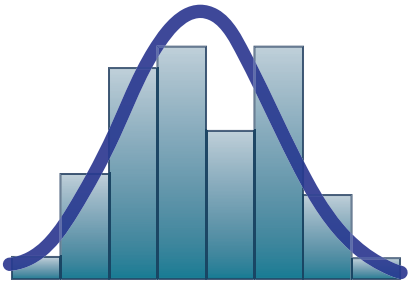
cannot overlap too much. As shown in Schilling et al.[25], for the two distributions to be sufficiently far apart, the distance between the means of the two distributions needs to exceed 2σ. This, however, assumes that the two distributions have the same variance. More formally, if the two subdistributions do not have the same variance, then for their sum to be bimodal, the following must hold[26]:

$$2^{\frac{1}{2}} \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2 + \sigma_2^2)}} > 2$$

2.1. Histograms can deceive

Consider this histogram of sepal widths for the Iris species *versicolor,* taken from the Wikipedia page on "normal distribution"[27]:

The data has two peaks, but the data is considered to be



sampled from a normal distribution. If we were to try and model this data as the mixture of two normal distributions, the two subdistributions would be too close together to produce two distinct peaks. The simplest way to model this data is as a normal distribution, especially as this is consistent with biological theory.

Remember that what we see in a histogram is a result of how we select the bins. It is possible to bin this data in a way that does not have two 'peaks' (for example, by using larger bin intervals, or shifting the bin boundaries). With grade distributions, ceiling effects are common: if you take normally distributed data, and then lower the values above 100% down to 100%, you may wind up seeing a second "peak" in your histogram's top bin. For an illustration, see distribution 6 in Figure 1.

3. STUDY 1: GRADES ANALYSIS

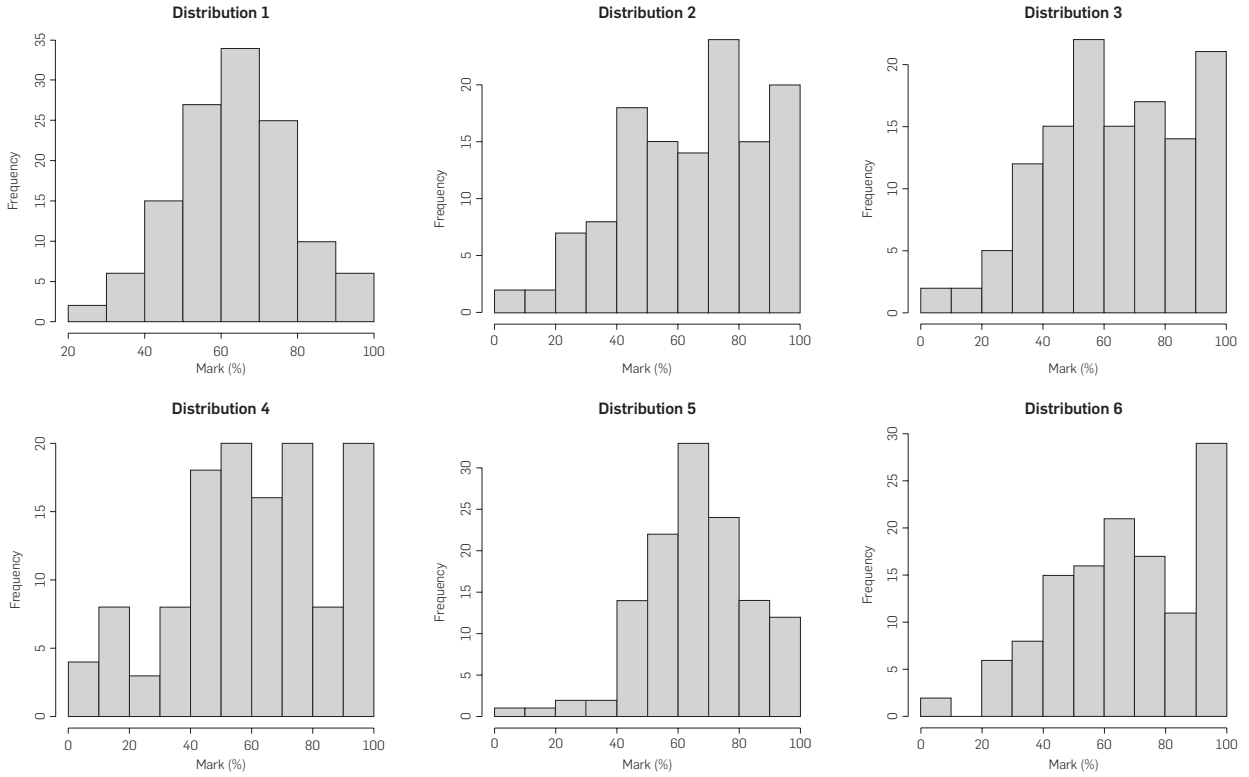
Are CS grades bimodal, or unimodal? To test this, we acquired the final grade distributions for every undergraduate CS class at the University of British Columbia (UBC), from 1996 to 2013. This represents 778 different lecture sections, containing 30,214 final grades (average class size: 75). We analyzed this data to see what distribution(s) it may have most likely come from. Frequentist null-hypothesis testing is the standard in computer science education research; for readers who are unfamiliar interpreting p values from null-hypothesis tests, we recommend consulting Goodman.[4]

3.1. Testing for multimodality

We began by computing the kurtosis for each class. *Kurtosis* is a measure of how 'tailed' the data is: high kurtosis means a distribution has a sharp peak and short tails, whereas low kurtosis implies low peak(s) and long tails.

If you look back at the illustration of adding two normal

Figure 1. The six histograms shown to participants, all of which were generated using GNU R's `rnorm` function. A ceiling of 100% was used, which is most evident in Distribution 6. Each generated distribution had 100 points, and was generated with an average of 60 and standard deviation of 5 and displayed as a histogram with bins of size 10.



distributions together, for the bimodal example, the distribution winds up being rather spread out horizontally. That distribution has low kurtosis. Indeed, for a distribution to be spread out far enough horizontally to allow for multimodality, it necessarily will have low kurtosis.

The normal distribution has a kurtosis of three. A distribution with a kurtosis greater than three cannot be bimodal.[26] We found that 323 of the 778 classes had a kurtosis less than 3. This means that 455 (58%) of the classes were not bimodal, and that *at most* 323 (42%) classes could be bimodal.

Hartigan's Dip Test. Hartigan's Dip Test is a test for testing whether data is multimodal (bimodal, trimodal, etc.). It looks at whether there is a "dip" in between the possible means and how deep the "dip" is (essentially: whether there is a concave-up section in the distribution). We applied Hartigan's Dip Test to the 323 classes that had a kurtosis less than 3. We chose to apply the test only to these 323 classes rather than the full 778 set in order to reduce the likelihood of false positives. For Hartigan's Dip Test, the null hypothesis is that the population is unimodal. As such, **our null hypothesis for each of the 323 tests was that a given class is unimodal**.

Results of the Dip Test. Of the 323 classes that had a kurtosis below three, 45 classes yielded a p value from Hartigan's Dip Test that was below our α value of 0.05. This is 13.9% of all the classes on which we ran Hartigan's Dip Test, or 5.8% of all the classes in our data set.

We chose the standard α value of 0.05. This means that *if* the null hypothesis is true (unimodal), the chance of a false null hypothesis rejection is 5%.[4] If the null hypothesis is false (multimodal), the chance of a false null hypothesis rejection is 0%, because we cannot falsely reject a false null hypothesis.[4] Until it is known whether the null hypothesis is true or not, the chance of a false positive lies between 0% and 5%.[a4]

It could be the case that all 45 classes where we can reject the null hypothesis are indeed multimodal. But also given the noisiness of grading, ceiling effects, and small sample sizes, it could still be the case that all of these 45 classes are indeed unimodal.[4]

Although we cannot give conclusive determinations on a given null hypothesis test, the results here do provide information. Even in the unlikely case that all 45 of these classes are indeed multimodal, we see that multimodal distributions are far from being typical.[b]

[a] To give the reader a sense of the reliability of Hartigan's Dip Test, we generated 100,000 distributions with R's `rnorm` with n=100, μ=60, and σ=5. A total of 133 distributions (1.3%) were tested as multimodal per Dip Test. This gives us some indication that false positives will occur with the test, but likely less than 5.8%.

[b] Many people have asked whether first-year classes are more likely to be multimodal than upper-level classes. Given how few classes passed the test of multimodality, we do not have sufficient data one way or the other to properly evaluate this. More data and replication at other universities would be needed to properly test if multimodal distributions occur more often in lower-level courses.

3.2. Testing for normality

A variety of null hypothesis tests, such as Anderson-Darling, Shapiro-Wilk, and Pearson's chi-squared test determine whether a dataset is normal. We chose Shapiro-Wilk, because it has been found to have the highest statistical power.[21]

Shapiro-Wilk test. For the Shapiro-Wilk test, **the null hypothesis is that the population is normally distributed.** So, if $p < \alpha$, we can reject the null hypothesis and have evidence that the population is not normally distributed. We could reject the null hypothesis for 106 classes. This indicates that 13.6% of the classes in the data set are *not* normally distributed. As with the results of Hartigan's Dip Test, this does not mean that the null hypothesis is necessarily false in these cases. There are many reasons a distribution could not be normal: for example, it could be too skewed, it could be the wrong shape (e.g., triangular and uniform), or it could be multimodal.

It is worth noting that of the 45 classes where we rejected the null hypothesis that they were unimodal, for 44 of these classes we also rejected the null hypothesis that they were not-normal. As such, 44 of the 106 (41.5%) of the classes that were tested as being not-normal were also tested as being multimodal.

For the 86.4% of classes where we failed to reject the null hypothesis, we cannot guarantee that they are actually normal (type II error). To give an estimate of how many are actually normal, we bootstrapped a likely beta value. This yielded an estimated false negative rate of 1.48%.

From our data, we estimate that 85.1% of the final grades in UBC's CS classes are normally distributed. This indicates that grades from a computer science class are typically normal—not bimodal.

Skewness. Although most of the distributions appear to be normally distributed, it is worth noting that the average skewness of all the distributions is −0.33, whereas a normal distribution should have a skewness of zero. If we only consider the distributions whose test results indicated normality, the average skewness is −0.13. This provides some sanity checking on our normality testing: the "normal" distributions are not particularly skewed. For the classes where we rejected the null hypothesis of normality (i.e., probably not normal), the average skewness was higher. Likely, this is why many of these classes were indicated by Shapiro-Wilk as not normal. Higher skewness could also be a result of the ceiling effect in grade distributions.

3.3. Discussion

We only examined final grades: our analysis did not include term grades. And as grades only came from one institution, one may wonder about generalizability. We tried to acquire grade distributions from other institutions, but generally found it difficult to gather the same scale of data. What stood out for us is that our colleagues (both at UBC and elsewhere) would routinely assert that their CS grades are bimodal, and our analysis gives evidence to the contrary. Although we cannot assert from this analysis that every university has the same distributions as UBC, the large scale of data both in numbers and time-span is compelling. Our interpretation is also not alone: our results support Lister's argument that CS grades are generally not bimodal.

We invite readers to replicate our findings at other institutions.[c] The code to replicate the analysis is available online at https://github.com/patitsas/bimodality.

4. STUDY 2: HUMAN INTERPRETATION OF DISTRIBUTIONS

So if CS grades are rarely bimodal, why does the belief in bimodality persist? An insight came one day when generating some random normal distributions in R: with only 100 data points, the resulting histogram often had more than one peak and could be easily erroneously perceived as "bimodal". A typical "large class" does not have a large enough sample size to consistently provide a smooth curve. Indeed, many of the distributions produced by R's `rnorm` looked very much like the grade distributions we had seen in our own classes and called "bimodal."

Interested in whether instructor perceptions affect the interpretation of noisy distributions, we designed an experiment wherein participants are presented with histograms of distributions produced by R's `rnorm` function, and asked to categorize the distribution (normal, bimodal, uniform, etc.). We initially had two research questions:

1. Do CS instructors who believe in innate ability categorize more noisy distributions as bimodal?
2. If we prime participants that CS distributions are commonly thought to be bimodal, are they then more likely to see bimodal distributions in the noise?

Once we analyzed our data for those two research questions, a third research question arose:

3. If instructors label noisy distributions as bimodal, are they more likely to agree with the idea of innate CS ability? (i.e., is there a possible feedback loop between looking at distributions and instructors' beliefs?)

4.1. Experimental design

A difficulty in studies looking at priming effects is that you cannot state the purpose of the study in the consent form. If you do, then you are priming participants, even the participants you want in your control group. To disguise our study, we presented it as one asking people how often they saw various distribution shapes in their own classes.

We presented each participant with the six histograms as shown in Figure 1, all of which we generated using R's `rnorm` function. We generated a few dozen histograms and selected the six histograms from that pool: one to be clearly normal (distribution 1), one that was mildly skewed as though students who were failing were pushed up to 50% (distribution 5), one where the ceiling effect was visible (distribution 6), and three noisy distributions which had multiple peaks (distributions 2–4).

We asked each participant whether they saw this shape of

[c] Since the original ICER publication, our findings have been replicated at a university in the United States.[2]

distribution in their own classes ("very often" to "never" on a Likert scale), and how they would categorize the distribution (normal, bimodal, multimodal, uniform, and others). We randomly assigned participants to one of two treatments:

Treatment 0: participants were asked whether they agreed that CS ability is innate, then asked to categorize the distributions, and were not being primed to think about bimodality.

Treatment 1: participants were primed to think about the common-held belief about CS grade distributions, before they saw the distributions; after that, we asked whether they agreed that CS ability is innate.

The survey's five pages are described in Table 1. For each question, we created a shorthand label, shown in sans-serif, for use in our analysis.

Because so many of the potential participants were our colleagues, we deliberately did not collect names and identify information about participants. We did not want to know who was or was not a participant, nor how they responded to the survey.

As a courtesy, we offered participants the option of having their email recorded on a separate platform if they wanted us to follow up with them about the results of the study.

We did not look at this email list until after our analysis was complete.

4.2. Participants

We recruited 60 CS instructors, mostly from the SIGCSE members' list. Some participants were recruited from other online CS education communities, and some were recruited at ICER 2015. Fifty-three participants completed every question on the survey; twenty-eight were in Treatment 0 (the nonprimed group), and twenty-five were in Treatment 1 (the primed group). The participants who had provided their emails for follow-up purposes were debriefed. As fewer than half of the participants had provided their email, we posted open debriefing statements to the online communities where we had recruited participants.

Table 1. The pages of the survey.

1. Questions about how large their typical class was ("class-size") and how long they had been teaching ("years-experience").
2. A priming question: 'It is a commonly held belief that CS grade distributions are bimodal. Do you find this to be the case in your teaching?' ("have-bimodal")
3. Questions on how often they look at their grade distributions:
   - 'When teaching, how often do you look at histograms of your students' grades? (This applies both to term work and final grades.)' ("look-histo")
   - 'How often do you look at how many students fall into each letter category (A, B, etc.)? (This applies both to term work and final grades.)' ("look-letter")
4. Six histograms, all generated with GNU R's `rnorm`, shown in Figure 1. For each histogram, we asked two questions:
   - 'How often do you see the shape of [this distribution] in your classes?'
   - 'What sort of distribution would you describe [this distribution] as?'
5. Likert-style questions on innate ability (5 points, Strongly Agree to Strongly Disagree):
   - Nearly everyone is capable of succeeding in computer science if they work at it. ("all-succeed")
   - Some students are innately predisposed to do better at CS than others. ("innately-predisposed")

Pages 2 and 5 were swapped for a random half of the participants. We chose the all-succeed question because it had been used previously in the literature.

4.3. Results

For each participant, we computed a value we call "seeing-bimodality," which is how many of the six distributions the participant had categorized as bimodal or multimodal. In our data, seeing-bimodality ranged from 0 to 5.

Regression on seeing-bimodality. We wanted to see if seeing-bimodality could be predicted by participants' responses to our questions. The regression we performed was to model seeing-bimodality as a function of innately-predisposed, all-succeed, look-histo, and look-letter (shorthand names from Table 1).

When visualizing the results, we noticed that the relationship between seeing-bimodality and the Likert questions varied between the two treatments. As a nonparametric equivalent of ANCOVA, we performed an ordinal logistic regression on the two treatments separately using the `polr` function from R's `MASS` library, and then used the `Anova` function from the `car` package to compare the two. This allowed us not only to test whether there were relationships between seeing-bimodality and the Likert questions, but to see if these relationships were different for the two treatments. This approach required computing 28 p values. To reduce the chance of false positives from using multiple statistical tests, we applied a Šidák correction, which reduced our α level to 0.002 for this section of our analysis.

In both our regressions on Treatment 0 and on Treatment 1, we found a significant relationship between seeing-bimodality and participants' responses to the questions related to innate ability (all-succeed and innately predisposed).[d]

We then looked to see if this relationship was stronger in one treatment than the other. In both questions about innate ability, the effect was significantly stronger in the treatment where subjects were primed to think about CS grades being bimodal, as shown in Table 2.

Both regressions also revealed a statistically significant relationship between seeing-bimodality and how often participants reported looking at histograms of their grades (look-histo). This relationship was not statistically significantly different between the two treatment groups.

Perhaps unsurprisingly, there was a strong negative

[d] Regression tables are provided in the original ICER publication, and are omitted due to page limitations.

Table 2. Results of the `Anova` of the regressions on the two treatments; that is, does the relationship between a given factor and seeing-bimodality differ between the two treatments?

	LR Chisq	Df	Signif?
innately-predisposed	11.0	2	yes
all-succeed	14.8	3	yes
look-histo	4.1	4	no
look-letter	6.1	4	no

correlation between all-succeed and innately predisposed. Those who felt there was an innate predisposition to do well in CS also felt that not everyone could succeed in the field.

Regression on all-succeed. After finding a one-way relationship between grade perceptions and the innateness belief, we wanted to see if there was any evidence of a feedback loop between the two. Because all-succeed and innately predisposed correlated so highly, we found they were interchangeable as measures of belief in innate ability. As logistic regression involves only one dependent variable, we had to pick one of the two to use. We chose to do this analysis with all-succeed because the question item had been used in another study.[14]

Recall that our study was set up so that a random half of the participants categorized distributions and then were asked about innate ability (Treatment 1), whereas the other half were asked about innate ability and then categorized the distributions (Treatment 0). If there is a feedback loop here, we would expect that seeing-bimodality would predict all-succeed in Treatment 1, but not in Treatment 0.

Guidelines for statistical power in logistic regression suggest that an α level of 0.05 requires 10–20 data points per independent variable in your model.[16] Because this part of the analysis requires the statistical power to reject a null hypothesis, we modeled all-succeed as only a function of seeing-bimodality, and set $\alpha = 0.05$.

For Treatment 1, we found that seeing-bimodality was a statistically significant predictor of all-succeed. In Treatment 0, it was not. This indicates that there is a feedback loop between categorizing distributions as bimodal and agreement with the idea of innate ability. We hence have observed evidence for the relationships illustrated in **Figures 2** and **3**.

4.4. Discussion

With regard to the feedback loop between seeing-bimodality and all-succeed, we have some weak evidence that categorizing distributions as bimodal increases belief in the Geek Gene Hypothesis. We consider our evidence weak because our study was underpowered, and caution should be taken in interpreting the lack of significance in the second treatment.

We were initially surprised that regularly looking at histograms of grades was associated with a higher score for seeing-bimodality. This led us to add our third research question, based on the idea that it could be that the more often you look at your grades, the more it solidifies your conception of what your grades are like. This supports our observation that categorizing distributions as bimodal increases belief in innate ability. System justification theory explains how once you are forced to take a position on a subject, you are more likely to believe and defend it.[11] Our approach to priming—stating that it is a commonly held belief that CS grade distributions are bimodal—may have strengthened our participants' beliefs about the bimodality of CS grades. Because the survey presents us, the researchers, as authority figures, and we imply that grades may be bimodal, some participants could assume it to be true because of our endorsement.

When we piloted our survey, some participants opined that they believed that some students were predisposed because of prior experience, rather than inherent brilliance.

We did not have a representative sample of CS educators. The educators who participate in CS education communities are generally much more invested in their teaching than their peers who do not. Furthermore, some of our participants may be familiar with Ahadi and Lister[1], which could have influenced their responses. But we would expect the SIGCSE community to be less inclined to believe in innate ability than their non-SIGCSE peers. We still had enough participants who agreed with the hypothesis for us to conduct our analysis. Future work is needed to replicate our findings with a more representative sample of CS educators.

Supporting literature. Our findings agree with the psychology literature: people's biases affect their decision-making more when they are judging more ambiguous information.[10] For example, Heilman et al. found that resumes of extremely qualified candidates were likely to be judged worthy of a salary increase regardless of the gender listed on the resume—but for resumes of ambiguously qualified candidates, resumes with male names were more likely to be viewed positively than those with female names.[10] Eyesnck et al. studied the interpretation of written sentences as either threatening or nonthreatening by people who have anxiety and by a control group.[3] They found that unambiguously threatening or nonthreatening sentences were interpreted similarly between groups, but participants with anxiety were more likely than controls to label ambiguous sentences as threatening. Visual information is also subject to this phenomenon: Payne et al. showed participants a series of photos of people holding either guns or ambiguous objects, and participants were more likely to identify the ambiguous object as a gun if it was held by a black person.[19]

Figure 2. Individual-level feedback loop leading individuals to categorize ambiguous distributions as bimodal.

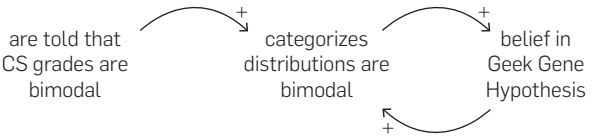
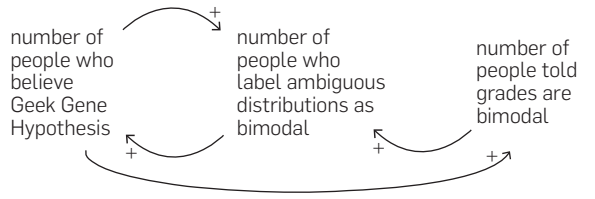


Figure 3. Social-level feedback loops leading individuals to categorize ambiguous distributions as bimodal.



Furthermore, belief can affect judgment regardless of ambiguity. For example, Kahan et al. found that participants were more likely to get a math problem incorrect if the correct result would disagree with their political beliefs.[12] It is hence plausible that a computer scientist who believes in the Geek Gene Hypothesis could look at an unambiguously unimodal distribution and still view it as bimodal.

5. THE GEEK GENE HYPOTHESIS AS A SOCIAL DEFENSE

Once again, our findings support Lister's hypothesis that CS grades are generally not bimodal and this perception stems from instructors expecting to find bimodal grades due to a belief in the Geek Gene Hypothesis. We now go a step further and argue that the perception of bimodality is a *social defense* in the CS education community.

In sociology and social psychology, a "social defense is a set of organizational arrangements, including structures, work routines, and narratives, that functions to protect members from having to confront disturbing emotions stemming from internal psychological conflicts produced by the nature of the work".[17]

5.1. Social defenses in teaching

Guzdial reports that teachers generally have a high level of self-efficacy (great confidence in their teaching ability) at the start of their career. This then plummets as they face the realities of classroom teaching but slowly returns with time.[9] Teacher self-efficacy is not necessarily tied to teaching ability: university educators often get little meaningful feedback on how their students are learning, given their large class sizes and lecture-based pedagogies.[9]

Guzdial notes that if an individual university-level CS educator has high self-efficacy, and sees evidence of students not learning, then it is rational for them to believe that the problem lies with the students and that the problem is innate to them—that is, beyond the ability of the teacher to influence.[9] Compounding this, Sahami and Piech have observed that CS educators are more aware of their top and bottom students than they are of their average students, giving educators a biased perception of their students' abilities.[24] Guzdial argues that CS educators have poor results, because we so frequently use ineffective teaching methods.[7] Zingaro et al. suggest that not only do CS educators frequently use ineffective pedagogies, they also frequently use ineffective assessment tools.[28, 20]

We theorize that the Geek Gene Hypothesis is a social defense: it is easier for computer science educators to blame innate qualities of their students for a lack of learning than it is for the educators to come to terms with the ineffectiveness of their teaching.

A social defense is a phenomenon on a *social* scale, in contrast to Guzdial's observation about individual teachers. When numerous educators bond over how their students just "do not have it," it allows for the Geek Gene hypothesis to go from one individual's suspicion to a social narrative. And as bimodal grade distributions sometimes do occur, those cases are used to argue that this is a common and inherent phenomenon in CS classes. The perception of bimodal grades provides evidence to the Geek Gene narrative that some students "have it" and some do not. And when new educators who have been primed to see bimodality then begin teaching and do not see all their students learning, these new educators can then see this as evidence of the Geek Gene. The reproduction of the Geek Gene Hypothesis is hence social in nature.

5.2. The "Geek Gene" is an equity issue

Debunking the "Geek Gene" is also important for equity reasons. Recent studies have found that academic disciplines in which "brilliance" is seen as necessary for success have less gender diversity.[13] Looking at the history of science, women and people of color were long denied entry and acknowledgment in science because they were seen as lacking the "brilliance" needed to do science.[23] If computing ability is viewed as being the result of a "Geek Gene," then educators may use this as a reason not to teach students who they perceive as lacking this "gene." Similarly, they could lower expectations of these groups and encourage them less—which is troubling given evidence that teacher expectations have an effect on student performance.[22]

6. CONCLUSION

Our analysis of one institution's CS grades indicates that although bimodal grade distributions can be found, they are far from typical. Much more commonly, grade distributions are normal (85.1% of cases) or highly skewed unimodal distributions. Our psychology experiment found that participants who were more likely to label ambiguous distributions as bimodal were also more likely to report a belief in an innate ability to succeed in CS. This suggests that instructor beliefs play a role in the perception of bimodality.

Priming participants to think about the common perception of bimodal grades also led to participants being more likely to label ambiguous distributions as bimodal. This suggests that confirmation bias plays a role in the belief that bimodal grades are typical.

Given that the belief that CS ability is innate is widespread among CS educators, there is likely a social element to the confirmation bias. This belief in bimodality appears related to the belief in innate ability, which in turn has been implicated in the under-representation of women and minorities in computing. We encourage educators reading this paper to take time to analyze the grades in their own classes, and bring the same level of rigor and skepticism we would use in our research to understand our own teaching.

Acknowledgments

The first author received funding from the Social Science and Humanities Research Council of Canada. We would also like to thank our anonymous reviews, Aditya Bhargava, Jinghui Cheng, Jeff Forbes, Jin Guo, Mark Guzdial, Ray Lister, Andrew Petersen, Greg Wilson, and Dan Zingaro for their feedback and suggestions on this line of investigation. C

References

1. Ahadi, A., Lister, R. Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant? In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research*, ACM, 2013, 123–128.
2. Basnet, R.B., Payne, L.K., Doleck, T., Lemay, D.J., Bazelais, P. Exploring bimodality in introductory computer science performance distributions. *EURASIA J. Math. Sci. Technol.*, 14 (2018), 10.
3. Eysenck, M.W., Mogg, K., May, J., Richards, A., Mathews, A. Bias in interpretation of ambiguous sentences related to threat in anxiety. *J. Abnorm. Psychol. 2*, 100 (1991), 144.
4. Goodman, S. A dirty dozen: twelve p-value misconceptions. In *Seminars in Hematology*, Volume 45. Elsevier, 2008, 135–140.
5. Gould, S.J. *The Mismeasure of Man*. WW Norton & Company, 1996.
6. Guzdial, M. Anyone can learn programming: Teaching > genetics, 2014.
7. Guzdial, M. Teaching computer science better to get better results, 2014.
8. Guzdial, M. Learner-centered design of computing education: Research on computing for everyone. *Synth. Lect. Hum. Cent. Inform. 6*, 8 (2015), 1–165.
9. Guzdial, M. Source of the "geek gene"? *Teacher beliefs: Reading on Lijun Ni, learning from Helenrose Fives on teacher self-efficacy*, 2015.
10. Heilman, M.E., Block, C.J., Stathatos, P. The affirmative action stigma of incompetence: Effects of performance information ambiguity. *Acad. Mgmnt. J. 3*, 40 (1997), 603–625.
11. Jost, J.T., Banaji, M.R., Nosek, B.A. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Polit. Psychol. 6*, 25 (2004), 881–919.
12. Kahan, D.M., Peters, E., Dawson, E.C., Slovic, P. Motivated numeracy and enlightened self-government. *Yale Law School, Public Law Working Paper*, (307), 2013.
13. Leslie, S.-J., Cimpian, A., Meyer, M., Freeland, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science 6219*, 347 (2015), 262–265.
14. Lewis, C. Attitudes and beliefs about computer science among students and faculty. *SIGCSE Bull. 2*, 39 (2007), 37–41.
15. Lister, R. Computing education research geek genes and bimodal grades. *ACM Inroads 3*, 1 (2010), 16–17.
16. McDonald, J.H. *Handbook of Biological Statistics*, Volume 2. Sparky House Publishing, Baltimore, MD, 2009.
17. Padavic, I., Ely, R.J. The work-family narrative as a social defense, 2013.
18. Park, T.H., Saxena, A., Jagannath, S., Wiedenbeck, S., Forte, A. Towards a taxonomy of errors in HTML and CSS. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research*, ACM, 2013, 75–82.
19. Payne, B.K., Shimizu, Y., Jacoby, L.L. Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *J. Exp. Soc. Psychol. 1*, 41 (2005), 36–47.
20. Petersen, A., Craig, M., Zingaro, D. Reviewing CS1 exam question content. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, SIGCSE'11 (New York, NY, USA, 2011). ACM, 631–636.
21. Razali, N.M., Wah, Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal. 1*, 2 (2011), 21–33.
22. Rosenthal, R. Teacher expectancy effects: A brief update 25 years after the pygmalion experiment. *J. Res. Educ. 1*, 1 (1991), 3–12.
23. Rossiter, M.W. *Women Scientists in America: Struggles and Strategies to 1940*, Volume 1. JHU Press, 1982.
24. Sahami, M., Piech, C. As CS enrollments grow, are we attracting weaker students? In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE'16 (New York, NY, USA), 2016. ACM, 54–59.
25. Schilling, M.F., Watkins, A.E., Watkins, W. Is human height bimodal? *Am. Stat. 3*, 56 (2002), 223–229.
26. Wikipedia. Multimodal distribution—wikipedia, the free encyclopedia, 2016 [online; accessed 6-April-2016].
27. Wikipedia. Normal distribution—wikipedia, the free encyclopedia, 2016 [online; accessed 6-April-2016].
28. Zingaro, D., Petersen, A., Craig, M. Stepping up to integrative questions on cs1 exams. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*. ACM, 2012, 253–258.

Elizabeth Patitsas (elizabeth.patitsas@mcgill.ca), McGill University Montreal, Québec, Canada.

Jesse Berlin (jesse.berlin1@gmail.com), Toronto, Ontario, Canada.

Michelle Craig and Steve Easterbrook ({mcraig,sme}@cs.toronto.edu), University of Toronto Toronto, Ontario, Canada.