Evidence That Computer Science Grades Are Not Bimodal

Elizabeth Patitsas, Jesse Berlin, Michelle Craig, and Steve Easterbrook
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
patitsas,mcraig,sme@cs.toronto.edu and jesse.berlin@mail.utoronto.ca

ABSTRACT

Although it has never been rigourously demonstrated, there is a common belief that CS grades are bimodal. We statistically analyzed 778 distributions of final course grades from a large research university, and found 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 and asked them to categorize the distributions. 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 [1]). But these discussions do not include statistical testing of whether the CS grades are bimodal in the first place.

From what we've seen, people take a quick visual look at their grade distributions, and then if they see two peaks, they say it's bimodal. But eyeballing a distribution is unreliable; for example, if you expect the data to have a certain distribution, you're more likely to see it.

Anecdotally, we've seen new instructors and TAs (and students) 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. Every community of practice has a knowledge base of beliefs that inform their practice [13], and these beliefs may or may not be based on empirical evidence.

ICER '16 September 08-12, 2016, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-4449-4/16/09.

DOI: http://dx.doi.org/10.1145/2960310.2960312

1.1 Explanations of Bimodality

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.

1.1.1 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 it [1].

High school CS is not common in many countries, and so students enter university CS with a range of prior experience. However, this explanation fits students into two bins. Prior experience is not as simple as "have it" vs. not – there is a large range on how much prior experience students can have programming, and practice with non-programming languages like HTML/CSS could also be beneficial [21].

1.1.2 Learning Edge Momentum, Stumbling Points, and Threshold Concepts

One family of explanations could be summarized as that some CS concepts are more difficult for students to learn, and if they miss these concepts, they fall behind while their peers advance ahead of them [1]. Because CS1 as it is typically taught builds on itself heavily, once a student falls behind, they continue to fall further and further behind [1].

One might think of this explanation as a variant of the prior experience explanation, where the students who succeed have better study skills, and those who fall behind do not.

1.1.3 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" in his writing [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." [15].

However, there seems to be little evidence that there is indeed a "Geek Gene", and that plenty of evidence that effective pedagogy allows for all students to succeed [8].

1.1.4 Lousy Assessment

Another line of explanation implicates instructors' assessment tools as the source of bimodally distributed grades [33, 23]. A common trend on CS exams is to ask a series of long-answer coding questions. Zingaro et al. found that these questions are coarse in terms of the information given 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 this student has of a concept.

As an alternative, Zingaro et al. experimentally compared using short answer questions which build upon each other to having one isomorphic long-answer question. When the different conceptual parts of the question were broken up, the resulting grades were normally distributed, whereas the long-answer questions led to grades that the authors described as bimodal [33].

1.1.5 Or perhaps CS grades are not bimodal?

A competing view of CS grades argued by Lister is that the grades are not, in fact, bimodal [17]. Lister observed that CS grades 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 see bimodality where there is none [17]. Lister's argument was theoretical, and based on statistical theory; in our paper we will test his argument by statistically analysing real world grades 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.

Most standard continuous probability distributions have a mean, a median, a mode, and some measure of the distribution's width (variance). Standard distributions most people might be familiar with include the normal (Gaussian), Pareto, Poisson, Cauchy, Student's t, and logistic distributions. When we plot them with a histogram, we see what's called their probability density.

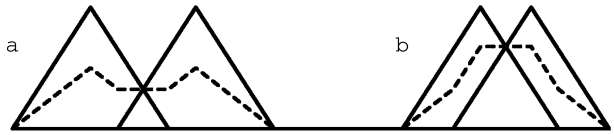
All of these distributions have a single mode, and have a probability density that can be modelled with a function that has a single term. For example, the normal distribution's PDF is:

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

In this function, a represents the height of the curve's peak, b is the position of the centre of the peak, and c represents the width of the curve [31].

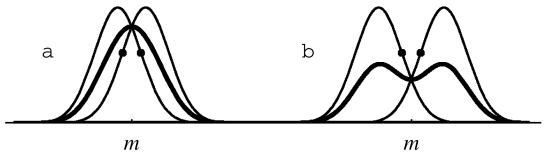
In contrast, a bimodal distribution has two *distinct* modes. A 'multimodal' distribution is any distribution with multiple distinct modes (two or more).

For an example, consider these examples from [28]. 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 sub-distributions are far away (example **a**), we get a distribution with two peaks. But when the two sub-distributions 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 can be seen for normal distributions (also from [28]):



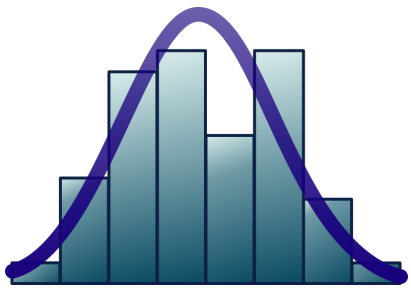
For a distribution to be bimodal, the sub-distributions can't overlap too much. As shown in [28], for the two distributions to be sufficiently far apart, the distance between the means of the two distributions needs to exceed 2σ. This, however, assumes the two distributions have the same variance.

More formally, if the two sub-distributions do not have the same variance, then for their sum to be bimodal, the following must hold [30]:

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

2.1 Real World Data

Consider this histogram of sepal widths for the Iris species *versicolor*, taken from the Wikipedia page on "normal distribution" [31]:



These data have two peaks, but it is considered a normal distribution. If we were to try and model these data as the mixture of two normal distributions, the two sub-distributions would be too close together to produce two distinct peaks. The simplest way to model these data is as a normal distribution.

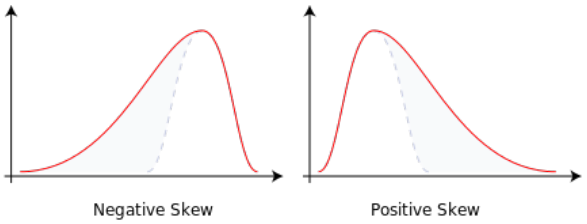
Finally, it must be stressed that what we see in a histogram is a result of how we bin the data. It is possible to bin these data in a way which do not have two 'peaks' (for example, using larger intervals for the bins, or shifting the intervals).

2.2 Skewness and Kurtosis

By definition, a normal distribution is symmetric around its mode (which is also its mean and median). However, many real world data which produce a bell curve when graphed as a histogram do not fit these properties.

2.2.1 *Skewness*

Skewness is a measure of how asymmetric the data are. A distribution with a skewness of zero is perfectly symmetric. In comparison, a distribution with a negative skewness will have a longer 'tail' on the left side than on the right side; the opposite is true of positive skewness [32]:

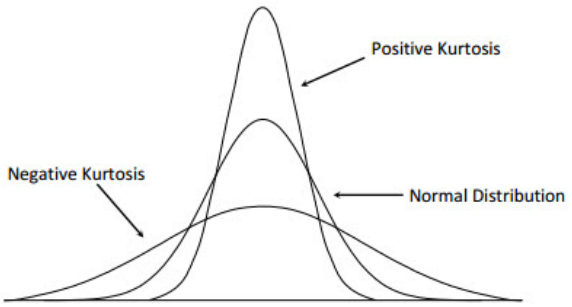


One may expect grades distributions to be skewed. One cause of skewness is the ceiling effect: if students are performing well (and this is normally distributed), and we set a maximum grade of 100%, this will cause the students at the top of the class to be bunched together.

By convention, if the absolute value of the skewness is greater than 1, a distribution is considered highly skewed; an absolute value of skewness between 0.5 and 1 is considered moderately skewed; less than 0.5 is considered approximately symmetric [32].

2.2.2 *Kurtosis*

Kurtosis is a measure of how 'tailed' the data is. A distribution with high kurtosis has a sharp peak and short tails. A distribution with low/negative kurtosis has a low peak and long tails. The normal distribution has a kurtosis of 3. A distribution with a kurtosis greater than this cannot be bimodal [30].



If you look back at the illustration of adding two normal 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.

3. STUDY 1: STATISTICAL ANALYSIS OF GRADES

Are CS grades bimodal, or unimodal? To test this, we acquired the final grades distributions for every undergraduate CS class at the University of British Columbia (UBC), from 1996 to 2013. This represents 778 different lecture sections, containing a total of 30,214 final grades (average class size: 75).

3.1 Testing for normality vs. bimodality

There are a number of ways to test whether some data are consistent with a particular statistical distribution.

One way is to fit your data to whatever formula describes that distribution. You can then eyeball whether your resulting curve matches the data, or you could look at the residuals, or even do a goodness-of-fit test.

Another is to use a pre-established statistical test which will allow you to reject/accept a null hypothesis on the nature of your data. We used this approach, for the ease of checking hundreds of different distributions and comparing them.

There are a large variety of tests for whether a distribution is normal, such as Anderson-Darling and Pearson's chi-squared test. We chose Shapiro-Wilk, since it has been found to have the highest statistical power [25].

There are few tests for whether a distribution is bimodal. Most of them essentially work by trying to capture the difference in means in the two distributions that are in the bimodal model, and testing whether the means are sufficiently separate. We used Hartigan's Dip Test, because it was the only one available in GNU R at the time of analysis.

We also computed the kurtosis for every distribution due to the necessary (but not sufficient) condition of kurtosis < 3 for bimodality [30]. To minimize false positives, we only performed Hartigan's Dip Test on distributions where the kurtosis was less than 3.

We chose the standard alpha value of 0.05. Given that we performed thousands of statistical tests, false positives are inevitable – we expect 5% of our tests will yield a false positive.

3.2 Test results

3.2.1 *Unimodality vs. Multimodality*

Beginning with kurtosis, 323 of the 778 lecture sections had a kurtosis less than 3. This means that 455 (58%) of the classes were definitely not bimodal, and that at most 323 (42%) classes could be bimodal.

Next we applied Hartigan's Dip Test to the 323 classes which had a kurtosis less than 3. For this test, the null hypothesis is that the population is unimodal. As a result, if $p < \alpha$, then we may reject the null hypothesis and conclude we have a multimodal distribution. This was the case for 45 classes (13.9% of those tested, 5.8% of all the classes).

Of the 45 classes which were multimodal, 16 were 100-level classes (35%), 5 were 200-level (11%), 12 were 300-level (27%), and 12 were 400-level (27%). For comparison, in the full set of 778 classes, 171 were 100-level (22%), 165 were 200-level (21%), 243 were 300-level (31%), and 199 were 400-level (26%).

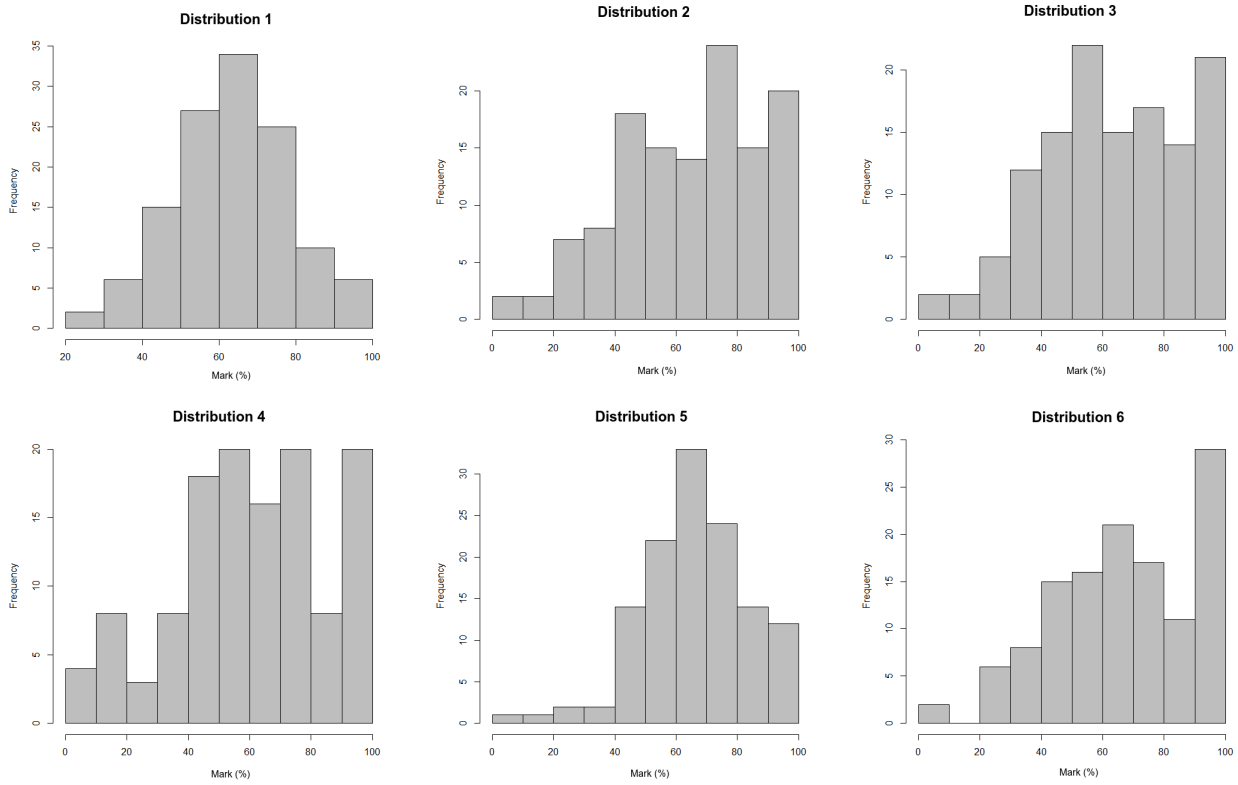


Figure 1: The six histograms shown to participants, all of which were generated using GNU R's `rnorm` function. A ceiling of 100% was used, which is most evident in Distribution 6. Each generated distribution had 100 points, and was generated with an average of 60 and standard deviation of 5.

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 grades 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 grades 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. Questions on the 'Geek Gene':
   - 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")

Table 1: The pages of the survey. Pages 2 and 5 were swapped for a random half of the participants. We chose the all-succeed question because it had been used in [16].

3.2.2 Normality

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 say the population is not normally distributed. This was the case for 106 classes.

44 of the 45 classes which were previously determined to be multimodal were among the 106 classes which the Shapiro-Wilk test indicated weren't normally distributed. In short, 13.6% of the classes aren't normally distributed, many of which are known to be multimodal.

For the 86.4% of classes where we failed to reject the null hypothesis, we can't guarantee that they are actually normal, because of type II error. Fortunately, we have a large sample size and good statistical power. We bootstrapped a likely beta value, providing an estimated false negative rate of 1.48%.

In short, an estimated 85.1% of the final grades in UBC's undergrad CS classes are normally distributed. If CS grades were typically bimodal, we would expect far more than 5.8% of classes to test as bimodal.

3.2.3 Skewness

While most of the distributions appear to be normally-distributed, it is worth noting that the average skewness of all the distributions was -0.33, ranging from -2.30 to 1.02. For just the distributions we'd determined to be normal, the average skewness was -0.13, ranging from -1.11 to 0.84. It is therefore likely that for many of the distributions which are unimodal but not normal, their non-normality is because they are too skewed to pass a test of normality. This may be a result of the ceiling effect in grade distributions.

3.3 Discussion

It is worth noting that we only examined final grades: our analysis did not include term grades.

As grades only came from one institution, one may wonder about the generalizability. We tried to get access to grades distributions from other institutions but generally found it difficult to gather the same scale of data. Analyzing five grades distributions from the University of Toronto, we found them to be normally-distributed.

While we can't assert that every university has the same grades distributions as UBC, the large scale of data both in numbers and time-span gives does give us a great deal of information. More work should be done to replicate our findings at other institutions.

What stood out for us is that at both UBC and UToronto, the CS faculty would routinely assert that their CS grades are bimodal – and we now had evidence to the contrary.

Our results support Lister's argument that CS grades are generally not bimodal, and that the perception of bimodality comes from instructors expecting their grades to be [17].

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, there's often more than one peak. The multiple peaks may be erroneously perceived as "bimodal". A typical "large class" does not have a large enough sample size to consistently provide a smooth bell curve. Indeed, many of the distributions produced by R's `rnorm` looked very much like the grade distributions we'd seen in our own classes and called "bimodal."[1]

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 the Geek Gene 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'd analysed 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 Geek Gene hypothesis? (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 shown in Figure 1, all of which we'd 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 (distribution 5) as though students who were failing were pushed up to 50%, 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 distribution in their own classes (very often to never on a Likert scale), and then how they would categorize the distribution (normal, bimodal, multimodal, uniform, other).

We randomly assigned participants to one of two treatments:

Treatment 0: participants were asked whether they agreed with the Geek Gene Hypothesis, 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 grades distributions, before they saw the distributions; after that we asked them whether they agreed with the the Geek Gene Hypothesis.

The survey had five pages, which are described in Table 1. For each question we created a shorthand, in bold, for use in our analysis.

[1]One may wonder how many of the distributions generated by `rnorm` will test as bimodal per Hargigan's Dip Test. We generated 100,000 distributions with n=100, μ=60, σ=5 and only 133 distributions (1.3%) tested as multimodal per the Dip Test.

	Treatment 0				Treatment 1			
Parameter	2	3	4	5	2	3	4	5
innately-pred		-2.2 (1.2)	-22 (4.5e-2)*		0.2 (1.8)	2.8 (1.8)	5.6 (2.3)*	
all-succeed	-37 (14)*	-35 (14)*	-39 (14)*		3.5 (2.6)	4.6 (2.8)	6.9 (3.2)*	
look-histo	7.0 (57)	6.0 (57)	7.8 (57)	-22 (3.1e-6)*		-2.6 (2.4)*	-3.8 (2.1)*	-6.4 (3.1)*
look-letter	32 (2.7)	1.4 (2.1)	1.0 (2.1)	-4.1 (3.2)		27 (1.9)	29 (0.9)	32 (1.8)

Table 2: Coefficients from the `polr` regression on seeing-bimodality for each treatment; standard errors are in parentheses; * denotes statistical significance.

	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

Table 3: Results of the `Anova` of the regressions on the two treatments; i.e., does the relationship between a given factor and seeing-bimodality differ between the two treatments?

Because so many of the potential participants were our colleagues, we deliberately did not collect names and identifying information about the participants in the survey. 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 to 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[2]. 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. 53 participants completed every question on the survey; 28 were in Treatment 0 (the non-primed group), and 25 were in Treatment 1 (the primed group).

The participants who had provided their emails for follow-up purposes were debriefed. Since fewer than half of the participants had provided their email, we posted open letters to the online communities where we had recruited participants.

4.3 Results

For each participant, we computed a value we'll call "seeing-bimodality," which is the number of distributions they had categorized as bimodal/multimodal. In our data, seeing-bimodality ranged from 0 to 5.

4.3.1 *Regresion on* seeing-bimodality

We wanted to see if seeing-bimodality could be predicted by participants' responses to the questions we'd asked. The regression we performed was to model seeing-bimodality as a function of innately-predisposed, all-succeed, look-histo, and look-letter, using the shorthands from subsection 4.1.

When visualizing the results, we noticed that the relationship between seeing-bimodality and the Likert questions varied between the two treatments. To perform a non-parametric 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.

In doing so we expected to compute 28 p values. Applying a Šidák correction to the standard alpha level of 0.05, we used 0.002 as our alpha level for this section of our analysis.

We found a statistically significant relationship between seeing-bimodality and participants' responses to the questions relating to the Geek Gene hypothesis (all-succeed and innately-predisposed), as shown in Table 2. Furthermore, when it came to all-succeed, the effect was statistically significantly stronger in the treatment which was primed to think about CS grades being bimodal, as shown in Table 3. We also observed there was a strong negative correlation between all-succeed and innately-predisposed.

We also found 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.

4.3.2 *Regression on* all-succeed

After finding a one-way relationship between grade perceptions and the Geek Gene Hypothesis, 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 the Geek Gene. Since 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 [16].

Recall that our study was set up so that a random half of the participants categorized distributions then were asked about the Geek Gene (Treatment 1), and the other half were asked about the Geek Gene and then categorized the distributions (Treatment 0). If there's 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 are that for an alpha level of 0.05, you need 10–20 data points per independent variable in your model [18]. Because this part of the analysis requires the statistical power to reject a null hypothesis, we modelled 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, as shown in Table 4. In Treatment 0, we found that it was not. This indicates that there is a feedback loop between categorizing distributions as bimodal and agreement with the Geek Gene Hypothesis.

We hence have observed evidence for the feedback loops illustrated in Figure 2.

[2]The survey was on SurveyMonkey; signing up for follow-up emails was via Google Forms.

Treatment 0				Treatment 1				
Parameter	1	2	3	Parameter	1	2	3	5
seeing-bimodality	-0.2 (0.9)	-1.1 (1.0)	-0.7 (1.1)	seeing-bimodality	0.6 (1.0)	0.9 (1.2)	1.4 (1.0)	1.7 (3.2e-7)*
intercepts	-3.8 (1.2)	-2.0 (0.8)	-0.3 (0.6)	intercepts	-2.6 (1.1)	0.2 (0.7)	1.5 (0.8)	

Table 4: Coefficients from the `polr` regression on all-succeed for each treatment; standard errors are in parentheses; * denotes statistical significance. p values were calculated from z values using `coeftest`.

4.4 Discussion

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 the Geek Gene Hypothesis.

Our approach to priming may have led participants to believe more that grades are bimodal. Because the survey presents us, the researchers, as authority figures, and we imply that grades are thought to be bimodal, some participants could assume it to be true since we said so.

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 had hoped to recruit a larger number of participants; however, recruiting a large number of CS educators to fill out the survey turned out to be infeasible with our resources. It must be noted that 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 [2], which could have influenced their responses.

But we would expect the SIGCSE community to be *less* inclined to believe in the Geek Gene hypothesis 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.

4.4.1 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]. As another example, Eyesnck et al. studied the interpretation of sentences as either threatening or non-threatening by people who have anxiety and by a control group [4]. They found that unambiguously threatening/non-threatening sentences were interpreted similarly between groups, but participants with anxiety were more likely to label ambiguous sentences as threatening than participants in the control group. Visual information is subject to this phenomenon also: Payne et al. showed participants a series of photos of black and white 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 [22].

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.

As for our evidence that looking at histograms reinforces belief in the Geek Gene Hypothesis, *systems justification theory* explains that once you are forced to take a position on a subject, you're more likely to believe and defend it [11].

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 would go a step further and argue that the perception of bimodality is a *social defense* in the CS education community.

5.1 What is a Social Defense?

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" [20].

For example, Padavic et al. [20] found that the "work-family" narrative in business is an example of a social defense: people will say that women leave the workplace because of "family", despite the large amount of evidence that women leave their jobs because of inadequate pay or opportunities for advancement [20], particularly when they see male co-workers promoted ahead of them. The "work-family" narrative is a more palatable explanation rather than to confront sexual discrimination in the workplace, and so the narrative continues.

5.2 Teacher Self-Efficacy

Guzdial reported that, per Fives [9], 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. With time, their self-efficacy slowly increases again. [9]

Teacher self-efficacy is not necessarily tied to how well they can teach: university educators often get little meaningful feedback on how their students are learning, given their large class sizes and lecture-based pedagogies. [9]

Guzdial reasoned that if an individual university-level CS educator has high self-efficacy, and sees evidence of students not learning, then it's rational for them to believe that the problem lies with the students and that the problem is innate to them—i.e., beyond the ability of the teacher to improve

it [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 [27].

Relatedly, Guzdial noted that CS educators have poor results, because we so frequently use ineffective teaching methods [7]. Indeed, Porter et al. recently found that performance on early assessments in CS1 correlate highly with final grades, indicating that surprisingly little learning goes on in CS1 [24]. The results of Zingaro, Petersen, and Craig would add that not only do CS educators frequently use ineffective pedagogies, they also frequently use ineffective assessment tools [33, 23].

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 "don't 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 phenomena in CS classes. When administrators accept this narrative and do not mandate professors to improve their teaching, the narrative can continue unchallenged.

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 begin teaching, do not see all their students learning, and have been primed by colleagues to see bimodality, the new educator can then see this as evidence of the Geek Gene. The reproduction of the Geek Gene Hypothesis is hence social in nature.

Recent studies have found that academic disciplines in which "brilliance" is seen as necessary for success have less demographic diversity [14]. Looking at the history of science, women and people of colour were long denied entry and acknowledgment in science because they were seen as lacking the "brilliance" needed to do science [26].

If computing ability is viewed as being the result of a "Geek Gene", then educators may use this as an reason not to teach students who lack this "gene". Similarly they could lower expectations of these groups and encourage them less. Research on implicit biases consistently find that implicit biases against seeing women and people of colour as being brilliant scientists [29]. Students with disabilities or attention disorders could also be affected, or whoever else a particular educator might see as lacking the "gene". The "Geek Gene" narrative can also contribute to how women and minorities feel they do not belong in CS classes. It has been documented that underrepresented groups feel demotivated when their more experience peers boast that CS is "easy", and this could trigger stereotype threat [3].

6. CONCLUSIONS

Our analysis of UBC's grades indicates that while bimodal grade distributions can be found, they are far from typical (at most 5.8% of cases given type I error). Much more commonly, grade distributions are normal (85.1%) or skewed.

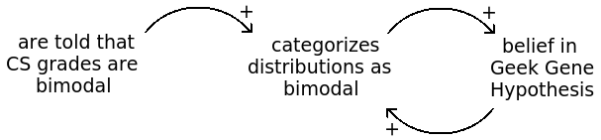


Figure 2: Individual-level feedback loops leading individuals to categorize ambiguous distributions as bimodal.

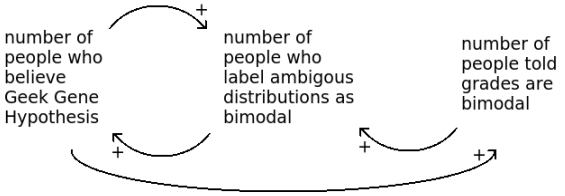


Figure 3: Social-level feedback loops leading individuals to categorize ambiguous distributions as bimodal.

Our psychology experiment found that priming participants to think about the common perception of bimodal grades leads to participants being more likely to label ambiguous distributions as bimodal. This indicates confirmation bias plays a role in the belief that bimodal grades are typical, when our (more rigourous, less anecdotal) evidence is that they are uncommon.

We also found that participants who reported beliefs consistent with the Geek Gene Hypothesis were more likely to label ambiguous distributions as bimodal. This indicates instructor beliefs play a role in perception of bimodality.

We observed that instructors who report looking at histograms of their grades were more likely to label ambiguous distributions as bimodal. As well, the random half of participants who labelled distributions as bimodal and then were asked about the Geek Gene Hypothesis were more likely to agree with it than the random half of participants who had been asked about the Geek Gene first.

Both our analysis of UBC's grades and our psychology experiment provide evidence for Lister's hypothesis that CS grades are not typically bimodal.

We theorized that the perception of bimodal grades in CS is a social defense. It is easier for the CS education community to believe that some students "have it" and others do not than it is for the community to come to terms with the shortfalls of our pedagogical approaches and assessment tools. A belief in the Geek Gene gives educators an easy way out from confronting these issues and being pushed to do better. In order for efforts to have CS taught "for all" to succeed, the CS education community needs to develop and use pedagogical approaches and assessment tools that will benefit all students.

7. ACKNOWLEDGMENTS

The first author received funding from the Social Science and Humanities Research Council of Canada. We would also like to thank our anonymous reviewers for their feedback, as well as Andrew Petersen, Jeff Forbes, and Aditya Bhargava for their suggestions.

8. REFERENCES

[1] A. Ahadi and R. Lister. 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*, pages 123–128. ACM, 2013.

[2] A. Ahadi and R. Lister. 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*, pages 123–128. ACM, 2013.

[3] C. Ashcraft, E. Eger, and M. Friend. *Girls in IT: The Facts*. National Center for Women & Information Technology, 2012.

[4] M. W. Eysenck, K. Mogg, J. May, A. Richards, and A. Mathews. Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of abnormal psychology*, 100(2):144, 1991.

[5] S. J. Gould. *The mismeasure of man*. WW Norton & Company, 1996.

[6] M. Guzdial. Anyone can learn programming: Teaching > genetics, 2014.

[7] M. Guzdial. Teaching computer science better to get better results, 2014.

[8] M. Guzdial. Learner-centered design of computing education: Research on computing for everyone. *Synthesis Lectures on Human-Centered Informatics*, 8(6):1–165, 2015.

[9] M. Guzdial. Source of the "geek gene"? teacher beliefs: Reading on lijun ni, learning from helenrose fives on teacher self-efficacy, 2015.

[10] M. E. Heilman, C. J. Block, and P. Stathatos. The affirmative action stigma of incompetence: Effects of performance information ambiguity. *Acad. of Mgmnt. J.*, 40(3):603–625, 1997.

[11] J. T. Jost, M. R. Banaji, and B. A. Nosek. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology*, 25(6):881–919, 2004.

[12] D. M. Kahan, E. Peters, E. C. Dawson, and P. Slovic. Motivated numeracy and enlightened self-government. *Yale Law School, Public Law Working Paper*, (307), 2013.

[13] J. Lave and E. Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[14] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[15] C. Lewis. Attitudes and beliefs about computer science among students and faculty. *SIGCSE Bull.*, 39(2):37–41, June 2007.

[16] C. Lewis. Attitudes and beliefs about computer science among students and faculty. *SIGCSE Bull.*, 39(2):37–41, June 2007.

[17] R. Lister. Computing education research geek genes and bimodal grades. *ACM Inroads*, 1(3):16–17, 2010.

[18] J. H. McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

[19] D. H. Meadows. *Thinking in systems: A primer*. Chelsea Green Publishing, 2008.

[20] I. Padavic and R. J. Ely. The work-family narrative as a social defense, 2013.

[21] T. H. Park, A. Saxena, S. Jagannath, S. Wiedenbeck, and A. Forte. Towards a taxonomy of errors in HTML and CSS. In *Proceedings of the ninth annual international ACM conference on International computing education research*, pages 75–82. ACM, 2013.

[22] B. K. Payne, Y. Shimizu, and L. L. Jacoby. Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, 41(1):36–47, 2005.

[23] A. Petersen, M. Craig, and D. Zingaro. Reviewing CS1 exam question content. In *Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education*, SIGCSE '11, pages 631–636, New York, NY, USA, 2011. ACM.

[24] L. Porter, D. Zingaro, and R. Lister. Predicting student success using fine grain clicker data. In *Proceedings of the Tenth Annual Conference on International Computing Education Research*, ICER '14, pages 51–58, New York, NY, USA, 2014. ACM.

[25] N. M. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[26] M. W. Rossiter. *Women scientists in America: Struggles and strategies to 1940*, volume 1. JHU Press, 1982.

[27] M. Sahami and C. Piech. As CS enrollments grow, are we attracting weaker students? In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 54–59, New York, NY, USA, 2016. ACM.

[28] M. F. Schilling, A. E. Watkins, and W. Watkins. Is human height bimodal? *The American Statistician*, 56(3):223–229, 2002.

[29] J. G. Stout, N. Dasgupta, M. Hunsinger, and M. A. McManus. Steming the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (stem). *Journal of personality and social psychology*, 100(2):255, 2011.

[30] Wikipedia. Multimodal distribution — wikipedia, the free encyclopedia, 2016. [Online; accessed 6-April-2016].

[31] Wikipedia. Normal distribution — wikipedia, the free encyclopedia, 2016. [Online; accessed 6-April-2016].

[32] Wikipedia. Skewness — wikipedia, the free encyclopedia, 2016. [Online; accessed 6-April-2016].

[33] D. Zingaro, A. Petersen, and M. Craig. Stepping up to integrative questions on cs1 exams. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pages 253–258. ACM, 2012.