INVESTIGATING STUDENTS' REASONING ABOUT SAMPLING DISTRIBUTIONS THROUGH A RESOURCE PERSPECTIVE

KELLY FINDLEY
Florida State University
kfindley@fsu.edu

ALEXANDER LYFORD
Middlebury College
alyford@middlebury.edu

ABSTRACT

Researchers have documented many misconceptions students hold about sampling variability. This study takes a different approach—instead of identifying shortcomings, we consider the productive reasoning pieces students construct as they reason about sampling distributions. We interviewed eight undergraduate students newly enrolled in an introductory statistics course. Taking a grounded theory style approach, we identified 10 resources that students used when reasoning about the sampling distribution for the average within two contexts: penny years and dice rolls. Students had varied success in their responses as they made choices about how to represent their resources in their constructions. Successful constructions exemplified careful blending of resources, while less successful constructions reflected disjoint perceptions and tensions between seemingly conflicting resources. Our findings stress the importance of framing students as capable reasoning agents by describing student resources that were used while solving tasks related to sampling distributions. We also discuss the influence of context and problem setting in students' reasoning and resource elicitation.

Keywords: *Statistics education research; Constructivism; Statistical reasoning; Conceptual blending*

1. INTRODUCTION

The heart of statistical reasoning lies in connecting the core statistical concepts of sampling, variability, and distribution into a unified conceptualization of sampling distributions (Garfield & Ben-Zvi, 2008). When students struggle to reconcile these ideas meaningfully, they may resort to procedural, cookbook approaches when applying formal inferential methods (Garfield, delMas, & Zieffler, 2012).

The statistics education literature has traditionally discussed students' statistical reasoning in terms of common misconceptions (e.g., Chance, delMas, & Garfield, 2004; Cooper & Shore, 2008; Lane-Getaz, 2017; Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). *Misconceptions*, as defined by Sotos et al. (2007), may represent "any sort of fallacies, misunderstandings, misuses, or misinterpretations of concepts, provided that they result in a documented systematic pattern of error" (p. 99).

As educational researchers and instructors, it is often easier to discuss what students do *not* know rather than what they do know (Sewell, 2002). Sewell explained that constructivist-based philosophies toward learning that truly value students' pre-existing ideas are greatly challenged if we believe students need to eradicate their wrong ideas in order to adopt correct ones (e.g., Eaton, Anderson, & Smith, 1984). That is not to say that labeling common misconceptions is inherently problematic in educational research, but rather that such terminology should be used thoughtfully:

> Unfortunately, most people … [see] misconceptions as impediments, and in this way, misconceptions research has had almost the opposite effect from the researchers' intentions. Instead of raising respect for students' prior understandings, it has convinced many educators that students

© International Association for Statistical Education (IASE/ISI), May, 2019

> are worse than blank slates; they're slates with wrong ideas written on them in hard-to-erase chalk. (Hammer & Van Zee, 2006, p. 15)

An alternative research perspective is to investigate the productive steps students take in their reasoning by identifying the *resources* they bring to a task (Smith, diSessa, & Roschelle, 1994). These resources may be conceptualized as *Knowledge in Pieces*, representing fine-grained intuitions drawn from experiences and activated in multiple contexts where the learner identifies potential connections (diSessa, 1988). diSessa presented the notion "closer means stronger" as a common resource students utilize to make sense of the physical world (e.g., sound or heat is heard or felt stronger as one moves closer to the source). This fundamental pattern, however, can also be applied in inappropriate contexts (e.g., the temperature is hotter in summer because our hemisphere faces the sun more directly, not because Earth is closer to the sun). Resources, when blended appropriately, serve as building blocks for constructing complex conceptions (Smith et al., 1994). It follows that researchers and instructors should be responsive to the resources students apply to statistical problems. "Learning is enhanced when teachers pay attention to the knowledge and beliefs that learners bring to a learning task, use this knowledge as a starting point for new instruction, and monitor students' changing conceptions as instruction proceeds" (Garfield & Ben-Zvi, 2009, p. 73).

By analyzing interviews with eight undergraduate students enrolled in an introductory statistics course at a large, public university, we examined the productive ideas students shared that could be leveraged toward conceptual understanding of the sampling distribution of the mean. Students completed two similar tasks with different contexts. Each task asked students to construct a sampling distribution for the sample averages and explain how its shape would change as the sample size increased. Our research questions are:

1) What resources do newly-enrolled students in introductory statistics elicit when reasoning about sampling distributions?
2) How do students reconcile multiple, seemingly contradictory resources in their attempts to make sense of sampling distributions?
3) How—if at all—does the context and problem setting influence students' resource elicitation?

2. LITERATURE REVIEW

2.1. CURRICULAR INCLUSION OF SAMPLING DISTRIBUTIONS

Our research questions assume that undergraduate students should develop a deep, conceptual understanding of sampling distributions and the Central Limit Theorem (CLT). However, a growing number of statistics educators question the value of inferential statistical testing (IST) in the applied statistics curriculum (White & Gorard, 2017). For this reason, we first examine whether an in-depth exploration of sampling distributions and central tendency is useful in an introductory statistics course.

We acknowledge that the needs of every student and course are different. Thus, any argument for curricular inclusion should be made contextually rather than universally. Many proposed introductory course formats that focus predominately on basic research design and statistical literacy reasonably preclude much theoretical content (e.g., Baglin, Reece, & Baker, 2015; Prodromou & Dunn, 2017). Foregoing or limiting course time on sampling distributions and the CLT seems reasonable for such courses.

Shared concerns over teaching IST center on the frequency for which IST assumptions are unmet in research contexts, as well as the convuluted meaning of *p*-values and confidence intervals in these paradigms (Nicholson & Ridgway, 2017; White & Gorard, 2017). Regardless of the future for quantitative research methods, we believe that any curriculum designed to prepare students to participate in quantitative research would be remiss to avoid acquainting students with the statistical methods found in their research literature for the previous several decades. Arguably, any movement away from IST methods in quantitative research methods should require students to have *more* focused instruction on the CLT and the assumptions of IST to understand the reasoning for such a paradigm-shifting decision for their discipline.

Among alternative methods proposed for introductory courses, instructors may employ simulation-based inference (e.g., permutation tests) as a safer choice based on its more intuitive *p*-value interpretations and lack of a normality assumption (e.g., Garfield et al., 2012). Even if simulation-based

inference replaced parametric testing in the introductory course curriculum, students must still reason about sampling distributions in this testing context, albeit with decreased need to understand central tendency. Garfield and Ben-Zvi (2008) emphasized that coursework involving sampling distributions should give ample opportunity for students to reason about the phenomenon rather than assuming students can merely accept the definition and representation without later content difficulty.

2.2. STUDENT REASONING ABOUT SAMPLING VARIABILITY

On the topic of sampling variability, many researchers have found that students commonly express one of two extreme perspectives on sample-population relationships. The first is believing that a sample will represent the population perfectly, and the second is believing that samples are unpredictable and unrepresentative of the population (Braham & Ben-Zvi, 2017; Prodromou & Pratt, 2006; Shaughnessy, 2007; Watson, Callingham, & Kelly, 2007). Prodromou and Pratt (2006) discussed the former as a modeling perspective—students reason about long-term, overall results. The latter is termed a data-centric perspective—students recognize the inherent variability from outcome to outcome.

Pratt (2000) and Pratt, Johnston-Wilder, Ainley, and Mason (2008) found that the success of 10- and 11-year-olds in reasoning about sample likelihood is anchored in their balancing of both local and global perspectives. A local perspective, as defined by the authors, involves eliciting resources about trial-by-trial outcomes. These short-term results accentuate the unpredictability inherent in small samples. In contrast, a global perspective allows students to see predictability and stabilization in results. In Pratt et al., manipulating the parameters in a certain computer micro-world allowed students little control over short-term results, but prominent control over the long-term. The authors reported students struggling to take a global perspective as they typically did not take large samples on their own to see how their sample statistics would stabilize and converge over time. Only when students were guided to take very large samples (i.e., over 200) were they eliciting both local and global reasoning. Additional scaffolding from the instructor and contextual cues from the simulation environment aided students in reconciling both perspectives.

Another layer of complexity is added when students must reason about the characteristics of multiple independent samples. In their varied tasks, Shaughnessy and colleagues (e.g., Noll, Shaughnessy, & Ciancetta, 2010; Watson & Shaughnessy, 2004) investigated students' beliefs about the reasonableness of a sampling distribution for proportions given certain characteristics about the population. Under the constraint of taking a limited number of samples (e.g., 10 or 100), the researchers found that students at all levels typically overestimated the likelihood of extreme sample proportions. Much of this error seems to be rooted in students' difficulty to employ proportional reasoning in their responses—failing to see the occurrence of an extreme result being accompanied by a large number of non-extreme results. Similar to the findings from Pratt et al. (2008), we view these students as balancing the local and global perspectives, but from the perspective of a distribution of sample averages rather than observations from a single, growing sample.

Braham and Ben-Zvi (2017) studied how middle-school students developed correct conceptions of sampling distributions across several weeks of activities. By alternating between activities with real samples in the context of an unknown population and simulated samples from a known population, students expressed different perspectives about the composition of the sampling distribution. Students typically started with a relativistic perspective of sampling distributions (i.e., unpredictable), followed by a uniformly representative sampling distribution, and finally a narrowing bell-shaped sampling distribution.

2.3. THEORETICAL PERSPECTIVE—RESOURCES AND MISCONCEPTIONS

One theoretical perspective for approaching student understanding is to identify the conceptual resources students demonstrate—ideas that have productive applications in certain contexts, but may not necessarily be universally correct (Smith et al., 1994). These may include approaches, perspectives, and strategies that can be used in the construction of a conception. For example, the works of Kahneman and Tversky (1972) and Konold, Pollatsek, Well, Lohmeier, and Lipson (1993) explored the "heuristics" students would apply as they reasoned about likely events from repeated coin flips. These heuristics represented fundamental building blocks students elicited from their broader conceptual

knowledge as they made arguments about the likelihood of certain results. Furthermore, Pratt et al. (2008) more recently used a resource framing to understand how students make connections between theoretical probability and long-term sampling results.

The common misconceptions that have been identified in the literature contribute another research perspective that offers important insight on student thinking. Chance et al. (2004) summarized several misconceptions in student thinking on the topic, including (p. 302):

- [Thinking] sampling distributions should look more like the population as the sample size increases (generalizes expectations for a single sample of observed values to a sampling distribution).
- [Predicting] that sampling distributions for small and large sample sizes have the same variability.
- [Believing] sampling distributions for large samples have more variability.
- [Confusing] one sample (real data) with all possible samples (in distribution) or potential samples.
- [Thinking that] the mean of a positive skewed distribution will be greater than the mean of the sampling distribution for samples taken from this population.

As was the case with Kahneman and Tversky's (1972) heuristics, many of these misconceptions simultaneously hint at productive observations that students make.

We find these perspectives on student thinking to be incomplete without proper investigation of how the ideas stem from more fundamental conceptual resources. Garfield, Le, Zieffler, and Ben-Zvi (2015) noted the limited value of *only* identifying misconceptions, describing these findings as drawing attention to gaps in students' understanding rather than their potential for conceptual progress towards more expert thinking. The reflections of Garfield et al. on this matter echo the call from Hammer and van Zee (2006) that researchers should not view common student errors as mistakes to be erased, but ideas to be nurtured.

Still, identifying productive seeds in students' thinking is not as simple as avoiding the term *misconception*. For example, Zieffler et al. (2008) described the heuristics identified by Konold et al. (1993) as the "systematic and persistent errors people make when attempting to make decisions involving chance and uncertainty" (p. 3). There is nothing remarkable about the activation of conceptual resources without a corresponding attempt from the learner to make deeper conceptual sense of these resources. We point to Fauconnier and Turner's (2003) notion of conceptual blending as the theoretical basis for this point. Fauconnier and Turner explained that knowledge is created as individuals learn to link distinct objects together to form an updated conceptual basis, leading to new schema from which to analyze and make sense of incoming information. Thus, it is not the presence of resources themselves that lead students to conceptual gains, but rather the meaningful blending of resources into new conceptual structures (Smith et al., 1994).

Our work presents an initial snapshot of college students' intuitions and ideas as they enter the introductory statistics course. We discuss the resources they used as they treaded unfamiliar conceptual territory. Similar to Pratt et al. (2008), we consider whether context and problem characteristics are related to how students reason statistically. Hjalmarson, Moore, and delMas (2011) discussed context as an entry point to a task that may elicit inspiration with regard to constructing measures and methods to make sense of data. In the context of statistical and probabilistic reasoning, researchers report that students use different and sometimes contrarian approaches in similar tasks with different characteristics (e.g., Konold et al., 1993). In response, our research examines student reasoning across two similar problems with different contextual characteristics.

3. METHODS

3.1. SETTING

This study was conducted with eight students enrolled in an introduction to applied statistics course (enrollment of 500) at a large, public university in the Southeastern United States. The students responded to an in-class invitation for participants who were taking their first college statistics course. Each participant completed a 30-minute interview during the second week of class during which they

were asked to complete two open-ended tasks involving the sampling distribution of the sample average. At the time of each interview, the course had only covered summary statistics and visual displays of data.

Details about the participants are shown in Table 1. Noah and David were two and four years removed from their high school AP statistics courses respectively.

Table 1. Participants

Pseudonym	Major	Statistics Coursework
Noah	Psychology & Criminology	AP Statistics
Elijah	Exercise & Physiology	None
Holland	Pre-Med	None
Mika	Psychology	None
Lily	Family & Child Sciences	None
David	Geology	AP Statistics
Polina	Psychology	None
Robert	Family & Child Sciences	None

3.2. DATA COLLECTION

Data sources included videos from each of the eight interviews, interview transcriptions, and student drawings/written work. Each task—henceforth referred to as the Penny task and Dice task—contained three prompts (full text of the prompts for both tasks can be found in the Appendix). The first prompt of the Penny task asked students to think about the production years of pennies in circulation, including the range of years one might reasonably see and the number of pennies one would expect to find for each of these years. Using the *x-y* axes on the page, students were then guided by the interviewer (first author) to draw a graph representing the population distribution of the production years of pennies. Elijah's drawing for this prompt is shown as an example (Figure 1). Although the focus of this study was on reasoning about sampling distributions and not population distributions, we wanted students to take ownership of the task from the beginning rather than to provide them with a pre-determined population shape.

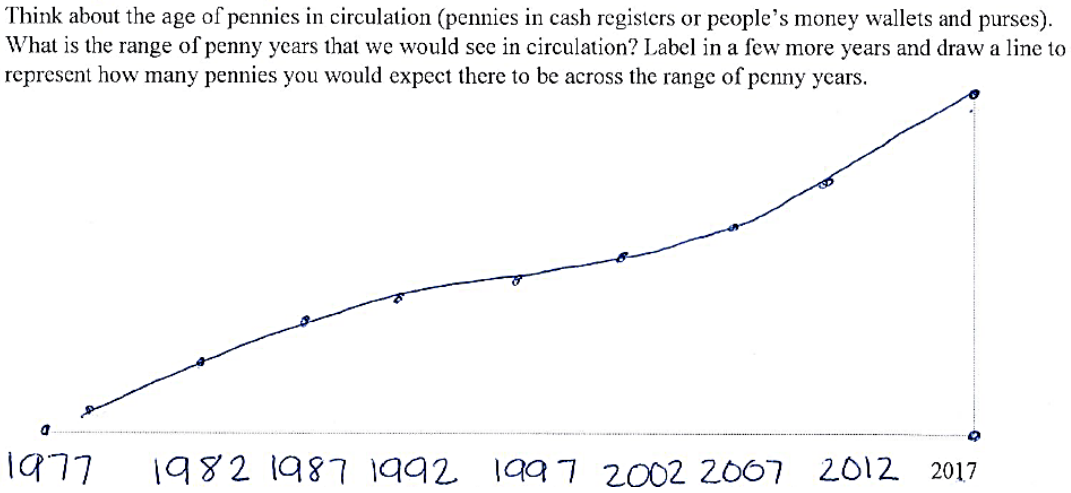


Figure 1. Elijah's penny population

The second prompt of the Penny task asked students to think about taking two pennies randomly from the population and recording their average year of production. Students were then asked to consider the range of possible two-penny averages they would reasonably expect to see and to draw a line graph on a new set of *x-y* axes to represent which averages were more likely than others. The third prompt asked students to repeat the second prompt, but to consider taking 10-penny averages. For each

of the three prompts, the interviewer encouraged students to share their thinking out loud, offering additional scaffolding for each task as needed to ensure students understood what was being asked.

The Dice task was similar to the Penny task and involved reasoning about outcomes from six-sided dice. The first prompt asked students to imagine one million people each rolling a six-sided die and recording the number that came up. Students drew a bar graph or line across an *x* axis labeled 1 through 6 to report approximately how many rolls they would expect to see across this range of outcomes. The second and third prompt were identical to that of the Penny task, except in the context of rolling dice instead of selecting pennies out of the population.

These tasks were chosen carefully to capture student reasoning in two different settings to ensure that the resources identified were not specific to the context.. The Penny task, which we hypothesized would be more difficult and unfamiliar, was used intentionally to motivate students to draw on their own resources. This task was always presented first, because we expected students to be more familiar with the dice context and didn't want students to be able to simply carry over patterns from the Dice task to the Penny task without fully engaging in the context of the latter.

3.3. METHODS OF ANALYSIS

In our analysis, we sought to discover the resources students brought into these tasks on sampling distributions. Rather than begin with a pre-existing framework (e.g., Chance et al., 2004, list of misconceptions), we took a Grounded Theory-based approach (Corbin & Strauss, 2008) by looking at our data and asking, "What is this student getting right?" We began by examining points in the interview where students provided a clear claim and justification for that claim. For example, in response to the third prompt of the Penny task, Polina claimed "I think the larger sample you take, the more representative it will be of all the pennies." Claims like these, followed by the explanations and additional observations students related to these claims, served as the units of analysis for this investigation. These statements were identified and categorized through a process of open coding. After open coding three interviews separately, the authors then convened to discuss the categories and independently coded two additional interviews, meeting again to discuss coding agreement and emergence of new codes. The remaining interviews were coded in a similar fashion.

Next, we summarized and discussed students' claims and justifications in greater detail. In the early stages, we had numerous resource codes, with some focusing more on productive elements and some focusing more on shared claims. By examining each interview more comprehensively (beyond the clear claims and justifications), we gradually became more comfortable condensing our codes, concluding with the list of 10 resources outlined in the Results section. This collection of resources helps answer our first research question.

To answer the second research question, we worked to understand how students coordinated multiple resources. The first author read through each interview transcript again and summarized each interview in terms of reasoning phases. Each phase was characterized by the resource(s) used in the construction of a claim or the direct comparison of two claims. New phases were identified when the student changed perspective, or moved on to a new prompt. By examining phases, we identified interesting interview vignettes that demonstrate how students blended resources in their constructions or grappled with seemingly opposing resources. We constructed explanations to make sense of how each student was using resources to come to conclusions and how tension among competing resources were ultimately resolved—if resolution occurred. We share three interview vignettes in this paper to demonstrate how resources were used collectively to construct more complex understanding, or seemed to be irreconciable in other cases.

In deconstructing the interview phases, the task context appeared to be a central influence on students' reasoning. We created a comparison table (Table 2) to display the resources students used in the interviews when reasoning in the Penny task versus the Dice task, revealing marked differences in the reasoning used in each context. The answer to our third research question emerged as we attended to understanding differences in student reasoning between contexts. We address this question by providing a narrative of David's reasoning, including careful exmination of his work in both task settings. Through this narrative, we provide insight into how characteristics of each task apparently influenced his reasoning, both from our inferences and David's own reflections.

Table 2. Identified resources

Categories	Descriptions	Productive application
Repeated Data Values	Student recognizes that getting the same data value (or type) several times in a row is unlikely (e.g., I probably would not pick up two 2017 pennies in a row or two really old pennies in a row).	Helps student recognize averages at the extremes have fewer possible sample combinations to produce them.
Modeling Likelihoods	Student calculates or reasons about likelihoods to generate a response (e.g., when rolling dice, there is one way to get an average of 1, two ways to get an average of 1.5, etc.).	Orients student to a theoretical justification for a certain shape.
Average Relates to Middle	Student associates average with middle, such as the middle of a sample or middle of the population range (e.g., the average will be in the middle).	A symmetrical sample will average at the midpoint, which may be an entry point to considering what happens with non-symmetrical samples.
Average Relates to Peak	Student associates average with the peak (mode) of the population distribution, believing that the position of the peak is relevant to predicting where averages will cluster (e.g., there are a lot more 2000 pennies, so we will see more averages at 2000).	The position of a peak is a visual cue that may help the student decipher where the population balances.
Sampling Distribution Resembles Population Shape	Student sees the sampling distribution being responsive to the population shape. This may be demonstrated as identical reflection between the two, or sharing shape attributes (e.g., the sampling distribution will also have a dip after 2001).	The position and shape of a sampling distribution for a small sample size depends on the shape of the population.
Growing Possibilities	Student recognizes that there are more potential unique samples when we take larger samples (e.g., If you just take 2, I think it will be clustered around the middle, but if you take 50, it could be anywhere).	The number of additional sample possibilities grows exponentially as n increases. Student may find that values near the edges have fewer relative sample combinations that produce averages there.
Widening Range of Values	Student associates larger samples with having better representation across the range, including outliers (e.g., With 10 pennies, now I would expect to see some really old pennies).	Recognizing the growing range of values may be a gateway to seeing how a sample that properly reflects the range will average at a more consistent balance point.
Stabilizing	Student uses language that suggests stabilizing to describe what happens when we take averages from larger samples (e.g., the sampling distribution will smooth out).	Student may relate the settling nature of larger samples to the consistency of averages clustering at the balance point.
Better Representing Population	Student associates larger samples with better representation of the population. (e.g., With 50 pennies, I would expect that to go back to the population shape).	Student may see that better representation of the population produces sample averages that better represent the population average.
Becoming More Accurate	Student describes larger samples as producing and representing something more accurate (e.g., With that many pennies, my averages should be more specific; The average would not be an outlier.).	Student connects the accuracy of larger samples to sample averages that more accurately converge to the population average.

4. RESULTS

4.1. RESOURCES IDENTIFIED

In Table 2, we list the 10 resources that emerged from our analysis, each pertaining to students' ideas or approaches to reasoning about the form and behavior of (population and) sampling distributions. Alongside, we offer brief descriptions and examples of how the resources were evidenced in our data. Also in Table 2, we discuss how each resource could be used productively in developing an enhanced understanding of sampling distributions. Although not every productive application in the table was observed in the data (e.g., *Average Relates to Middle*, *Average Relates to Peak, Widening of Range of Values*), we include them to demonstrate how these resources may potentially support deep, conceptual thinking. Sections 4.2 and 4.3 demonstrate student examples involving productive applications embedded within richer narratives.

4.2. RESOURCES AS BUILDING BLOCKS

As discussed in Section 2.3, exclusive reliance on individual resources may lead the learner toward or away from building a conception that cannot hold under new situations (Smith et al., 1994). For example, a student who only associates average with middle will theoretically encounter more conceptual difficulties in certain contexts (e.g., viewing a skewed distribution) than a student who sees the population's middle, peak(s), and overall shape as valuable information for determining how sample averages are distributed. In this manner, resources may be viewed as tools, where a limited collection of tools constrains what students can build, whereas a varied collection of tools can afford students more nuanced constructions. Although the utilization of multiple resources does not guarantee a correct understanding (as we will see with David's work in Section 4.3), we do argue that the presence of multiple resources can be a platform for productive thinking, with students' thoughtful blending of resources being key to success.

In our analysis, we found that individual students had varying success in combining and reconciling multiple resources into their responses.We now examine moments from Lily, Polina, and Holland's interviews that exemplify both difficulty and success in blending resources.

Lily's Struggle to Reconcile Resources While reasoning about likely averages from two pennies (Penny task, second prompt), Lily faced difficulty determining the shape of the sampling distribution. At different points in the interview, she thought the shape of the sampling distribution for averaging two pennies might mirror the shape of the population distribution, flatten out, or narrow in the center. She felt overwhelmed by the many potential sample combinations, using her fingers to point out the seemingly scattered and unpredictable nature of taking samples. Reasoning about averages from 10 and 50 pennies only exacerbated the numerous possibilities and apparent chaos of representing likely averages.

Lily used the *Growing Possibilities* resource as a common justification for several different conclusions. She initially used it to support her feeling that the sampling distribution would return to the population shape: "I just feel like, it would kind of … go back to this [population distribution]." She also used this resource as justification for a uniform shape: "If you pick like a penny, 1900, 2017, and then you pick eight of them in here [between], I feel like, it kind of could go anyway, I feel like the range … would it be like more evenly distributed?"

The *Becoming More Accurate* resource was used separately from the *Growing Possibilities* resource. Lily noted that as the sample size increases, you are less likely to observe sample averages on the extremes of the domain:

> Lily: I feel like if you did more, I feel like compared to [$n = 2$], it would do that same little bell cluster, just because the major quantity is around, like around here [middle], so I feel like if you had 10 pennies, like what are the odds that you're going get 10 pennies all the way over here [left], or 10 pennies all the way over here [right].

Lily indirectly used the *Repeated Values* resource to justify this movement towards accuracy. However, Lily again wavered as she returned to her previous solution path:

Lily:	I feel like it would just kind of keep getting more narrow
Interviewer:	Ok, do you think you could draw what you imagine happening … with like 50 pennies at a time?
Lily:	Or I feel like honestly, it would kind of get more, even. Even.
Interviewer:	So when you say even, what kind of motivates that?
Lily:	If it's two, then it's more likely to be clustered around the median I guess, but if you're doing like 50, you could take 20 from [right], 10 from [middle], and another 20 from [left], and it would kind of even out, and still be in the middle. Or you could take 30 from [right], 5 from [middle], and 15 from [left].

In this exchange, Lily felt these two resources tugging her in different directions. However Lily never did reconcile the two, and ultimately selected the *Growing Possibilities* resource when justifying her final answer, rather than finding an answer that balanced both notions.

Polina's Tension Between Resources Polina experienced tensions in reasoning as she pondered the distribution of the sample average from rolling 10 dice. She was comfortable with the sampling distribution for two dice after utilizing the *Modeling Likelihoods* resource and generalizing about the emerging pattern. She drew a bell curve to represent her findings: "It's like there's less combinations to get an average of a 1 and a 6, and then there's more [combinations] as you go towards the middle."

As she pondered the averages from samples of size 10, Polina initially stated that "more of the middle ones" would be represented, to which the interviewer probed further:

Interviewer:	When you say represent more of the middle ones, you mean represent the same way that [n = 2] does, or do you mean represent it even more than it did before?
Polina:	Like 2 to … yeah even more than it did before probably … or no, it would more widely represent I think, the sixes and the ones because it's a higher sample, so it's going to be more representative of the population.

Here, Polina experienced a tension between different sets of resources she viewed as taking her in two different directions. She knew probabilistically with two dice, there were more combinations yielding averages in the middle of the domain. She also knew that larger samples could yield more ones and sixes, reflecting the *Widening Range of Values* resource. She ultimately reconciled these two divergent patterns of thinking:

Polina:	I think the more dice rolls that you have, there is going to be more like sixes rolled and, or averages and 1 averages…but I think in general, like even more so, it will show even more of this [middle].

By attempting to reconcile these different resources, Polina grappled with important conceptual ideas, creating a space for deep statistical reasoning. She realized that for all of the additional combinations that would produce extreme averages, there would be even more combinations added that represented middle averages. Thus, each of these resources could each be applied in this context and aligned with the conclusion that averages would cluster in the middle.

Holland's Successful Blending of Resources Holland articulated several different resources during the Penny task that seemed to fit together like puzzle pieces in her constructions. When she began reasoning about the sampling distribution for two pennies, she used the *Repeated Values* resource to reason that you were not likely to find averages at either end of the range: "I think that the chance of getting two 2017 ones in a row is kind of difficult. I really don't think I've ever had a penny where I'm like, 'oh this one's from 1972,' and 'oh this one's also from 1972.'" She initially drew a centered and fairly narrow bell curve, explaining that averages were more likely to appear in the middle (suggesting she used the *Average Relates to Middle* resource).

When she began her construction of the sampling distribution for selecting 10 pennies, she revised her thinking:

Holland: Oh, I wonder, here let's make [n = 2 distribution] a little more inclusive … my new graph for this one's going to be a little more inclusive and have, well maybe you did find really cool pennies! … and then this one's since it's 10 pennies, I think it's less inclusive.

She redrew her construction for n = 2 to be a much wider, rounded peak, then created a more well-defined peak for n = 10 along with wide, but flattening tails.

Holland: The average, I feel like, must be a lot more central if you're taking [larger samples] versus with [smaller samples]. You do have more flexibility, which is why I made [n = 10] to include more, and like, the widths I guess … but if you're taking 10 at a time, I think that there's, more, umm, more likely to get an average I guess, like it's more specific almost? ... it will be more difficult to get an average that's an outlier if you have a larger group to look at.

Holland balanced three resources that, together, created a powerful tool for reasoning. First, she discussed averages as *Becoming More Accurate* with large samples. She also discussed samples of size 10 as having more flexibility, reflecting that in her graph by letting the tails extend out across the entire range, but keeping them fairly low. This second statement was coded with the *Growing Possibilities* resource. However, Holland still balanced this greater sample flexibility with more defined peaking in the averages, describing these larger sample averages as being more specific. She paired her recognition of the *Growing Possibilities* resource to the resource she articulated earlier, that *Repeated Values* are harder to obtain. This synthesis of different ideas gave her confidence in her belief that averages would cluster more clearly at a central point.

4.3. INFLUENCES FROM PROBLEM SETTING AND CONTEXT

Overview As we look at how and where resources were articulated across the interviews, it became clear that the resources students elicited differed by context and problem setting. This contrast in contextual reasoning spaces prompted interesting discussions and comparisons for students as they sought to reconcile or explain the sometimes diverging conclusions they made in these two settings. In Table 3, we provide a snapshot of how often each resource was elicited across the interviews for both the Penny task and the Dice task.

Table 3. Number of students eliciting each resource in each task

Resources	Penny	Dice
Repeated Values	4	5
Modeling Likelihoods	0	4
Averages Relates to Middle	5	1
Average Relates to Peak	5	0
Sampling Distribution Resembles Population Shape	4	4
Growing Possibilities	2	3
Widening Range of Values	3	3
Stabilizing	2	0
Better Representing Population	2	2
Becoming More Accurate	6	5

One difference in resource elicitation between the two contexts was students' use of Modeling Likelihoods in the Dice task. We attribute this to several context-specific characteristics of the Dice task that facilitated the use of this resource. The first was the limited number of values that students had to consider when reasoning about dice rolls with n = 2. For example, Mika struggled to conceive of likely 2-penny averages in the second prompt, but oriented herself quite quickly to the same prompt in the Dice task: "In this situation, you can have, like, in between numbers right? Because … if they roll 1 and 3, the average would be … two!" By only needing to consider the different combinations possible with six values, Mika and others could manageably name and calculate the probability of each possible average, or reasonably approximate some values as less likely than others.

Furthermore, students expressed certainty about the dice population distribution (uniform). This was in stark contrast to the uncertainty students had with their penny population distributions: almost every student questioned and redrew their penny population after having moved on to the sampling distribution prompts. The Dice context allowed Mika and others to make quick average calculations, whereas the Penny task required a more abstract, conceptual approach.

The students were more likely to make associations between average and middle or average and peak when reasoning about pennies. The fact that students' dice population distributions were all uniform, and had no peak, is an easy explanation for why students did not typically elicit these resources in that context. Furthermore, students' inability to calculate probabilities with pennies left them to depend on characteristics of the population shape to justify their constructions. The remaining resources varied little in elicitation between contexts. We note that the *Stabilizing* resource was used only in the context of pennies—by only two students—but we see no reason why this resource could not have been elicited with dice.

As discussed in Section 3.2, all students completed the Penny task before the Dice task. It is possible, therefore, that students may have approached the Dice Task differently if completed first or independently. The students we interviewed, however, did not of their own accord refer to their work with pennies to justify their work with dice. Students only explicitly related their reasoning between the two contexts when prompted by the interviewer, after having completed the Dice task. Typically, when students began the Dice prompt, they recalled terms like experimental and theoretical probability, or mentioned an experience playing a game, as if changing gears in their head to orient themselves to the new task.

Whereas the overall accounting of resources revealed only limited insight on differences between the task contexts, we did find noteworthy differences when focusing on individual students. In particular, we look at David's reasoning. Although David was one of two students who had previously taken an AP statistics course, he was four years removed from that course and self-reported having little to no recall of the course content. Although we cannot truly gauge the influence of this prior course on his reasoning, we do note that David expressed much cognitive dissonance throughout the interview, suggesting that he was grappling with conceptual ideas rather than applying recalled answers. We chose to highlight David because he articulated interesting tensions in his reasoning and elicited a diverse set of resources in both tasks. Furthermore, we believe his responses can be clearly linked to certain characteristics within each context to demonstrate how context shaped his thinking.

David's Reasoning with Pennies Like others, David believed that recently minted pennies would appear most commonly in circulation, with the frequency decreasing steadily across time. His proposed population distribution is shown in Figure 2.

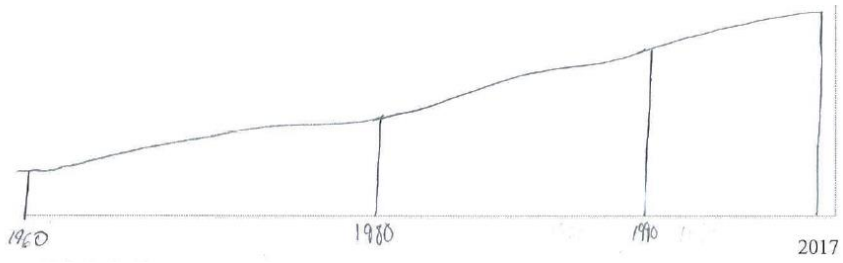


Figure 2. David's population distribution for pennies

After reading the second prompt (averages from two-penny samples), David drew a candidate distribution with little hesitation. His drawing demonstrates a clustering of sample averages closer to the middle-right of the range (Figure 3).

David:	If this initial one is correct that there's more in circulation in current times, you'll get a higher average than you would lower, so this [older end] should even be near 0, getting two of the same lowest year.
Interviewer:	[pointing] 2017 is also low, would that be for the same reason that 1960 is the lowest?

David: I'm taking 2017 as, it's like only that year, so it would be lower because finding a penny of the same year would be not statistically easy … so having two 2017 pennies will be harder to reach than getting say a 2001 and a 1991.

David used the *Sampling Distribution Resembles Population Shape* resource when he attributed the concentration of pennies on the right to influencing the placement of penny averages similarly. David constructed a graph that was not identical to the population (showing some recognition of the change in variabiltiy), but certainly responsive to its shape. He also implemented the resource *Repeated Values* by stating that getting two 2017 pennies was the only way to get an average of 2017, making that average hard to obtain. He likewise considered old pennies to be similarly unlikely.

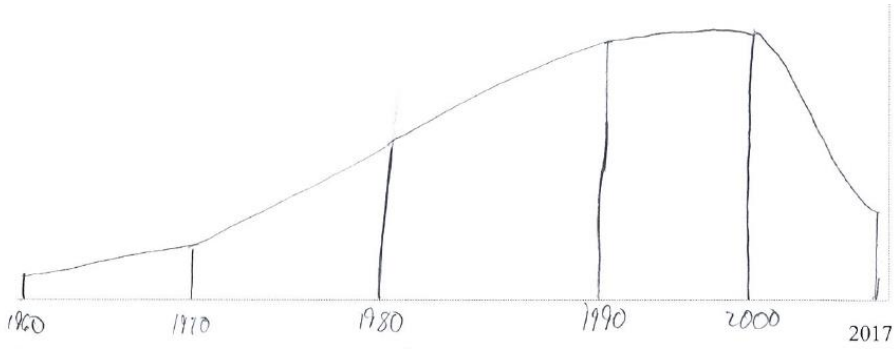


Figure 3. David's sampling distribution for n = 2

When David moved to reasoning about averages from samples of size 10, his reasoning shifted. He drew a sampling distribution that peaked more in the middle of the range and elicited a new set of resources to justify his construction.

David: So in [the *n* = 2 construction], there are more in circulation in these higher [years], so the chances of finding these is higher in a 2 survey, or a 2 average, so it's skewed higher. But since you're doing a 10 average here, there's more chances of getting '80, '70, and '60 coins.

Here, we see the *Widening Range of Values* . He then tied in an additional resource, explaining: "You just have more in the sample, so you get a better average rather than two pennies." In this response, David elicited the *Becoming More Accurate* resource to relate the increasing sample size to improved confidence in the sample averages produced. However, his conception of accuracy in this case was tied to finding averages better representing all parts of the range.

When asked about the shape of the sampling distribution for *n* = 50 or more, David described this curve as being more level, more gradual, and less steep, though he did not see this curve ever reaching a flat line (Figure 4). The language used here suggested he viewed *Stabilizing* as a resource for understanding the behavior of large sample averages. Furthermore, David explained that the larger samples would produce averages that peak closer to the center of the range as they would contain a wider selection of pennies, whereas the smaller sample averages would still peak slightly right of middle to reflect the population shape. Here, David drew from the *Becoming More Accurate* resource to justify

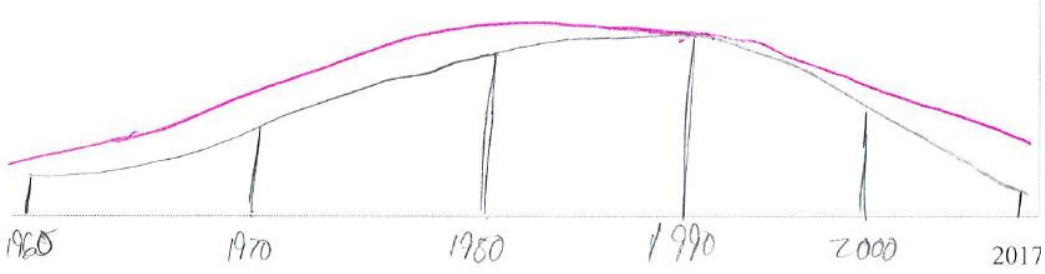


Figure 4. David's sampling distributions for n = 10 (black) and n = 50 (pink)

this central position while pulling from the *Widening Range of Values* and *Stabilizing* resources to justify the extended, smooth shape.

In summary, David began with the resource *Sampling Distribution Resembles Population Shape*, but with more clustering away from the ends due to his resource that *Repeated Values* are less likely. For the sampling distributions with larger sample sizes, David focused on how these samples would include a *Widening Range of Value*s while also *Stabilizing* in the form of a flatter distribution. He viewed these averages *Becoming More Accurate* in representing the range by beginning to peak in the center rather than to the right.

David's Reasoning with Dice David knew that when rolling one die, all six possible values were equally likely to appear. As he moved on to reasoning about averages from two dice rolls,the probabilities of each of the possible values became less clear.

David began responding to the $n = 2$ prompt with the *Sampling Distribution Resembles Population Shape* resource: "Wouldn't it just be the same because it's a 1 in 6 chance of hitting … because you're averaging them, not adding them together, so it's not a different probability of hitting higher numbers or lower numbers." He struggled to move on to the next prompt because "his gut" told him it should be bell-shaped. By imagining playing a game that involves rolling two dice, David expressed that rolling doubles was always less likely than rolling non-doubles (*Repeated Values*), leading him to wonder whether averages of 1 and 6 must be less likely. His drawing representing these thoughts is found in Figure 5.

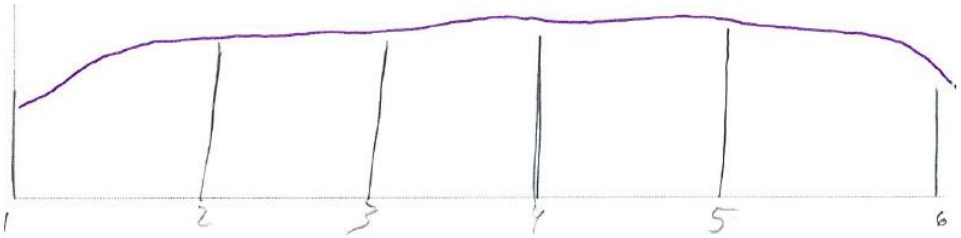


Figure 5. David's sampling distribution for n = 2

He began responding to the $n = 10$ prompt by stating, "I'm thinking [$n = 2$] is wrong, again … hitting, 1 ten times seems a lot lower than hitting 10 random numbers, which makes me think this [$n = 2$] should be slightly different. Similar to a bell curve distribution."

His next statement tied together the *Repeated Values* resource with the resources of *Better Representing Population*, *Widening Range of Values*, and *Becoming More Accurate*.

> David: Taking the average of all of them, it will be more likely to hit a bunch of different, each one of them equally, which should average out to between 3 and 4, so 3.5, then it would be to hit 6 or 1 or really any of them, 10 times in a row.

Here, David recognized that a larger sample would contain a more diverse collection of values ("hit a bunch of different [numbers]"), aided by the fact that all values have an equal chance of being rolled and represented in larger samples. From here, he was able to turn his attention to the likely position of the averages. Recognizing that rolling the same number again and again would be less likely helped pave the way to seeing large-sample averages clustering in the middle and not near the ends. When asked what he believed would happen as we increased the sample size, David said that the middle would continue to peak higher and the ends would continue to drop lower, leading to a divergence in conclusions between his responses in the Penny task and those in the Dice task.

Table 4 provides a comparison summary of David's elicited resources in both tasks—*italics* indicate differences in resources elicited by the tasks.

Table 4. A comparison of resources for David's reasoning with pennies and dice

Prompt	Penny Resources	Dice Resources
$n = 2$	- Repeated Values - Sampling Distribution Resembles Population Shape	- Repeated Values - Sampling Distribution Resembles Population Shape
$n = 10$	- *Sampling Distribution Resembles Population Shape* - Widening Range of Values - Becoming More Accurate	- *Repeated Values* - *Better Representing Population* - Widening Range of Values - Becoming More Accurate
$n = 50+$	- Widening Range of Values - Becoming More Accurate - *Stabilizing*	- Widening Range of Values - Becoming More Accurate - *Better Representing Population* - *Repeated Values*

Summarizing David's Reasoning The summary table reveals that David elicited a similar—but not identical—progression of resources in both reasoning contexts, yet ultimately came to different conclusions. David instinctively related the sampling distribution shape with that of the population as a first step in both tasks. He also built on the knowledge that getting the same data value repeatedly should be less likely. When reasoning about larger samples, he recognized that larger samples would include a wider, more "representative" collection of values, but found this resource to bring different implications in each task. With the Penny task, he believed this wider sample range should produce averages that would stretch to include all values, with accuracy coming into play to explain the position of the peak. In the Dice task, however, David believed large samples containing uniformly representative values from 1 to 6 would produce more accurate averages that visibly cluster in the middle of the range.

When given the opportunity to revisit the prompts in the Penny task after completing the Dice task, David said "Dice are about probabilities, and we know the probability of a 6-sided die, but here, at least I'm not familiar with the probability of picking up a penny and knowing what the year is going to be." It seemed the added assurance of knowing the uniform probabilities of dice rolls provided fewer unknowns to wrestle with. Without such assurance in reasoning about pennies, David struggled to look beyond the content of the samples he might get and conceptualize what averages would be produced from these wider samples.

When picturing what might happen as the sample size continued to get larger and larger, David continued to see the shape of the population as significant in the context of dice, but not pennies. The skewness and uncertainty of the penny population pushed him to conceive of a stabilizing and smoothing result. David did not discuss stabilization with dice, seemingly because no stabilization and smoothing was necessary. Dice averages were clustering in the center from even the smallest samples in his drawings, and the fact that the dice population was already perfectly smooth left no room for the sampling distributions to become smoother. We speculate that the simplicity of the dice context aided him. His surety in the uniform population distribution allowed him to use this resource confidently in his constructions. This contrasted with his uncertainty in the penny population distribution.

5. DISCUSSION

5.1. CONTRIBUTIONS AND LIMITATIONS

In this paper, we have used a *Knowledge in Pieces* approach (diSessa, 1988) to identify resources that students elicit when reasoning about sampling distributions. As we compare our findings with insights in the literature, we see that many of our resources resonate with common student misconceptions. For example, Chance et al. (2004) noted that students believe the sampling distribution should look more like the population as sample size increases, which corresponds to one of our identified resources. Our research extends insights from the literature by uncovering the more

fundamental observations behind these common misconceptions and investigating how these observations can be mended productively.

We take an approach similar to that of Kahneman and Tversky (1972) in looking for heuristics that can be used as building blocks for conceptual models. We believe the resources we identified complement more recent work that has identified different perspectives and modeling approaches students have taken with respect to reasoning about sampling distributions (e.g., Braham & Ben-Zvi, 2017; Pfannkuch, Arnold, & Wild, 2015; Prodromou & Pratt, 2006). Whereas these studies discussed larger frames of thinking that students expressed across instructional sequences, this paper presents more fine-grained resources we believe undergird the perspectives students take. For example, Prodromou and Pratt (2006) found many students reason about sampling distributions through either a data-centric perspective or a modeling perspective, rather than a thoughtful blending of the two. Research and instruction may understand students' perspectives with more depth by probing the specific resources students activate when taking each of these perspectives and how the illumination of alternative resources may bring about a more balanced perspective. For example, in the case of a student holding a data-centric perspective, the instructor might offer the notion of stabilization with large sample averages to elicit a new perspective for the student to consider. Students who focus on the shape of the population to determine the shape of the sampling distribution may further their thinking by considering how accuracy may be linked to larger sample sizes. Further research on instruction through the lens of resource blending may inspire new insights on learning trajectories.

Our findings support existing research by noting that conceptual progress depended on explicit reconciliation of multiple perspectives (Pfannkuch et al., 2015; Prodromou & Pratt, 2006). It was through thoughtful blending of resources that students were able to see how all of these pieces could fit together in a unified model. Interview excerpts from Holland and Polina provided detailed looks at how resources could be tied together to create new, more complex conceptions of the situation. For example, Holland's recognition of the widening range of values in larger samples paired with recognition that repeated extreme values were unlikely solidified her belief that the averages these samples produced would become more specific and cluster in the center. In contrast, Lily struggled with the tasks, hitting a roadblock in reconciling two resources that she believed could not both be true and cohesively tied together in a construction.

Additionally, our findings enhance current understandings of contextual influences on students' statistical reasoning (Pfannkuch, 2011). We noted characteristics about each problem that led students to notice different things and ultimately take dissimilar reasoning paths. David's work demonstrated contextual divergence as he used similar but distinct sets of resources to come to different conclusions about the sampling distribution shape for pennies and dice. David was more comfortable in the Dice task due to his confidence in the population shape and his discovery that averages from two dice were not probabilistically equal. These scaffolds in the dice context freed him to focus on the averages these samples would produce. The ambiguity of the penny population left him uncertain in his responses as he tried to juggle multiple considerations.

The scope of our findings, however, is somewhat limited by the design of our study. With a more diverse set of tasks, it is possible that students would have utilized a more varied set of resources. Additionally, our sample of only eight undergraduates from a large, public university is not representative of undergraduate students as a whole. As such, the resources used by these students are likely not comprehensive and may differ in prevalence from the resources used by students in different schools, majors, or academic levels.

Finally, each student received the penny task first and the dice task second. Our particular study was not designed to compare differences in ordering of tasks, which then limits our understanding of how task ordering influences elicitation and development of resources.

5.2. CONCLUSION

In her book *The Having of Wonderful Ideas*, Duckworth (2006) stressed the importance of being sensitive to student thinking by giving students the resources, space, and trust to engage with their own questions. By searching for the resources students activate, rather than simply the misconceptions they articulate, we believe instructors will be better attuned to support students' conceptual progress. Effective instruction should aptly respond to students' ideas, creating opportunity for students to

conduct thought experiments that may include incorrect notions along the way (Garfield & Ben-Zvi, 2009). Learning environments that appropriately elicit and support students' ideas toward productive ends—as opposed to writing over students' incomplete notions and questions—will be rich with possibilities for growth (Cobb, 1994).

Although our study did not specifically investigate the role of computer simulations in students' conceptual understanding, we view the resources we identified as worthwhile considerations across an instructional sequence that includes simulations. As Garfield et al. (2015) discussed from the extensive research on this topic, informal reasoning prior to the use of simulations often prompts students to conceptual ideas and limits the possibility of tacit acceptance of simulated results. Future work may investigate how the identification and prompting of an array of cognitive resources can support more productive student interaction and discussion of simulated results.

Our findings also point to the importance of choosing context carefully in the creation of instructional tasks. In particular, the familiar, manageable setting of the Dice task provided a fairly safe conceptual space for students to grapple with multiple resources, whereas the ambiguous Penny task required more abstract thinking. We wonder how a task in between the two (e.g., a fixed sample space with a non-uniform distribution) may elicit a different mixture of resources. Future research is needed to further understand how the settings of multiple contexts, or the sequencing of multiple tasks, can influence students' elicitation and blending of resources toward generalizable, conceptual models of sampling variation.

ACKNOWLEDGEMENTS

We would like to thank Lama Jaber for providing early inspiration and feedback on our work. We would also like to thank our reviewers for their substantive suggestions. Finally, we thank the students who participated in our study.

REFERENCES

Baglin, J., Reece, J., & Baker, J. (2015). Virtualising the quantitative research methods course: An island-based approach. *Statistics Education Research Journal*, *14*(2), 28–52. [Online: http://iase-web.org/documents/SERJ/SERJ14(2)_Baglin.pdf]

Braham, H., & Ben-Zvi, D. (2017). Students' emergent articulations of models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, *16*(2), 116–143. [Online: https://iase-web.org/documents/SERJ/SERJ16(2)_ManorBraham.pdf]

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher, 23*(7), 13–20.

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2). [Online: http://jse.amstat.org/v16n2/cooper.pdf]

Corbin, J., & Strauss, A. (2008). *The basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Erlbaum.

Duckworth, E. (2006). *The having of wonderful ideas and other essays on teaching and learning*. New York: Teachers College Press.

Eaton, J. F., Anderson, C. W., & Smith, E. L. (1984). Students' misconceptions interfere with science learning: Case studies of fifth-grade students. *The Elementary School Journal*, *84*(4), 365–379.

Fauconnier, G., & Turner, M. (2003). Conceptual blending, form and meaning. *Recherches en Communication*, *19*(19), 57–86.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics*, *31*(3), 72–77.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM–International Journal on Mathematics Education*, *44*(7), 883–898.

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*(3), 327–342.

Hammer, D., & van Zee, E. (2006). The beginnings of scientific reasoning. In D. Hammer & E. van Zee (Eds.), *Seeing the science in children's thinking: Case studies of student inquiry in physical science* (pp. 13–37). Portsmouth, NH: Heinemann.

Hjalmarson, M. A., Moore, T. J., & delMas, R. (2011). Statistical analysis when the data is an image: Eliciting student thinking about sampling and variability. *Statistics Education Research Journal*, *10*(1), 15–34.
[Online: http://iase-web.org/documents/SERJ/SERJ10(1)_Hjalmarson.pdf]

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education*, 392–414.

Lane-Getaz, S. J. (2017). Is the *p*-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. *Statistics Education Research Journal*, *17*(1), 357–399.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_LaneGetaz.pdf]

Nicholson, J., & Ridgway, J. (2017). A response to White and Gorard: Against inferential statistics: How and why current statistics teaching gets it wrong. *Statistics Education Research Journal*, *16*(1), 66–73.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_Nicholson.pdf]

Noll, J., Shaughnessy, M., & Ciancetta, M. (2010). Students' statistical reasoning about distribution across grade levels: A look from middle school through graduate school. In C. Reading (Ed.) *Proceedings of the Eighth International Conference on the Teaching of Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots8/ICOTS8_8B4_NOLL.pdf]

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1-2), 27–46.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics*, *88*(3), 343–360.

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education*, *31*(5), 602–625.

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107–129.
[Online: https://iase-web.org/documents/SERJ/SERJ7(2)_Pratt.pdf]

Prodromou, T., & Dunn, T. (2017). Statistical literacy in data revolution era: Building blocks and instructional dilemmas. *Statistics Education Research Journal*, *16*(1), 38–43.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_Prodromou.pdf]

Prodromou, T., & Pratt, D. (2006). The role of causality in the coordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69–88.
[Online: https://iase-web.org/documents/SERJ/SERJ5(2)_Prod_Pratt.pdf]

Sewell, A. (2002). Constructivism and student misconceptions: Why every teacher needs to know about them. *Australian Science Teachers Journal*, *48*(4), 24.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163.

Shaughnessy, M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on the teaching and learning of mathematics* (Vol. 2, pp. 957–1009). Charlotte, NC: Information Age Publishers.

Sotos, A., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113.

Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students appreciation of variation and expectation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, *9*(2), 83–130.

Watson, J. M., & Shaughnessy, J. M. (2004). Proportional reasoning: Lessons from research in data and chance. *Mathematics Teaching in the Middle School*, *10*(2), 104–109.

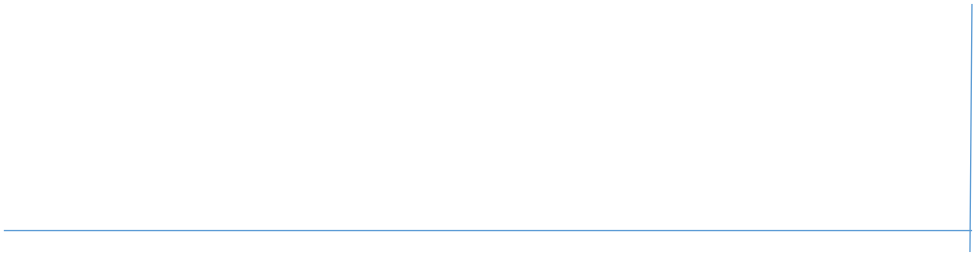
White, P., & Gorard, S. (2017). Against inferential statistics: How and why current statistics teaching gets it wrong. *Statistics Education Research Journal, 16*(1), 55–65.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_White.pdf]

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education, 16*(2).
[Online: http://jse.amstat.org/v16n2/zieffler.pdf]

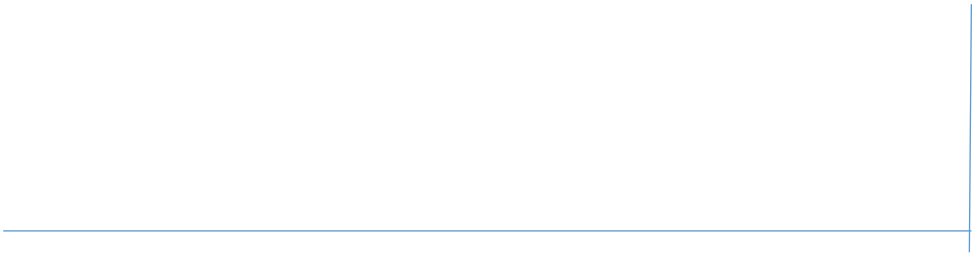
KELLY FINDLEY
Stone Building, 1114 W. Call St.
Tallahassee, FL 32306 USA

APPENDIX

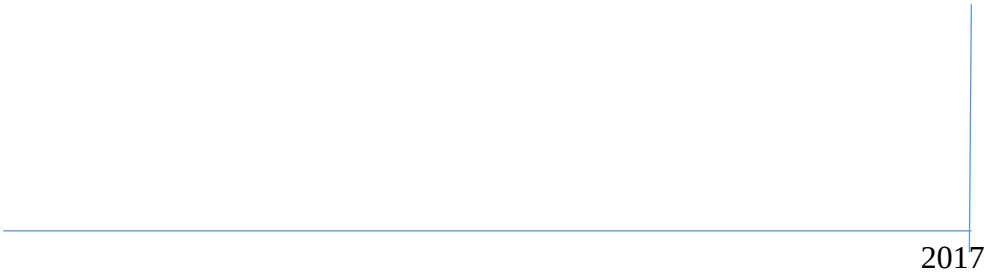
Think about the age of pennies in circulation (pennies in cash registers or people's money wallets and purses). What is the range of penny years that we would see in circulation? Label in a few more years and draw a line to represent how many pennies you would expect there to be across the range of penny years.



Now think about if we were to take 2 pennies randomly from circulation and find their average age. If we repeatedly did this and collected a list of 2-penny averages, then what would be the range of averages we would see? Fill in more years and again draw a line to represent how many of each average would we see.

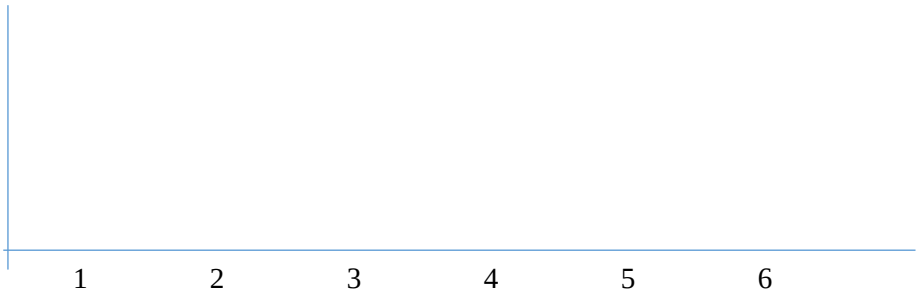


Do the same task we did above, but this time think about taking 10 pennies at a time instead of 2.

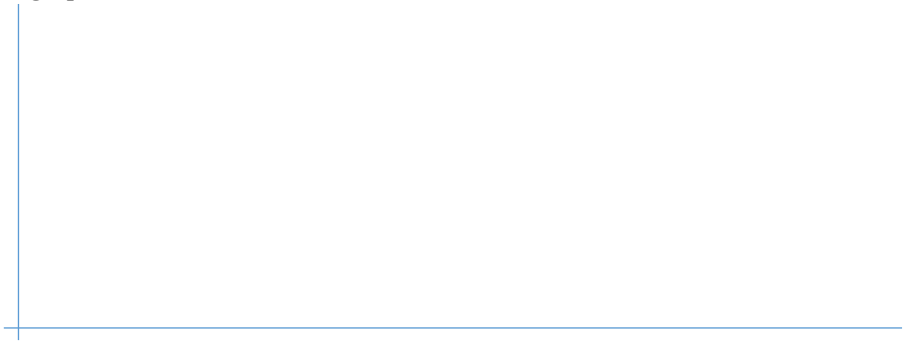


Think about rolling a regular, six-sided die. If we had one million people each roll a six-sided die, how many of each outcome would you expect to see?

Draw a column graph below to represent how many of each outcome you would expect.



Now imagine each of these one million people each rolled 2 dice and took the average value. Think about the averages we would see. How many of each 2-dice average do you expect to see? Create a column graph.



Now imagine each of these one million people each rolled 10 dice and took the average value. Draw a line plot (density curve) to represent the averages we would expect to see most often.