USING GUIDED REINVENTION TO DEVELOP TEACHERS' UNDERSTANDING OF HYPOTHESIS TESTING CONCEPTS

JASON DOLOR
Portland State University
jdolor@pdx.edu

JENNIFER NOLL
Portland State University
noll@pdx.edu

ABSTRACT

Statistics education reform efforts emphasize the importance of informal inference in the learning of statistics. Research suggests statistics teachers experience similar difficulties understanding statistical inference concepts as students and how teacher knowledge can impact student learning. This study investigates how teachers reinvented an informal hypothesis test for categorical data through the framework of guided reinvention. We describe how notions of variability help bridge the development from informal to formal understandings of empirical sampling distributions and procedures for constructing statistics and critical values for conducting hypothesis tests. A product of this paper is a hypothetical learning trajectory that statistics educators could utilize as both a framework for research and as an instructional tool to improve the teaching of hypothesis testing.

Keywords: *Statistics education research; Statistical inference; Sampling distribution; Test statistic; Hypothetical learning trajectory*

1. INTRODUCTION

The American Statistical Association's (2012) *Guidelines for Assessment and Instruction in Statistics Education (GAISE) report* outlines a progressive approach toward the teaching and learning of introductory statistics. According to the GAISE report, the goal of an introductory statistics course is "for students to focus more on conceptual understanding and attainment of statistical literacy and thinking, and less on learning a set of tools and procedures" (p. 10). GAISE advocates a focus on teachers as facilitators in a pedagogical process that encourages student construction of fundamental statistical concepts and methods. One of these fundamental concepts is that of statistical inference.

Statistical inference is the science of using sample information to draw conclusions regarding the population (Brase & Brase, 2012). GAISE suggests students should understand the basic ideas of statistical inference, which include: "the concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the

Statistics Education Research Journal, 14(1), 60-89, http://iase-web.org/Publications.php?p=SERJ

idea of standard error)" and "the concept of statistical significance, including significance levels and *p*-values" (ASA, 2012, p. 12).

Hypothesis testing is an essential component of statistical inference. Hypothesis testing methods are employed in all fields (e.g., education, sciences, medicine, social sciences, business, etc.) that utilize data (Bluman, 2012). Concepts of hypothesis testing, therefore, should be a priority for statistics education researchers. In addition, teacher knowledge has a direct impact on student learning (Ball, Hill, & Bass, 2005; Hill, Sleep, Lewis, & Ball, 2007; Liu & Thompson, 2009; Shulman, 1986); thus, statistics education researchers must also understand the development of teachers' knowledge of hypothesis testing. One particularly important area of research is analyzing how to bridge the gap between intuition and prior knowledge with formal statistical concepts. Therefore, the goal of this paper is to address this gap by investigating these research questions:

1. How do in-service and pre-service teachers (IPSTs) move from informal intuitions toward more formal concepts of hypothesis testing for categorical data using a guided reinvention approach?
2. What role does IPSTs' intuition about variability play in constructing empirical sampling distributions and developing procedures for finding a statistic and critical value used in a hypothesis test for categorical data?

This paper is organized into six main sections. In the first section, related research literature on students' and teachers' understanding of hypothesis testing is reviewed. Second, instructional strategies designed to improve the teaching of hypothesis testing are synthesized. Third, the framework of guided reinvention and how it led to a preliminary hypothetical learning trajectory (HLT) for supporting teachers' reinvention of a hypothesis test for categorical data is explicated. Fourth, the teaching experiment, focused on hypothesis testing in a statistics classroom for IPST, is explained. Note that for the remainder of this paper IPSTs are referred to as students for ease of reading. Fifth, data from the teaching experiment with respect to how these data addresses the research questions are shared. Lastly, this paper concludes with a discussion of results and suggestions for future research.

2. LITERATURE REVIEW

2.1. STUDENT AND TEACHER UNDERSTANDING OF HYPOTHESIS TESTS

A review of the literature indicates that students struggle to understand concepts of hypothesis testing (Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos, Vanhoof, Noortgate, & Onghena, 2007; Falk, 1986; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Thompson, Liu, & Saldanha, 2007; Vallecillos, 2002; Vallecillos & Batanero, 1997). Some research has focused on students' misinterpretation of *p*-value and level of significance (Batanero, 2000; Batanero & Diaz, 2006; Garfield & Ben-Zvi, 2008). Batanero (2000) observed that students could calculate a correct *p*-value and correctly conclude whether to reject or fail to reject the null hypothesis based on their calculated *p*-value. Yet, the students in her study could not adequately interpret a *p*-value. Vallecillos and Batanero (1997) revealed that students have difficulties identifying the null and alternative hypotheses when confronted with contextual problems where the hypotheses are not clearly delineated.

Statistics education researchers have also found that teachers struggle to understand hypothesis testing (Heid, Perkinson, Peters, & Fratto, 2005; Thompson, Liu, & Saldanha, 2007). Thompson and his colleagues investigated eight high school teachers in a professional development seminar. They noted that some of the teachers appeared to understand the logic of hypothesis testing, but could not appropriately identify problems that could be answered through hypothesis testing. Furthermore, some teachers in the study did not show a strong understanding of the role sampling distributions play in hypothesis testing. This phenomenon was also observed in a study conducted by Heid and her colleagues (2005), who investigated eighteen prospective secondary teachers enrolled in a course designed to broaden understanding of statistical concepts. In their research, they discovered that while some prospective teachers could accurately describe and construct a sampling distribution, the teachers were inconsistent when trying to articulate distinctions between distributions of samples, populations and sample means. Heid and her colleagues suggested that the teachers' inability to make these important distinctions led them to reason deterministically rather than probabilistically when applying sampling distributions to real data. This is an important point because concepts of sampling distributions are important prerequisite knowledge for a deeper understanding of hypothesis testing.

The research literature reviewed here reveals challenges students and teachers have understanding *p*-values, level of significance and sampling distributions. There is little research studying how students or teachers might come to derive a test statistic from observed sample data and construct a distribution of statistics generated under a null hypothesis. Nor is there much research on how students or teachers might conceive of the relationship between these ideas and inferences made through hypothesis testing. This research aims to make headway into these important ideas. To begin, current pedagogical approaches suggested by statistics education researchers are reviewed.

2.2. IMPACT OF RESEARCH ON PEDAGOGICAL APPROACHES

Traditional pedagogical approaches to the teaching of hypothesis testing have focused on formal approaches such as *t*-tests and *z*-tests. However, in the past decade, statistics education researchers have developed new approaches to teach statistical inference that are designed to improve student understanding. Three interrelated strategies have received the bulk of the attention: emphasizing (1) sampling distributions, (2) simulations, and (3) informal inference.

The first strategy suggested by researchers is emphasizing the importance of sampling distributions in the learning of statistical inference (Batanero, 2000; Garfield & Ben-Zvi, 2008; Heid et al., 2005; Lipson, 2003; Liu & Thompson, 2005; Makar & Confrey, 2004; Saldanha & Thompson, 2002). These researchers have argued that because sampling distributions form the basis for understanding the relationship between samples and probability in statistical inference, educators should emphasize sampling distributions as a core component of hypothesis testing during instruction.

The second strategy suggested by researchers (e.g., Chance, delMas, & Garfield, 2004; Chance, Ben-Zvi, Garfield, & Medina, 2007; delMas, Garfield, & Chance, 1999; Erickson, 2006; Garfield & Everson, 2009; Garfield & Ben-Zvi, 2008; Heid et al., 2005; Zieffler, Garfield, delMas, & Reading, 2008; Zieffler, Garfield, Alt, Dupuis, Holleque, & Chang, 2008) is having simulations serve as an important tool in teaching hypothesis testing because they provide a visual (via a computer) and/or physical experience (via hands-on simulations using dice, coins, spinners, etc.) of the sampling process as a means to generate an empirical sampling distribution. Chance, delMas, and Garfield (2004) have asserted that simulations can build foundational ideas of sampling distributions by allowing students to explore intuitively how samples and sampling distributions behave with regard to a population.

The third strategy suggested by researchers is that instruction should begin with an informal approach to statistical inference before introducing formal statistical inference (e.g., Chance & Rossman, 2006; Erickson, 2006; Garfield & Ben-Zvi, 2008; Garfield, delMas, & Zieffler, 2012; Rubin & Hammerman, 2006; Weinberg, Wiesner, & Pfaff, 2010; Zieffler, Garfield, delMas, & Reading, 2008). By building on informal approaches, educators can better support students' learning of statistical inference concepts because new knowledge is constructed from the learner's prior knowledge, and thus making formal procedures more accessible.

The three main strategies outlined in the research literature are interrelated and mutually supportive. Our synthesis of the literature (in particular, Erickson, 2006; Garfield et al., 2012; Chance & Rossman, 2006; Liu & Thompson, 2005; Makar & Confrey, 2004; Weinberg, Wiesner, & Pfaff, 2010; Zieffler et al., 2008) suggests these three strategies support a particular pedagogical approach where students begin their study of statistics by experiencing the sampling process (physically or electronically) to generate empirical sampling distributions. Students then discuss unusual samples using an empirical sampling distribution as a tool for a statistical argument. Simulations are encouraged because they allow inferences to be made without probability theory and assumptions about types of probability distributions. Terms such as *unusual*, *rare*, *typical*, and *representative* replace the formal language of *p*-value, test statistic, and significance level at the beginning of instruction.

Utilizing this new pedagogical approach makes statistical discussions accessible to students because words like *unusual* have everyday meaning. Activities are designed to build on students' prior knowledge of these words through statistical discussions where students construct methods of analyzing the unusualness of sample data. Instruction focuses on building connections between empirical sampling distributions, the unusualness of a statistic obtained from the observed data and building inferences from this information. Unusualness is reflected by the probability of an observed statistic (i.e., the likelihood of occurrence of an observed statistic ranging from expected to unexpected), which is dependent on the position of the statistic within an empirical sampling distribution based on the null hypothesis. Instruction, with a focus on informal ideas, may successfully support students in thinking about the unusualness of an observed statistic on a sliding scale, ranging from representative (highly probable) to unusual (not very probable). Students can use the graph of an empirical sampling distribution to see that the tails of the distribution represent areas of unusual statistics while the middle region corresponds to representative statistics. Figure 1 illustrates a simulated empirical sampling distribution of sample proportions from a population of equally proportioned males and females and highlights the unusualness (or representativeness) of an observed sample proportion of 0.70 males by its position in the empirical sampling distribution.

Based on this particular pedagogical strategy, we outline an informal hypothesis testing (IHT) approach as a 7-step process: (1) Begin with a research question and state the relevant hypotheses; (2) Collect relevant data; (3) Generate/Design a procedure for determining a statistic for the sample based on the research question; (4) Generate an empirical sampling distribution (based on the null hypothesis) through simulation; (5) Determine the unusualness of the observed statistic in relation to the empirical sampling distribution based on the null hypothesis; (6) Conclude the results of the hypothesis test; and (7) Interpret the results of the test. It is important to acknowledge that students' prior knowledge must include an understanding of simulations and graphs because these concepts support the creation and discussion of empirical sampling distributions in IHT. Table 1 presents an example of the IHT where the research question of interest is whether College A has more male students than female students.

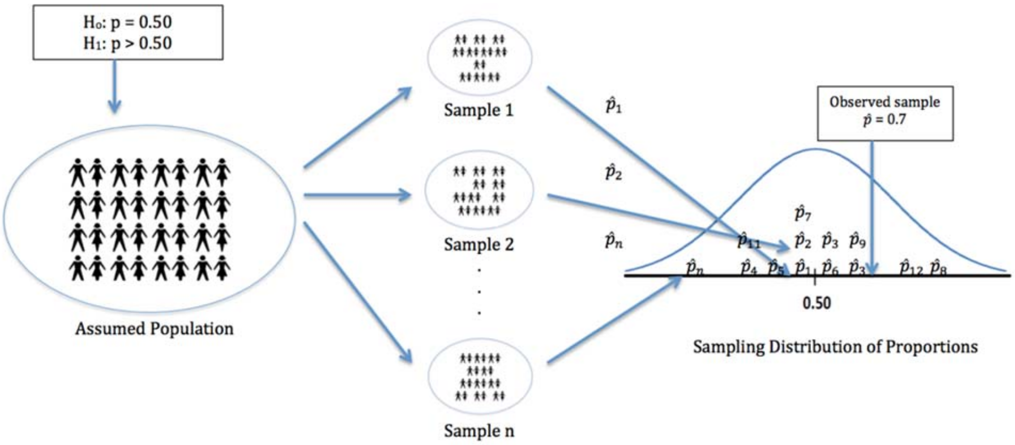
H_o: p = 0.50
H_1: p > 0.50
Assumed Population
Sample 1
Sample 2
Sample n
$\hat{p}_1$
$\hat{p}_2$
$\hat{p}_n$
Observed sample
$\hat{p} = 0.7$
$\hat{p}_7$
$\hat{p}_{11}$ $\hat{p}_2$ $\hat{p}_3$ $\hat{p}_9$
$\hat{p}_n$ $\hat{p}_4$ $\hat{p}_5$ $\hat{p}_1$ $\hat{p}_6$ $\hat{p}_3$ $\hat{p}_{12}$ $\hat{p}_8$
0.50
Sampling Distribution of Proportions

Figure 1. Relationship between population, sample, and sampling distribution.

We conjecture that through carefully designed activities, using simulations, IHT and incorporating the framework of guided reinvention, students' prior knowledge and intuition can be capitalized on so that students can reinvent fundamental concepts of hypothesis testing.

2.3. REALISTIC MATHEMATICS EDUCATION AND GUIDED REINVENTION

According to Gravemeijer (2004), the problem with the transmission model of teaching (traditional lecture) is that mathematics in general does not make sense to students because what is common sense for a mathematician is not common sense for a student. A core principle of realistic mathematics education (RME) is that mathematics should be learned naturally through invention and discussion as students are involved in solving mathematical problems that are realistic to them. He argues that in order to foster rich student learning opportunities, educators must take the approach of helping students (re)invent significant mathematics through a process of *guided reinvention*. Guided reinvention is a component of RME that emphasizes the development of tasks and pedagogical practices that allow opportunities for students' reinvention of mathematical concepts. This is done by having teachers establish a culture of inquiry within the mathematics classroom where the teacher's role is to present a relevant task, ask questions and lead discussion to facilitate students' ability to construct meaningful mathematics. The classroom is student-focused. Student work and student responses become the main topics of discussions.

Table 1. Steps of informal hypothesis testing

Hypothesis Test Steps	Explanations / Example
Step 1. Begin with research questions and state relevant hypotheses.	Students begin with a research question and state hypothetical claims based on their research question. *Example*: Research Question: Are there more male students than female students at College A? Null Hypothesis: The proportion of males is 0.50. Alternative Hypothesis: The proportion of males is greater than 0.50.
Step 2. Collect relevant data.	Students collect relevant data that would provide evidence for their claims. *Example*: To answer the research question, students either collect data or are given sample data about the gender of students at College A.
Step 3. Construct a procedure to find a statistic; use it to measure unusualness of sample data.	Students construct a numerical method to represent a sample's unusualness in light of the null hypothesis. *Example*: statistic = sample proportion.
Step 4. Generate an empirical sampling distribution based on null hypothesis.	Students use a simulator to generate an empirical sampling distribution of statistics drawn from a probability distribution that represents the null hypothesis. *Example*: Using a simulator, generate multiple samples of size 100 from a population distribution representing 0.50 males. Students graph the sampling distribution of sample proportions.
Step 5. Determine the likelihood of an observed sample.	Determine the position (and thus likelihood) of the observed sample's statistic relative to the sampling distribution generated in step 4. *Example*: If a sample with a proportion of 0.70 males is observed, then determine its likelihood using the sampling distribution generated in step 4. If the observed sample proportion is unusual, the sampling distribution would show a distribution that tails off around 0.70.
Step 6. Conclude the results of the hypothesis test.	Based on the results of step 5, the students determine whether there is sufficient evidence to reject the null hypothesis. If the observed statistic is highly unusual then there is evidence to reject the null hypothesis. If the observed statistic is not unusual then we fail to reject the null hypothesis. *Example*: In the previous step, we have shown that getting a sample proportion of 0.70 males is unusual. Thus, there is sufficient evidence to reject the claim that the proportion of males in the population is 0.50.
Step 7. Interpret the results.	Interpret the decision based on the results of step 6. *Example*: Since the sample proportion of 0.70 males has a small chance of occurring assuming the population proportion of males is 0.50, then we can claim that there are more males in the college population.

The RME design heuristics rely heavily on planning. Preparing an instructional sequence involves developing a hypothetical learning trajectory (HLT) for the class as students participate in activities. "The notion of a hypothetical learning trajectory entails that the teacher has to envision how the thinking and learning, in which the students might engage as they participate in certain instructional activities, relate to the chosen learning goal" (Gravemeijer, 2004, p. 8). The components of a HLT consist of establishing learning goals, envisioning students' mental processes and designing instructional tools (Gravemeijer, 2004). Instructional sequences can be divided into stages in order to achieve sub-goals, with the intention of satisfying overall learning goals. The teacher must take the perspective of the student, envisioning the mental activities of students as they participate in instructional tasks so that the teacher can plan for possible student responses and developing conceptions. For instance, teachers must consider what instructional tools, tasks and questions would be most helpful in developing student thinking. Since curriculum design is not always perfect, contingencies must also be planned in case students stray from the conjectured HLT.

Mathematics educators, and to a lesser extent, statistics educators (e.g., Bakker, 2004; Cobb, McClain, & Gravemeijer, 2003; Doorman & Gravemeijer, 2009; Gravemeijer & Bakker, 2006; Gravemeijer & Van Galen, 2003; Larsen, Johnson, Rutherford, & Bartlo, 2009; Oehrtman, Swinyard, Martin, Hart-Weber, & Hah Roh, 2011; Swinyard, 2011; Van Etten & Adendorff, 2007) have used RME as a lens for designing teaching experiments to study the learning processes of students. In the framework of RME, realistic mathematical problems are contextually and intellectually accessible from the perspective of students based on their prior knowledge and intuition. Using RME and IHT generates avenues for educators to support the learning of hypothesis testing by building on the learner's intuition and prior knowledge. A primary conjecture of this study is that the instructional strategies outlined in the literature review could be enhanced with a RME philosophy as a means to investigate students' development of hypothesis testing. This approach may yield new insights into the way students construct knowledge of hypothesis testing as well as provide empirical support for particular curricular approaches in the teaching and learning of hypothesis testing.

3. METHODS

Using the framework of RME and an IHT approach to the study of hypothesis testing, we conducted a teaching experiment at a large urban university in the Pacific Northwest. This section outlines our methodology beginning with a description of the participants, followed by a description of our HLT, including the tasks used in the teaching experiment and how RME served as an important theoretical lens from which we analyzed our results.

3.1. PARTICIPANTS

The study was conducted in a statistics course for teachers and had a prerequisite of an introductory statistics course. The class consisted of 13 graduate and upper division undergraduate mathematics students with a mix of pre-service and in-service teachers. Background information on the participants is provided in Table 2. All graduate students had bachelor's degrees in mathematics. Twelve of the students allowed the research team to videotape and analyze their written work. Only one of the students requested not to be videotaped but allowed his written work to be used as data. The research presented in this paper is based on classroom discussion related to hypothesis testing activities. Therefore, the student that did not wish to be videotaped is excluded from this discussion.

Table 2. Information on participants' backgrounds

Name	Educational Status
Amy	Masters of Science (MS) in Mathematics for Teachers (major) and graduate Teaching Assistant
Bernard	Undergraduate Mathematics (major)
Charles	Undergraduate Mathematics (major)
Danny	MS in Mathematics for Teachers (major)
Darla	Undergraduate Mathematics (major)
Eddie	MS in Mathematics for Teachers (major), MS in Mathematics (major), and graduate teaching assistant
Jonah	MS in Mathematics (major) and graduate teaching assistant
Margaret	Undergraduate Mathematics (major)
Martha	MS in Mathematics for Teachers (major) and pre-service teacher
River	MS in Mathematics for Teachers (major) and graduate teaching assistant
Rory	Undergraduate Mathematics (major)
Rose	MS in Mathematics for Teachers (major) and in-service teacher

The main topics of the course included sampling, sampling distributions, confidence intervals and hypothesis testing. The design of the course followed an RME framework where classroom activities focused on supporting students' reinvention. The class met twice a week for 75-minute sessions over a ten-week period.

The research team consisted of the principal investigator (the second author) who was the primary instructor for the course, and two graduate research assistants. During classroom sessions on hypothesis testing, the first author (graduate research assistant) developed the instructional sequence with consultation from the research team and instructed the class during this sequence. The role of the instructor was to lead class discussions centered on the instructional activities, to ask questions and to use student solution strategies as the basis for more in depth discussion of statistical concepts. The two assisting researchers videotaped classroom sessions, asked clarifying questions in small and large group discussions, and led some of the classroom teaching sessions. After each session, the research team met to discuss the progress of the teaching experiment. Meetings consisted of evaluating learning outcomes of past teaching episodes, and designing new tasks for the upcoming teaching cycles.

In the first half of the course, students were introduced to activities centered on sampling distributions and simulations. The course design emphasized sampling using computer simulators and physical sampling tools (e.g., buckets of beans, slips of papers, spinners, etc.) to help strengthen student conceptions of sampling distributions. Following the class sessions on sampling distributions, the teaching experiment progressed towards statistical inference topics. The tasks involving statistical inference focused on building informal and formal knowledge of confidence intervals and hypothesis tests. The research reported here focuses on the results of the classroom teaching experiment centered on the students' reinvention of an informal hypothesis test for categorical data.

3.2. INITIAL HLT: DEVELOPING AN INFORMAL HYPOTHESIS TEST FOR CATEGORICAL DATA

A primary goal of the study was to analyze how guided reinvention could be used to investigate students’ development of a hypothesis test for categorical data. The activity can be approached using a chi-squared goodness-of-fit hypothesis test. However, because this teaching approach utilized the theory of guided reinvention, we were not concerned if students arrived at the actual chi-squared goodness-of-fit test, only that they developed a meaningful informal hypothesis test (IHT) for categorical data. The choice of developing an activity where the problem could be solved using a chi-squared goodness-of-fit test was because of the intuitive nature of the test and the shape of the sampling distribution associated with the test. Prior to the start of the hypothesis test activities, these students spent significant time working with approximately normal-shaped sampling distributions. We wanted to observe how students would react when working with sampling distributions that were not normally distributed. In addition, we were interested in how students would address a problem that contained four categories in the population, rather than a population that could be analyzed using a single proportion (i.e., two categories). Given that these students had completed introductory statistics, they would have already seen hypothesis testing problems. Also, in the weeks leading up to the hypothesis test activities, students revisited sampling distribution concepts, and the idea of using sampling distributions as a tool to make informal inferences about the unusualness of an observed statistic for a single population proportion.

Prior to designing the instructional goals and sequence for the teaching experiment the research team hypothesized two initial strategies of students with a basic introductory statistics background. One student strategy we envisioned: students (forgetting or perhaps never having seen the chi-squared hypothesis test from their prior introductory statistics course) would consider each category of the population separately. The second student strategy we envisioned: students would create a procedure to summarize the observed data to generate a statistic (perhaps recalling a procedure from their prior introductory statistics course such as the chi-squared hypothesis test). We illustrate these two strategies with an example research question: Is a university’s student population equally distributed among freshman, sophomores, juniors and seniors?

In the first strategy, a student would generate four empirical sampling distributions of proportions for individual categories. Strategy 1, shown in Figure 2, illustrates the results of generating four empirical sampling distributions, one per category. We hypothesized that Strategy 1 might be intuitively appealing for students coming from introductory statistics where most of the class focuses on one-population hypothesis tests, but it turns out to be theoretically and intuitively problematic. Determining unusualness for the observed sample requires four statistics, one for each category in the sample, and each statistic needs to be compared to its respective empirical sampling distribution. We assumed once students attempted to make comparisons with four distributions, they would realize this strategy was not efficient. More importantly, Strategy 1 suffers from a misinterpretation of the Type-I error. In traditional hypothesis testing, the level of significance is set around a single sampling distribution. In IHT, the level of significance corresponds to the region of unusual statistics. If we set a 5% level of significance, then using four sampling distributions means setting a 5% level of significance per sampling distribution. As a result, the 5% level of significance is no longer the significance level for the problem as a whole.

Alternatively, Strategy 2, shown in Figure 2, uses one statistic to represent data from the observed sample, and this leads to a single empirical sampling distribution. This approach corresponds to traditional methods where a statistic is used to summarize sample information and a single empirical sampling distribution is produced under the null hypothesis. We anticipated at least one or two students would recognize the chi-squared nature of the problem and recall this strategy.

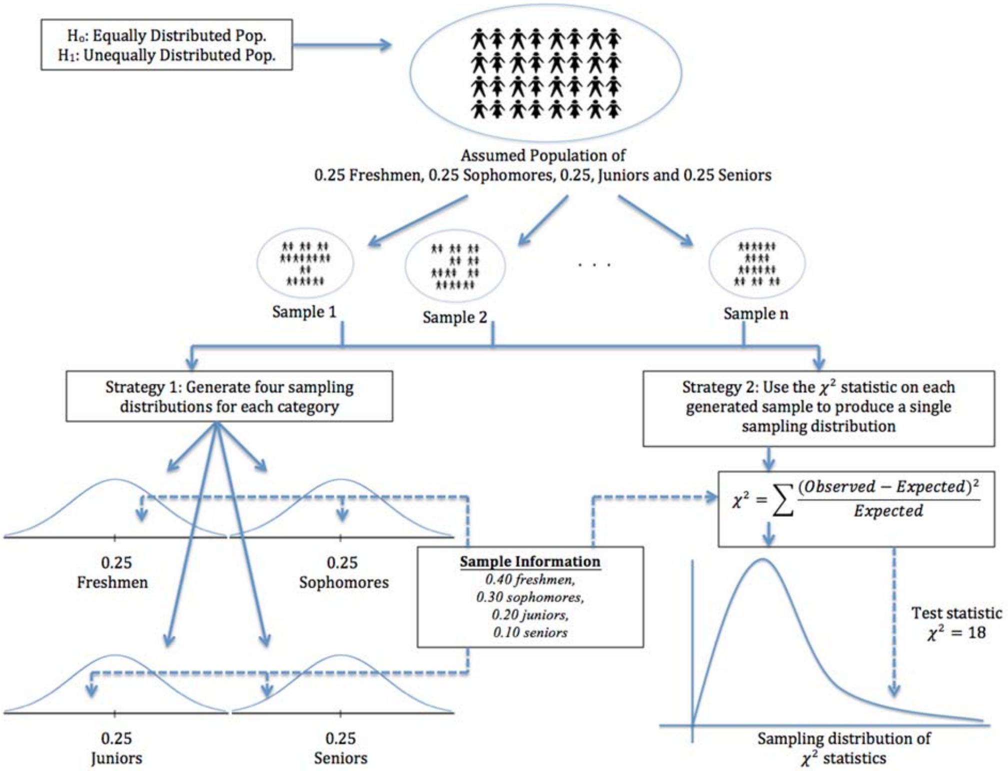
H_0: Equally Distributed Pop.
H_1: Unequally Distributed Pop.
Assumed Population of
0.25 Freshmen, 0.25 Sophomores, 0.25, Juniors and 0.25 Seniors
Sample 1
Sample 2
. . .
Sample n
Strategy 1: Generate four sampling distributions for each category
Strategy 2: Use the χ^2 statistic on each generated sample to produce a single sampling distribution
$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$
0.25
Freshmen
0.25
Sophomores
Sample Information
0.40 freshmen,
0.30 sophomores,
0.20 juniors,
0.10 seniors
Test statistic
$\chi^2 = 18$
0.25
Juniors
0.25
Seniors
Sampling distribution of
χ^2 statistics

Figure 2. Two sampling distribution strategies for a population with four categories.

Our larger instructional goal was to get students to generate Strategy 2. This meant that instruction needed to support student construction of a single statistic for a sample of categorical data. Given the mathematics background of these students, we assumed that if no one initially recalled a chi-squared test then at least a few students would construct a viable alternative method (though we did not know in advance what that method might be). We hypothesized that once students arrived at Strategy 2, or a viable alternative strategy that produced a statistic for the observed sample, discussion of the statistic's unusualness would follow the IHT approach (recall Table 1). We hypothesized that since students had already spent time doing this with a single proportion that they would be comfortable applying the same tools to a slightly new situation with a population of four categories.

3.3. BUILDING ON STUDENTS’ PRIOR KNOWLEDGE TO REFINE HLT

The design of the HLT and subsequent activities were constructed with the expectation that students would draw upon their prior knowledge of hypothesis testing, sampling distributions, simulations, and variability within a sample and between samples in an empirical sampling distribution. A primary conjecture of our teaching experiment was that instruction could begin with students’ intuition about variability in a sampling situation, which would transition to students’ quantifications of variability in a sampling situation, and then finally to student constructed procedures for finding a statistic to quantify the unusualness of sample data. Students would also generate an empirical sampling distribution based on the null hypothesis through simulations (building on prior knowledge of constructing sampling distributions from simulations in other situations). Finally, instruction would support the position of an observed sample’s statistic in an empirical sampling distribution as a tool for making a statistical inference. Erickson (2006) recommended this instructional approach, suggesting that students should be given the opportunity to construct methods to summarize numerically the unusualness of an observed sample and generate empirical sampling distributions as a vehicle to discuss unusualness. Table 3 outlines the approach designed to support students’ development of Strategy 2 in Figure 2.

Table 3. HLT: Learning activities and learning goals for constructing a hypothesis test for categorical data

Stage	Instructional Task	Learning Goals
1	Given a jar that contains four colors of beans in equal proportions (0.25 red, 0.25 white, 0.25 black, 0.25 brown), what would you expect the proportion of colors to be in a sample of 100 beans?	Explicit reasoning about the null hypothesis under an assumption that a jar contains equal proportions of black, brown, red, and white beans.
2	In your group, decide on a representative or usual range for the number of brown beans in a sample of 100. Repeat for other 3 colors.	Quantify representative intervals for each color. Consider expected variability from sample to sample.
3	Consider 8 different samples (see Table 4 for sample information) assumed to be from the jar of beans in equal proportions of red, white, black, and brown. Rank the samples from most to least unusual.	Construct methods for ranking the unusualness of the 8 different samples. Students quantify unusualness based on their methods.
4	Using the “best” (as decided by the class) method from the ranking task and the null hypothesis, generate a sampling distribution to model possible sample statistics from the assumed population.	Construct a simulation to model the population under the null hypothesis. Have class discussion on what the data values represent in the sampling distribution generated by the simulation.
5	Using the “best” method, compare the statistics constructed from the ranking task (Stage 3) to the sampling distribution constructed in Stage 4.	Determine a cutoff value for unusual versus representative samples and relate that value to an observed statistic. Make inferences based on location of the observed statistic within an empirical sampling distribution.

The context of the activity centers on a jar of beans with 4 different colors. This context was chosen because students had already worked with bean tasks in their previous sampling distribution activities (e.g., physically sampling from a bucket containing a population of two colors of beans: black and white). The new bean activity transitioned to a population consisting of four colors of beans. Stage 1 began with student discussions on the underlying hypothesis for the bean task. In Stage 2, students were asked to generate an interval for the number of brown beans in a sample of 100 that would be representative. The choice of which color to create a representative interval for is arbitrary since the jar contains equal proportions of red, brown, black, and white beans. This task served two underlying purposes. First, the intervals allowed the research team to gain information regarding the students' initial notions of variability in this context. We conjectured that students would intuitively generate a symmetric interval centered at 25 (e.g. 20-30 or 15-35) in which they would expect all brown (or black or red or white) beans should fall since the population is assumed to be equally distributed. We anticipated any justifications for the interval would be subjective, because each student likely has different perceptions on the expected variability of colors in any given sample. Second, we wanted to use student-generated intervals as a place to begin discussions for determining a cutoff value for quantifying unusualness in Stage 5.

Stage 3 of the teaching experiment was designed to facilitate students' construction of a statistic through student-generated methods of ranking samples from most to least unusual (Table 4 shows the sample information for the 8 samples). We anticipated that this would be a pivotal stage of the teaching experiment, supporting students' construction of a single statistic rather than four statistics for each color as noted in Strategy 1, Figure 2.

We conjectured that students would use the interval generated in Stage 2 as a tool to measure unusualness by counting the number of colors that fell within their expected interval. For instance, if students considered the interval (20, 30) to be their "expected" interval for each of the four colors, then they would rank the samples according to how many colors fell within this interval. In this example, Sample A would be ranked as the most representative because all the color counts are within the expected interval. Whereas Sample G would be ranked least representative because three of the color counts (black, red, and brown) are outside the expected interval. Anticipating some students would develop the strategy outlined above, the research team intentionally constructed samples to result in ties (e.g., Sample E

Table 4. Table of beans (sample size n = 100) provided for the ranking exercise

Sample	Count of Black	Count of Red	Count of White	Count of Brown
A	23	27	25	25
B	18	24	29	29
C	18	23	28	31
D	14	27	29	30
E	20	14	36	30
F	25	24	16	35
G	10	15	23	52
H	16	25	23	36

and H). We expected that the discovery of ties would cause cognitive conflict for the students, leading them to construct more refined methods of determining unusualness.

Because all the students had basic statistics and strong mathematics backgrounds, we expected them to develop methods to fix “ties.” Specifically, we expected students to develop methods that computed the difference between the observed colors (the sample’s distribution) and the expected colors (assumed distribution based on an equal proportion of each color). For example, we anticipated that students would make connections to the concepts of standard deviation and variance – thus, developing a formula similar to a chi-squared statistic (see Figure 3).

$$\sum (Observed - Expected)^2$$

Figure 3. Hypothesized student-constructed procedure for a population with four categories that are equally proportioned.

Once students generated a method for ranking the unusualness of the samples, the teaching experiment was designed to transition them toward generating an empirical sampling distribution based on the null hypothesis (Stage 4) where the elements of the sampling distribution would consist of the statistic generated in Stage 3 (hopefully a ‘modified chi-squared’ as in Figure 3 or an actual chi-squared). Discussions in class during this stage included: (1) the meaning of points in the sampling distribution; (2) how the null hypothesis was related to the generated empirical sampling distribution; (3) the relationship between the empirical sampling distribution and the student constructed statistic; and (4) the shape of the sampling distribution. We anticipated discussion around the shape of the sampling distribution to be challenging because of the right-skewed shape. We also expected students to struggle identifying the peak of the sampling distribution. For a chi-squared distribution, the peak is dependent on the degrees of freedom. However, for the purposes of our teaching experiment, we conjectured that through simulations students could visually explore the properties of the empirical sampling distribution they generated.

Stage 5 was designed to create an explicit discussion of the unusualness of sample statistics through an investigation of the correspondence between their initial interval (Stage 2), their statistic generated from their work in Stage 3 and their empirical sampling distributions (i.e., Stage 4). For example, if the students decided on intervals of (20, 30), then the maximum difference between the observed and expected values for each color bean would be 5. Squaring each value of 5 per color and summing would result in a value of 100 (Figure 4). The value 100 could be used as a *cutoff score*, the value used to distinguish unusual versus representative statistics in the sampling distribution. Finally, we hypothesized that the class could then compare the samples given in the ranking task and discuss the likelihood those samples came from a population with an equal proportion of black, brown, red, and white beans. Students could then make informal inferences through a consideration of the sampling distribution generated under the null hypothesis and their quantification of the unusualness of an observed sample statistic.

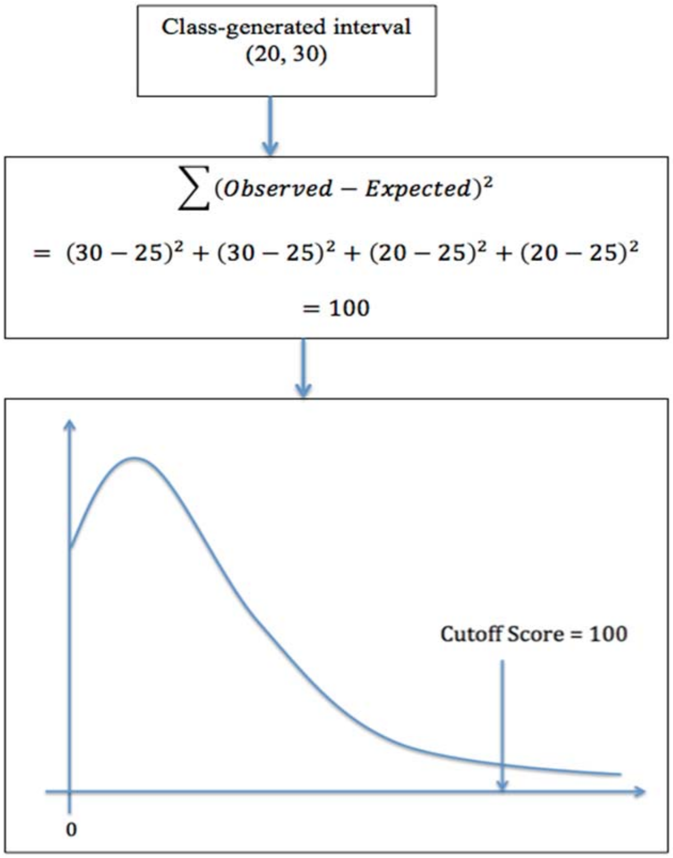

Figure 4. Relationship between interval, cutoff scores, and sampling distribution.

4. RESULTS

Through the instructional strategies outlined in the methods section, the research team gathered empirical evidence of students constructing an informal hypothesis test for categorical data (close to that of a chi-squared goodness-of-fit test). The methodological approach was valuable for allowing us deeper insight into the role variability played in these students' developmental processes. The analysis of the teaching episode revealed particularly important moments in students' development of a procedure for constructing a statistic for categorical data from Stage 3 through Stage 5. Therefore, our results focus on Stages 3-5. Five different student methods are presented in the results. The section will culminate in a revised HLT (based on the activities utilized in this teaching experiment) for students' reinvention of an informal hypothesis test for categorical data.

A brief comment about the results of Stages 1-2 is required before delving into the results of Stages 3-5. As anticipated by the research team, students provided very informal representative intervals for the number of brown beans (or black or red or white) in a sample of 100 beans. After each small group discussed and decided on an interval for their group, the whole class discussed possible intervals. The majority of groups gave intervals that stayed within plus or minus 5 from the value of 25. After much discussion, the class voted on one representative interval, (21, 29), and agreed that anything outside that interval would be considered unusual.

4.1. STAGE 3: GENERATING METHODS TO MEASURE *UNUSUALNESS*

After agreeing on representative intervals of (21, 29) for each bean color, students were prompted to develop a procedure to quantify unusualness (through the introduction of the ranking task, recall Section 3.3). The research team identified four numerical methods constructed by the class to rank the samples based on unusualness. The four student-generated methods we titled as: (1) general interval approach (GIA); (2) pairwise deviation approach; (3) range approach; (4) and the JC number approach. The first three methods shared similar and more intuitive ideas than the fourth method. In addition, the fourth method is the closest to the recreation of an actual chi-squared statistic. Therefore, this section summarizes the first three methods and the bulk of the Results section is devoted to the details of the JC number. To begin, Table 5 shows counts for red, black, white, and brown beans in samples A-H alongside the rankings for the different student generated approaches. There is no column for the GIA because none of the groups strictly used this method. Instead, students used GIA as a starting point, while also incorporating range or pairwise approaches to determine their rankings.

Table 5. Ranking of samples based on IPSTs' methods to measure unusualness

Sample	*Count of Black*	*Count of Red*	*Count of White*	*Count of Brown*	*Pairwise Deviation Approach**	*Range Approach**	*JC #**
A	23	27	25	25	1	1	1
B	18	24	29	29	2	2	2
C	18	23	28	31	3	3	3
D	14	27	29	30	5	4	4
E	20	14	36	30	7	7	6
F	25	24	16	35	4	5	5
G	10	15	23	52	8	8	8
H	16	25	23	36	6	6	7

*Numbers indicate the sample's rank from least unusual (1) to most unusual (8).

Method 1: General Interval Approach The GIA method counts how many of the four colors in each sample fall within (21, 29). For example, in Sample A all four colors fall within (21, 29). Thus, Sample A is ranked 'most representative'. Whereas in Sample E, none of the counts fall within (21, 29) so that sample would be labeled as most unusual. Only one student in the class strictly used this method. Once she experienced a tie, she moved on to a new method. The excerpt below shows Martha's iterative three-step approach that utilizes a GIA approach as an initial step.

Martha: Well, the way I did it, I had three steps. So the first thing I did was how many of each color of bean was in the range (21, 29). So was it four out of four, three out of four. … So first, how many of each are in the range? And then I ranked them by that. Then if there was a tie, how many beans was I away from twenty-five across the board. And if that was tied, which it was for two of them, how many away from the range was I? ...Yeah, the last step was how many was I away from the range of twenty-one to twenty-nine?

Most other students in the class took issue with Martha's three-step approach because they felt that Sample G was more unusual than Sample E. River expressed the general sense of the class when she suggested that Sample G should be considered more unusual than sample E.

River: That's really interesting! … I really like that, but I have a really hard time with [Sample] G not being the last one because again, if we are going for, if all four of the true population is that they are even and one is ten and one is fifty-two. That's about as least representative that you can possibly be.

Instructor: So your problem is this fifty-two?

River: Yeah, the fact that that method leads to that one [Sample G] not being the least of all of them…

The fact that many students took issue with Martha's method because her technique favored inclusion within the expected interval over a consideration of the distances of observed values from the expected value or expected range of values is a positive sign that on an intuitive level student were thinking about deviation from some expected value. As illustrated by Martha, the GIA was only a stepping-stone that led students to construct other strategies to replace GIA or use GIA as part of an iterative process.

Method 2: Pairwise Deviation Approach Two student groups considered the distance of the two extreme values (of each sample) from the endpoints of the class-generated interval (21, 29). For example, in Sample G the two extreme values are the 10 black beans and 52 brown beans. The absolute distance of 10 from 21 is 11 and the absolute distance of 52 from 29 is 23. Students using this pairwise deviation approach considered those absolute differences as a way to determine unusualness. Table 6 highlights the pairwise deviation approach, showing the absolute differences between the minimum of a sample and the lower value of the interval (21) and the maximum of a sample and the interval's upper value (29).

The next excerpt highlights a pairwise deviation approach by one student who debated whether Sample D was more representative than Sample H (or vice versa).

Bernard: The min minus... so if you have the range and 21 and 29, it would be twenty-one minus the minimum. …Well, you want the absolute value or the maximum minus 29 and then whichever number is larger, that would be the error on that one and then you'd find the one that has, compare that to the same thing for another category. [36-29 and 21-16 for Sample H and 21-14 and 30-29 for Sample D, finding the largest difference to rank them]

Table 6 reveals that Samples D and H have at least one absolute difference of seven, so in those instances students considered both difference values in the pair. Sample D is ranked as more representative than Sample H because the other difference value in Sample D is 1 compared to 5 in Sample H.

Table 6. Ranking of pairwise deviation approach

Sample	Count of Black	Count of Red	Count of White	Count of Brown	|21–min|	|max-29|	Rank (Pairwise Deviation)*
A	23	27	25	25	|21–23|=2	|27–29|=2	1
B	18	24	29	29	|21–18|=3	|29–29|=0	2
C	18	23	28	31	|21–18|=3	|31–29|=2	3
D	14	27	29	30	|21–14|=7	|30–29|=1	5
E	20	14	36	30	|21–14|=7	|36–29|=7	7
F	25	24	16	35	|21–16|=5	|35–29|=6	4
G	10	15	23	52	|21–10|=11	|52–29|=23	8
H	16	25	23	36	|21–16|=5	|36–29|=7	6

*Numbers in the column indicate least unusual (1) to most unusual (8)

Method 3: Range Approach Another student strategy was based on subtracting the minimum from the maximum value in each of the samples to rank unusualness. This strategy was another method considered by Amy's group as shown in the excerpt below.

Amy: Yeah, it's similar to that [the GIA and the pairwise deviation approach]. I was looking at (Sample) D and seeing how my highest is thirty, for the brown. But my lowest is fourteen (for the black) so like.... That range has a really low range [30-14=16] compared to [Sample] H, where I have a high of 36, but a low of 16 [36-16=20], so I feel like it falls into a better range.

Table 7 shows the ranking of samples according to the range approach. Those samples with the largest overall range would be more unusual. Thus, Sample H was more unusual because its overall range is 20 compared to a range of 16 for Sample D. The ranking of the overall range approach (see Table 7) has a similar ordering as the pairwise deviation approach, with the exception of Samples D and F.

Table 7. Ranking of range approach

Sample	Count of Black	Count of Red	Count of White	Count of Brown	Max Count – Min Count	Ranking Range Approach*
A	23	27	25	25	27–23= 5	1
B	18	24	29	29	29–18=11	2
C	18	23	28	31	31–18=13	3
D	14	27	29	30	30–14=16	4
E	20	14	36	30	36–14=22	7
F	25	24	16	35	35–16=19	5
G	10	15	23	52	52–10=42	8
H	16	25	23	36	36–16=20	6

*Numbers in the column indicate least unusual (1) to most unusual (8)

Method 4: JC Number Jonah's procedure compares the differences between the observed and the expected values per color, squares the differences, sums the squared differences of all four colors and finally takes the square root the sum to get a numerical value of unusualness for a sample (Figure 5).

$$\sqrt{\sum(Observed - Expected)^2}$$

Figure 5. Jonah's description of the JC number.

The JC number was later modified during class because the class decided to disregard the square root for ease of calculation. The resulting ranking is illustrated in Table 8.

Table 8. Ranking of samples using the JC number procedure

Sample	Count of Black	Count of Red	Count of White	Count of Brown	Modified JC Number*	Ranking of the JC # procedure
A	23	27	25	25	8	1
B	18	24	29	29	82	2
C	18	23	28	31	98	3
D	14	27	29	30	166	4
E	20	14	36	30	292	6
F	25	24	16	35	182	5
G	10	15	23	52	1058	8
H	16	25	23	36	206	7

* Modified JC = (25–count Black)^2 + (25–count Red)^2 + (25–count White)^2 + (25–count Brown)^2

The students who constructed the JC number had an approach distinct from the other methods because they did not use the representative interval (21, 29) from Stage 2 to construct their procedure for ranking. Their method raised questions for the other students. The following excerpt shows the class discussion about how useful the class-generated interval was for the ranking task.

Jonah: We didn't take our pre-determined range [21 to 29] into account at all.
River: I almost think our range [21 to 29] kind of confused us a little bit. I don't know. I'm still stuck on the fact they [samples] are all supposed to be equal [all colors have the same proportion].

In this excerpt, we see students discuss the usefulness of the class-generated interval in their ranking methods. Jonah mentions that his method did not consider the class-generated interval. River also elaborates that the intervals might have caused confusion in the ranking task because they were not directly taking into account the assumption that each category should be equally distributed. Overall, the class favored the JC number because it appeared more mathematical and looked like a standard deviation calculation. In addition, their method gave them a single measure for ranking as opposed to methods (such as pairwise deviation) which gave two measures for ranking.

4.2. STAGE 4: SAMPLING DISTRIBUTION OF JC NUMBERS

After the class discussed Jonah and Charles' constructed JC number, the class decided to use this method as the preferred procedure for finding a statistic from the observed data. The instructional strategy shifted to the construction of a sampling distribution of JC numbers (from an assumed population of four colors of beans in equal proportions) and ways to create connections between the sampling distribution, observed statistic and informal inference ideas (Stages 4 and 5). At the start of this teaching episode, students were asked to: (1) imagine taking many samples of beans from the population of four colors with equal proportions; (2) imagine computing the JC number of each sample; and (3) construct a dot plot of the JC numbers based on their images of what the distribution would look like. The excerpt below is from one group's discussion.

Eddie: So that's our spread, from 0 to 75 squared. Now the center as far as distribution well if we expected to get 0...no.
Jonah: It can't be centered if...
Eddie: That's kind of weird. But what I'm saying is as far as repeating this over again we expect to get mostly representative samples right?…So if we're talking about the spread of this distribution of data it's going to look something like this [Figure 6] right? That is most of our values will be mostly around here [see * in Figure 6]
Jonah: So if we didn't square it we would have negatives…
Eddie: Exactly! Exactly. That's like the effect of squaring. …It's kind of lopsided. And I think that actually fits with the chi-squared values for low degrees of freedom. It's going there. If the degrees of freedom are low, like 4 or 3, you have that kind of lopsided distribution.

Eddie begins the discussion by directing the group's attention to the location of the least and most representative samples in his sampling distribution (Figure 6). He gets 75^2 by considering a case in which the entire sample is one color (he forgets about adding in three of the $(0 - 25)^2$ terms for the rest of the JC number computation). During this moment Eddie and his partner, Jonah, realize that there is something perplexing about the distribution at zero. However, the two move on to describing the general shape of a chi-squared distribution. It is interesting that Eddie now sees the connection to chi-squared distributions, though he did not recognize the ranking task as a situation in which to apply a goodness-of-fit test in Stages 1-3.

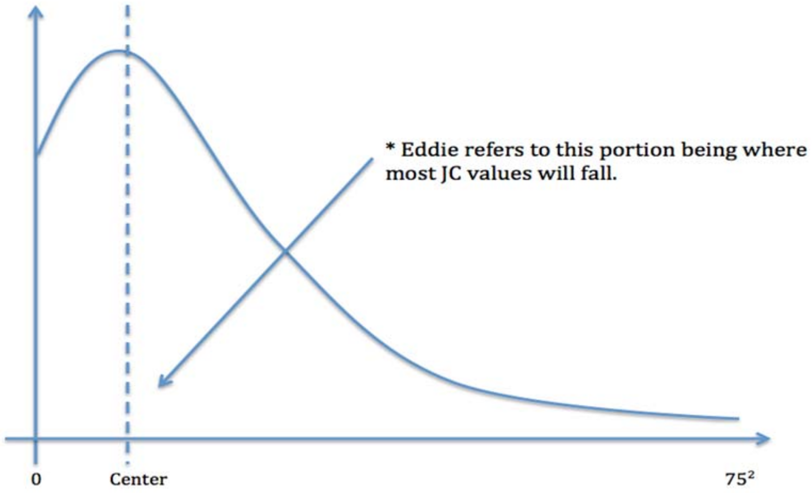

Figure 6. Eddie's conjectured sampling distribution.

After giving the students the opportunity to discuss and conjecture within their groups, Eddie presented his group's findings to the class.

Eddie: We started by considering the range because that seems like an easy place to start. So at best we pick a sample and it's perfectly representative, exactly 25% of everything. The value of the JC number we got would be 0 and at worst it is entirely composed of 1 color. It doesn't really matter which one. In this case, it would have a JC number of 75 squared.

Class: It should be plus 3 times 25 squared, 'cause zero minus…

Eddie: Oh yeah 0 minus 25. Oh shoot, so it could be even worse. Great! So… 75 squared plus 3 times 25 squared.

Here the class makes sense of Eddie's range idea for the most and least representative cases, and correct his calculation error. Building on Eddie's presentation, the class discusses the idea of a *perfectly representative* sample (25 of each color) and the *least representative* sample (100 of a single color). As Eddie continued to share his group's graph and the reasoning they used, discussions of shape and symmetry arose.

Eddie: And then uh...we uh... we would expect if we did this repeatedly over a long period of time that most of our samples would be representative within some range of uh that kind of golden JC number of zero. But it would be kind of weird to get an exactly representative sample. So we sort of fudge this right here [points at the peak of Figure 6]. There is going to be some sort of a magic number that's uh in the center of all this [Figure 6]. And uh...but as then you got more and more outlandish unrepresentative samples you expect to see that less and less often.

Instructor: Any comments or questions?

Danny: I kind of doubt that you get close to that many zeros...

Eddie: That’s a good point. I feel very uncertain of placing it here versus here [pointing to the frequency of the graph at 0, see Figure 6]. I mean the original we drew was kind of a straight drop but yeah...you are probably right it [the frequency of the plot at 0] is probably lower down here because it seems like picking an exact...I mean exactly 25 of each seems like kind of a problem.

Instructor: Okay so my question is how is this different from like the other distributions you have been seeing? Like say the normal distribution?

Eddie: Yeah right. The big...kind of obvious difference is that this [Figure 6] one is lopsided...it [refers to a normal curve] doesn't have a tail coming off at this end [points to the left side of the graph] while all the others distributions have something like that.

Jonah: Not symmetric.

In these excerpts, Eddie articulates some of the characteristics of a sampling distribution of JC numbers. At this point the class appears to converge upon the shape of a sampling distribution for the JC number, an approximate chi-squared graph. For example, Eddie mentions that the shape of the sampling distribution would be “lop-sided” and Jonah suggests a non-symmetric graph. The discussion continues with Rory challenging the expected frequency of a JC value of 0 and Eddie admits that his group was uncertain. Eddie’s group began with the mode at 0 and a graph that decreased from there. Other students in the class agreed that a JC number of 0 would be perfectly representative, but would rarely occur. By discussing the commonality of a JC number of 0, students questioned whether the graph they are sketching is theoretical or empirical.

Researcher: So can you say a little more Rory about when you said you are doing an experimental graph?

Rory: Yeah...I guess we were thinking about it sort of theoretically and then zero falls highest in the curve.

Eddie: Interesting. That’s...you know we...we started drawing that picture. We started here and dropped down here [pointing at Figure 6 and motioning at the graph’s height at 0].

Rory: As you did more and more samples it seems you’d get more zeros.

Eddie: Yeah I...I feel you there but I don’t feel like a good strong theoretical reason for that. I mean like...so...so maybe that’s true. I’d like to see it.

Rory seems to be arguing that the theoretical graph would have a mode at 0 because as more samples are taken, there would be more chances of getting representative samples, thus, JC numbers of 0. However, he also seems to argue that empirical graphs have a lower occurrence of 0’s and Eddie’s graph in Figure 6, being empirical, has too many counts at 0. Eddie does not necessarily disagree with Rory but suggests there should be a theoretical argument for this.

4.3. STAGE 5: DISCUSSING CUTOFF VALUES

After giving students the opportunity to discuss the sampling distribution of JC numbers, the instructor asked students to differentiate unusual versus representative samples using the JC number as their statistic (from Stage 3) relative to the empirical sampling distribution of statistics (Stage 4). The students were also asked to relate the class-generated interval (from Stage 2) with the JC number procedure (Stage 3) to create a JC number corresponding to a cutoff value for unusualness in their conjectured sampling distribution (Stage 4).

River: So we assumed that... the worst case scenario is each of the numbers are 4 away from 25 and that's the worst case scenario that our range said is still good.
Martha: And take the four of those squared and you get 64. So less than or equal to 64 [(i.e., $(21 – 25)^2 + (21 – 25)^2 + (29 – 25)^2 + (29 – 25)^2 = 64$)].
Rory: Is representative.
Bernard: The worst case would be 21, 21, 29, 29.
Rory: I think it shows us that our range is just...personally...we’re not there yet.
Instructor: What if you wanted to expand your range. ... How would that affect your JC number?
Rory: It would increase it.

In this excerpt, students describe a threshold of unusualness utilizing the endpoints of the class-generated interval (Stage 2) by describing the values of 21, 21, 29 and 29 as the worst possible sample within the interval. They then used the endpoint values as inputs to find a JC number of 64. Students also mention that increasing the size of the class-generated interval would generate a new critical cutoff value, illustrating the correspondence between the JC number procedure and the class-generated interval.

After giving students an opportunity to hypothesize about the sampling distribution of JC numbers and a cutoff value, they were shown a simulated sampling distribution of JC numbers generated from a population of equal proportions (Figure 7).

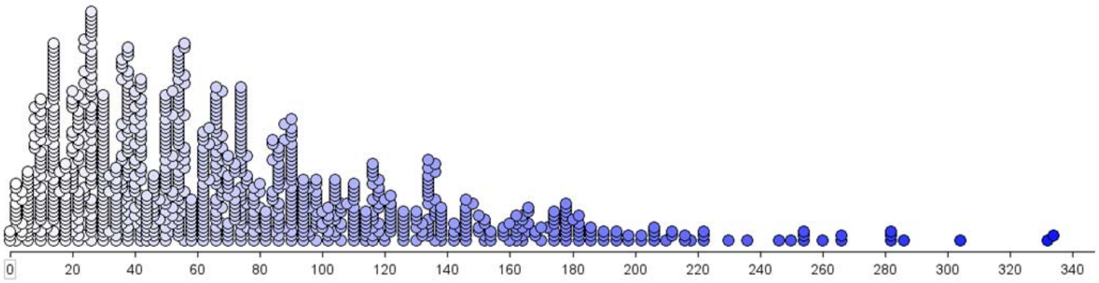

Figure 7. Sampling distribution of JC numbers from a uniformly distributed population

After viewing the empirical sampling distribution, many of the students disagreed with the previous cutoff value of 64 due to the large cluster of points around 64 in the dot plot. To give students a chance to discuss the unusualness of samples in light of the empirical sampling distribution, they were given the opportunity to discuss a new cutoff value and corresponding interval.

Darla: So if you are looking at an interval that was 18 to 32...so the max difference is 7. So, looking at the max cutoff. …I feel like...like then if we're cutting off these things (refers to points on sampling distribution to the right of the cutoff value) and we know that these are possible then...like.

Eddie: So the idea behind it is...is you are presented with a random sample and you don't know this...and you have no idea. And the question is what is the probability the random sample actually came from a population that looks like this which is not.... The way this is actually used is, you make a hypothesis about your population and then you take your sample and say if the hypothesis were true, how unlikely is this sample?

Jonah: Let's count the number that we get 10 of these [points on the sampling distribution] to the right? That will give us 5%, which would be more extreme.

In this excerpt, we notice students discussing percentages and potential cutoff values of JC numbers in relation to the empirical sampling distribution. Darla suggests widening their initial interval (18, 32). She also discusses deterministic versus probabilistic reasoning when she appears uncomfortable creating a cutoff point when outcomes outside that cutoff are still possible. Eddie attempts to explain the probabilistic nature of the problem and Jonah mentions the extreme 5% of JC numbers in the empirical sampling distribution. Eddie also describes the purpose of a cutoff value as a way to determine unusualness of a random sample by relating it back to his understanding of hypothesis testing.

5. DISCUSSION

The goal of this study was to examine how students might come to construct a viable hypothesis test for categorical data through guided reinvention. We also wanted to investigate the role variability might play in students' development of a procedure for finding a statistic from the observed data, the concept of a critical value and the concept of sampling distributions based on the null hypothesis. Through our research we uncovered some evidence that students were able to generate an informal version of a hypothesis test for categorical data and that student constructions could be used to leverage ideas of an empirical sampling distribution and cutoff values for unusual statistics. This section focuses on a synthesis of the results presented in the previous section with careful consideration of students' construction of a procedure for calculating a statistic from the observed data, students' construction of an empirical sampling distribution, and the class-generated interval.

Through guided reinvention we studied student-constructed statistics for measuring unusualness. Figure 8 summarizes the different methods outlined in the Results section as well as the true chi-squared statistic formula. Many of the methods generated by students

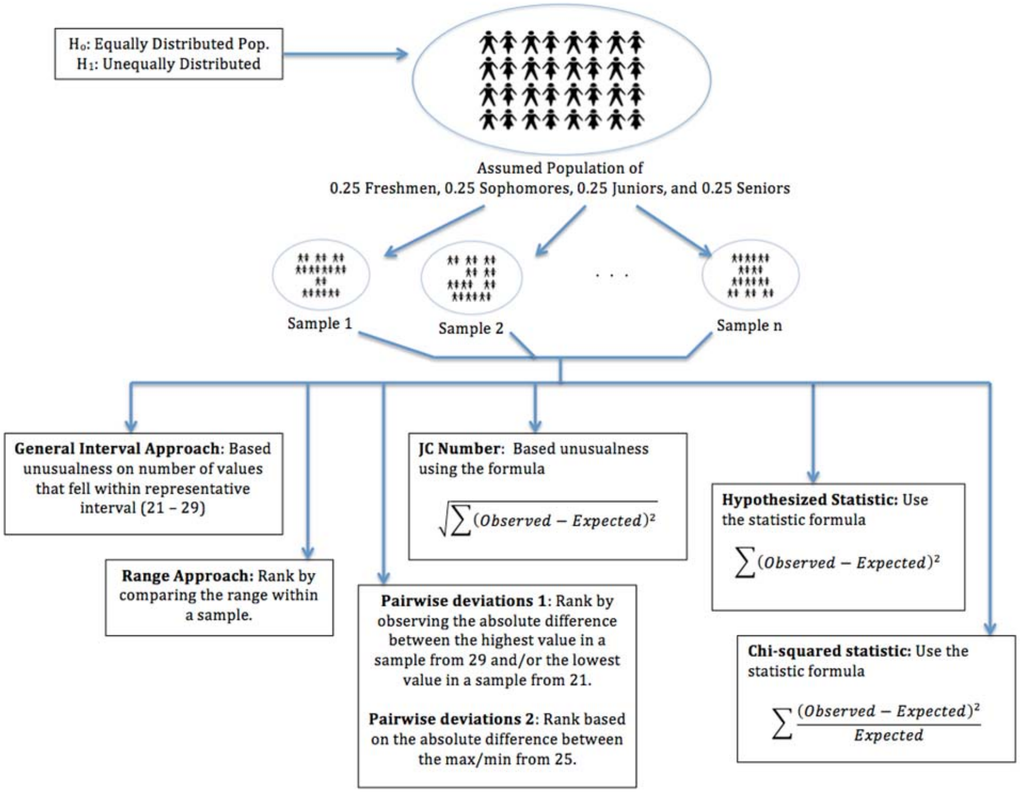
H_0: Equally Distributed Pop.
H_1: Unequally Distributed
Assumed Population of
0.25 Freshmen, 0.25 Sophomores, 0.25 Juniors, and 0.25 Seniors
Sample 1
Sample 2
. . .
Sample n
General Interval Approach: Based unusualness on number of values that fell within representative interval (21 – 29)
Range Approach: Rank by comparing the range within a sample.
Pairwise deviations 1: Rank by observing the absolute difference between the highest value in a sample from 29 and/or the lowest value in a sample from 21.
Pairwise deviations 2: Rank based on the absolute difference between the max/min from 25.
JC Number: Based unusualness using the formula
$\sqrt{\sum(Observed - Expected)^2}$
Hypothesized Statistic: Use the statistic formula
$\sum(Observed - Expected)^2$
Chi-squared statistic: Use the statistic formula
$\sum\frac{(Observed - Expected)^2}{Expected}$

Figure 8. *Students' methods to rank unusualness.*

have similarities to normative statistical measures of variation. For example, the range and pairwise deviation approaches have similarities to range, interquartile range, and absolute deviations. The JC number is very close to the actual chi-squared statistic and bears resemblance to the formula for standard deviation. In this respect, the teaching experiment was successful because the students were able to construct meaningful methods for ranking unusual samples, which supported their understanding of viable hypothesis tests for categorical data. We hypothesize that this instructional approach gave students a deeper understanding of the construction of statistics and their role in hypothesis testing. Even this study's students who did not create the JC number still used methods that created a platform for meaningful discussions regarding hypothesis testing concepts and the construction of empirical sampling distributions based on student-generated measures.

One of the key components of an IHT approach is to put empirical sampling distributions at the core of hypothesis tests. We theorized that supporting students' construction of a procedure to quantify the unusualness of a sample for categorical data could result in a deeper understanding of the empirical sampling distributions, which can be used as a tool in making inferences. This seemed to be the case. Once the class decided on the JC number as their statistic and they constructed a sampling distribution of JC numbers, they were able to discuss the properties of this sampling distribution in great detail and its role in making inferences. Evidence of students' deeper understanding of sampling distributions is illustrated in two particular class discussions. In the first discussion, the class developed a picture of the sampling distribution of JC numbers and engaged in meaningful discussions about the center, shape and spread of the distribution. Second, the class demonstrated an ability to adapt to changing cutoff values of unusualness in the sampling distribution. For example, students made connections between the class-generated interval (21, 29) and the corresponding JC number as a means to generate an initial cutoff value on the empirical sampling distributions for unusual samples. This approach illustrates students making connections between their initial intuitive notions of unusualness towards more formal measures they developed through their work and the empirical sampling distribution they constructed. We hypothesize that because students personally constructed the statistic (JC number) used to generate the empirical sampling distribution, we were able to leverage deeper thinking about a sampling distribution in a statistical inference argument. This discovery aligns with the research of Erickson (2006), who suggested that giving students the opportunity to create their own procedures for finding a statistic from the observed data could lead to a deeper understanding of hypothesis testing.

Finally, the class discussions around the class-generated interval illustrates the challenge students have coming up with an expected range of variability from sample to sample. Students do not have good internal measures for variability especially without experiences working with repeated samples. The representative interval voted on by the class (i.e., 21 to 29) ended up being extremely conservative. By the end of the teaching experiment, students were able to reflect on this initial interval. In particular, when students found the JC number corresponding to their interval values they realized their initial choice was too conservative. This was a pivotal and culminating moment in the teaching experiment because students had the opportunity to connect their intuitive choice for a representative interval to the class-constructed statistic (JC number) and the resulting empirical sampling distribution. This illustrates how a sampling distribution can support students' understanding of variability. This approach to learning hypothesis testing aligns with the research literature that suggests the importance of understanding sampling distributions in the learning of statistical inference (e.g., Chance & Rossman, 2006; Erickson, 2006; Garfield & Ben-Zvi, 2008; Weinberg et al., 2010). Rarely would a traditional course offer the opportunity for making these connections and allow students to reflect on prior assumptions. The class-generated interval also illustrates the tendency of students to consider only small changes in variability. The four-color bean task is especially difficult because variation must be determined among multiple dependent categories. By analyzing an empirical sampling distribution of statistics, students can openly discuss variability for simple and complex samples.

6. CONCLUSIONS AND FURTHER RESEARCH

In the beginning of the paper, we presented two research questions on hypothesis testing:

1. How do in-service and pre-service teachers (IPSTs) move from informal intuitions toward more formal concepts of hypothesis testing for categorical data using a guided reinvention approach?
2. What role does IPSTs' intuition about variability play in constructing empirical sampling distributions and developing procedures for finding a statistic and critical value used in a hypothesis test for categorical data?

The research presented here used a guided reinvention methodology to study students' construction of an informal hypothesis test for categorical data utilizing intuitions about variability and incorporating sampling distributions and simulations. Our work was successful in facilitating students' construction of these fundamental concepts of hypothesis testing. The HLT we developed (including learning goals and subsequent instructional activities) was useful in facilitating students' thinking about statistics, sampling distributions and informal inference. In addition, our work contributes to current research by illustrating a potential conceptual framework (Figure 8) of student development. This framework outlines a model on how students might develop a procedure for finding an observed statistic when working with categorical data.

While our work makes a contribution to research on student thinking about hypothesis testing there were limitations to the study. First, the sample size is small (12 students) and may not be representative of a larger population. Second, these students were prospective and/or current math teachers who had strong mathematics backgrounds. We do not know if our instructional activities would support introductory statistics students, with limited mathematics background, reinvention of an informal hypothesis test for categorical data. However, it is our conjecture that the intuitive nature of the chi-squared statistic in a goodness-of-fit test makes it a viable starting point for introductory statistics students and that the potential for guided reinvention among this population is possible. We recommend future work studying this HLT with this population.

A third limitation is that we wonder how the construction of representative intervals in Stage 2 impacted the methods students developed for ranking samples in Stage 3. One student (River) mentioned that the intervals confused her as she struggled to rank the unusualness of the samples (see excerpt in results). Furthermore, the group that developed the JC number mentioned that the class-generated interval was unnecessary for the method they constructed. The purpose of a representative interval was to generate discussions regarding intuitive ideas of unusualness and as a component to the construction of cutoff scores. Yet, it is possible that this approach restricted student thinking and subsequently impacted the types of methods they developed.

In a future research study, it would be interesting to see how students would approach the ranking task without Stage 2. One recommendation is to modify the HLT where Stage 2 is removed. A possible new stage 2 would consist of activities where students 'reinvent' mean absolute deviation, standard deviation, and/or other methods for measuring variability within data sets prior to the bean activity. The time spent developing students' reinvention of mean absolute and standard deviation may support their development of procedures for finding an observed statistic in a hypothesis testing context related to categorical data. In this study, we presented a case where students generated a procedure for finding an observed statistic for a population with four categories with equal proportions. Future research might have students expand their knowledge by considering a population with unequal proportions. This may allow students to investigate the limitations of the procedures they develop. For example, in the case of unequal proportions, the JC number would not be appropriate because having different proportions in each category results in differences in variability between categories. Students would need to consider how they could modify the JC number to accommodate this difference. Class discussions on these limitations could focus on the importance of developing procedures that account for weighted measures and potentially lead students towards generating the true chi-squared statistic.

Future research needs to study student development of a semi-formal construction of a hypothesis test for categorical data to the formal technique of a chi-squared goodness-of-fit test. The research presented here provides a new approach to the teaching and learning of hypothesis testing. It suggests that instruction could begin with informal tests for categorical data based on student intuitions of variability in a simple bean context and lead to the development of a formal chi-squared goodness-of-fit test. It is not only reasonable but also potentially beneficial for students to spend time reinventing an informal chi-squared goodness-of-fit test. Such reinvention may support students' understanding of sampling distributions, variability, test statistics, cutoff values and how to use these concepts to make inferences. Furthermore, we argue that guided reinvention is particularly important for future teachers of statistics because it has the potential to strengthen teachers' understanding of sampling distributions in hypothesis testing, support teachers in developing new approaches to their teaching that are more student centered and, subsequently, better support student learning through their new understandings.

ACKNOWLEDGEMENT

The authors express appreciation to Sonya Redmond (Portland Community College), who collaborated with the design of the research and data collection during her graduate research assistantship at Portland State University.

REFERENCES

American Statistical Association (2012). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author.
[Online: http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf]

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, The Netherlands: CD Beta Press.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator*, *29*(1), 14–17, 20–22, 43–46.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1-2), 75–98.

Batanero, C., & Diaz, C. (2006). Methodological and didactical controversies around statistical inference. *Proceedings of 38th Conference of the French Statistical Conference*. *Actes du 36iémes Journées de la Societé Française de Statistique*. CD-ROM. Paris: Societé Française de Statistique.

Bluman, A. G. (2012). *Elementary statistics: A step by step approach* (8th ed.). New York: McGraw Hill.

Brase, C. H., & Brase, C. P. (2012). *Understandable statistics: Concepts and methods* (10th ed.). Boston, MA, Brooks/Cole Cengage Learning.

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. V. den, & 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.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education Journal, 1*(1), 1–26.

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). Netherlands: Kluwer Academic Publishers.

Chance, B., & Rossman, A. (2006). *Investigating statistical concepts, applications, and methods*. Belmont, CA: Thomson Brooks/Cole.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1–78.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3).

Doorman, L. M., & Gravemeijer, K. (2009). Emergent modeling: discrete graphs to support the understanding of change and velocity. *ZDM – Mathematics Education, 41*(1), 199–211.

Erickson, T. (2006). Using simulation to learn about inference. *Proceedings of the Seventh International Conference on Teaching Statistics*.
[Online: http://iase-web.org/documents/papers/icots7/7G2_ERIC.pdf]

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*(1), 83–96.

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

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

Garfield, J., & Everson, M. (2009). Preparing teachers of statistics: A graduate course for future teachers. *Journal for Statistics Education, 17*(2), 1–15.

Gravemeijer, K. (2004). Creating opportunities for students to reinvent mathematics.
Proceedings of the Tenth International Congress in Mathematics Education (pp. 4–11).
Denmark.
[Online: http://www.staff.science.uu.nl/~savel101/edsci10/literature/gravemeijer1994.pdf]

Gravemeijer, K., & Bakker, A. (2006). Design research and design heuristics in statistics education. *Proceedings of the Seventh International Conference on Teaching Statistics*
[Online: http://iase-web.org/documents/papers/icots7/6F3_GRAV.pdf]

Gravemeijer, K., & Van Galen, F. (2003). Facts and algorithms as products of students' own mathematical activity. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 114–122). Reston, VA: National Council of Teachers of Mathematics.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–20.

Heid, M. K., Perkinson, D., Peters, S. A., & Fratto, C. L. (2005). Making and managing distinctions – The case of sampling distributions. In *Proceedings of the 27th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Vicksburg, VA.
[Online: http://citation.allacademic.com/meta/p18727_index.html]

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–155). Charlotte, NC: Information Age Publishing.

Larsen, S., Johnson, E., Rutherford, F., & Bartlo, J. (2009). A local instructional theory for the guided reinvention of the quotient group concept. Proceedings for the Twelfth Special Interest Group of the Mathematical Association of America on Research in Undergraduate Mathematics Education Conference on Research in Undergraduate Mathematics Education, Raleigh, NC.
[Online: http://sigmaa.maa.org/rume/crume2009/Larsen_LONG.pdf]

Lipson, K. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematics Education Research Journal*, *15*(3), 270–287.

Liu, Y., & Thompson, P. (2005). Teachers' understandings of hypothesis testing. In *Proceedings of the 27th Annual Meeting of the International Group for the Psychology of Mathematics Education*. Vicksburg, VA.
[Online: http://pat-thompson.net/PDFversions/2005PMENA%20HypTest.pdf]

Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies*, *4*(2), 126–138.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In Dani Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353–373). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Oehrtman, M., Swinyard, C., Martin, J., Hart-Weber, C., & Hah Roh, K. (2011). From intuition to rigor: Calculus students' reinvention of the definition of sequence convergence. *Proceedings of the Fourteenth Annual Conference on Research in Undergraduate Mathematics Education* (Vol. 3, pp. 137–140). Portland, OR: Special Interest Group of the Mathematical Association of America for Research in Undergraduate Mathematics Education.

Rubin, A., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance* (pp. 241–256). Reston, VA: National Council of Teachers of Mathematics.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Swinyard, C. (2011). Reinventing the formal definition of limit: The case of Amy and Mike. *Journal of Mathematical Behavior*, *30*(2), 93–114.

Thompson, P., Liu, Y., & Saldanha, L. (2007). Intricacies of statistical inference and teachers' understandings of them. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207–231). Mahwah, NJ: Lawrence Erlbaum.

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypothesis testing by university students. *Themes in Education*, *3*(2), 183–198.

Vallecillos, A., & Batanero, C. (1997). Conceptos activados en el contraste. De hipótesis estadísticas y su comprensión por estudiantes universitarios (Activated concepts in the statistical hypothesis contrast and their understanding by university students). *Recherches en Didactique des Mathématiques*, *17*(1), 29–48.

Van Etten, B., & Adendorff, S. (2007). Discovering Pythagoras' theorem through guided re-invention. Education Papers and Reports. Paper 26. Cape Peninsula University. [Online: http://www.academia.edu/7798261/RME_Discovering_the_Pythagorean_theorem_through_guided_reinvention]

Weinberg, A., Wiesner, E., & Pfaff, T. (2010). Using informal inferential reasoning to develop formal concepts: Analyzing an activity. *Journal of Statistics Education*, *18*(2), 1–24.

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), 1–25.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. [Online: http://iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf]

JASON DOLOR
Portland State University
Fariborz Maseeh Department of
Mathematics and Statistics
PO Box 751
Portland, OR 97207