Reasoning About One Population Hypothesis Testing: The Case of Steve

Annie Burns-Childers
University of Arkansas
Little Rock

Darryl Chamberlain Jr.
University of Florida

Aubrey Kemp
Georgia State University

Leslie Meadows
Georgia State University

Harrison Stalvey
University of Colorado
Boulder

Draga Vidakovic
Georgia State University

Hypothesis testing is a key concept included in many introductory statistics courses. Yet, due to common misunderstandings of both scientists and students, the use of hypothesis testing to interpret experimental data has received criticism. With statistics education on the rise, as well as an increasing number of students enrolling in introductory statistics courses each year, there is a need for research that investigates students' understanding of hypothesis testing. This paper describes results obtained from a larger study designed to investigate introductory statistics students' understanding of one population hypothesis testing. *In particular, we present on one student's understanding of the concepts involved in hypothesis testing, Steve, who provided us the best spectrum of different levels of knowledge according to APOS Theory, our guiding theoretical framework. Based on this data, we suggest implications for teaching.*

Keywords: Hypothesis Testing, Introductory Statistics, APOS Theory

Introduction

The use of statistics is crucial for numerous fields, such as business, medicine, education, and psychology. Due to its importance, according to the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*, more students are studying statistics, and at an increasingly younger age (GAISE College Report ASA Revision Committee, 2016). As a result, the *GAISE College Report* calls for nine goals for students in introductory statistics courses. One of these nine goals is that "Students should demonstrate an understanding of, and ability to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings" (p. 8).

Hypothesis testing is conducted in order to analyze a claim about a population parameter, based on sample statistics. It involves formulating opposing statements—the null hypothesis and alternative hypothesis—about the population parameter of interest. The goal of hypothesis testing is to determine whether or not to support the original claim, based on whether we reject the null hypothesis. To do so, a sample statistic is measured or observed and converted to a standardized value called the test statistic. The test statistic is then used to calculate the probability, called the *p*-value, of obtaining a test statistic at least as extreme, under the assumption that the null hypothesis is true. If the *p*-value is too low, then we reject the null hypothesis. Once a decision is made, a conclusion can be formed about the claim.

With statistics education reform on the rise, as well as an increasing number of students enrolling in introductory statistics courses each year, there is a need for research that investigates students' understanding of hypothesis testing, a concept taught in almost every introductory statistics course (GAISE College Report ASA Revision Committee, 2016; Krishnan & Idris, 2015). While previous research in this area has focused on students' misconceptions pertaining to hypothesis testing, our study sought to turn attention to what students understand and how they come to understand it. We focus our attention on the following research question:

How do students reason about the concepts involved in one population hypothesis testing while working two problems involving real-world situations?

In this paper, we focus on answering this question for one particular student, Steve, who elaborated the most in his interview, and thus, provided us with the richest data.

Literature Review

Research has revealed that although students are able to perform the procedures surrounding hypothesis testing, they lack an understanding of the concepts and their use (Smith, 2008). Providing a survey of research on students' understanding of statistical concepts, Batanero et al. (1994) stated that hypothesis testing "is probably the most misunderstood, confused and abused of all statistical topics" (p. 541). Students appear to experience a "symbol shock" (Schuyten, 1990), which provides an obstacle for students interpreting particular questions (Dolor & Noll, 2015; Liu & Thompson, 2005; Vallecillos, 2000). Vallecillos (2000) found that students have trouble with not only the symbols, but also with the formal language and meaning behind the concepts involved in hypothesis testing, including words such as "null" and "alternative" when referring to the hypotheses. Students interviewed were not able to accurately describe what these terms mean and how they impact the decision to either fail to reject or reject the null hypothesis (Vallecillos, 2000). Williams (1997) made a similar observation. She found that, due to the tedious process behind hypothesis testing, students were not able to connect the statistical concepts back to the context of the problem. She further stated that, "the biggest hurdle is reaching a statistical conclusion, and the real meaning of the original question may be forgotten in the process" (p. 591).

Students' difficulty with understanding hypothesis testing can oftentimes be attributed to how it is taught. Textbooks and instructors frequently give a specific step-by-step script to follow when performing hypothesis testing, which does not provide students the opportunity to see the process as a whole. Link (2002) described this as a six-part procedure, which leads many students to look for keywords and phrases as guides when solving hypothesis testing problems. He found evidence that students were able to correctly substitute values into a formula selected from a formula sheet, but they did not have an understanding of the logic behind the overall procedure of hypothesis testing.

Method

The focus of our larger study is on university students who are enrolled in an introductory statistics course at a large public institution in the southeastern United States. For this particular institution, students were required to spend three academic hours per week in a computer lab, completing assignments through Pearson's *MyStatLab.* Data collection took place during Fall 2014 and Spring 2015. All students enrolled in six sections of an introductory statistics course (approximately 240 students) were invited to participate in a problem solving session and semi-structured interview pertaining to hypothesis testing. Twelve students volunteered to participate. During the problem solving session, each participant worked alone on two hypothesis test questions, similar to problems they had already seen. They were encouraged to use Excel when needed, since the use of it was required as part of the class. The first question asked the student to conduct and interpret a hypothesis test for a single population proportion. The second question asked the student to conduct and interpret a hypothesis test for a single population mean. The questions were as follows:

1. In a recent poll of 750 randomly selected adults, 588 said that it is morally wrong to not report all income on tax returns. Use a 0.05 significance level to test the claim that 70% of adults say that it is morally wrong to not report all income on tax returns. Use the *P*-value method. Use the normal distribution as an approximation of the binomial distribution.
2. Assume that a simple random sample has been selected from a normally distributed population and test the given claim. In a manual on how to have a number one song, it is stated that a song must be no longer than 210 seconds. A simple random sample of 40 current hit songs results in a mean length of 231.8 seconds and a standard deviation of 53.5 seconds. Use a 0.05 significance level to test the claim that the sample is from a population of songs with a mean greater than 210 seconds.

Immediately following the problem solving session, the students participated in a semi-structured interview that was video-recorded. There were ten interviews, eight with one participant each and two with two participants each. During the interviews, participants were asked to elaborate on their solutions and thought processes. Conducting the interviews was divided among five members of the research team, who all followed the same protocol. The data (interview transcriptions, written work, and Excel files) were analyzed and coded according to the levels of conceptions in APOS Theory (described below). The research team deliberated until an agreement was made regarding the codes.

APOS Theory

Action–Process–Object–Schema (APOS) Theory is a constructivist framework for describing how an individual might develop his or her understanding of a mathematical concept (Arnon et al., 2014). It emphasizes the construction of cognitive structures called Actions, Processes, and Objects, which make up a Schema. These structures are constructed through reflective abstraction, particularly through the mental mechanisms of interiorization, reversal, coordination, encapsulation, and generalization. The construction of these structures signify levels in the learning of a mathematical concept. An Action is a transformation of Objects in response to external cues. The primary characterization of an Action is the external cue, which could be keywords or a memorized procedure. Reflection on a repeated Action can lead to its interiorization to a Process. While an Action is an external transformation of Objects, a Process is an internal transformation of Objects that enables an individual to think about the transformation without actually performing it. Once a Process is conceived as a totality and the individual can perform transformations on it, the Process is said to have been encapsulated into an Object. While a component of APOS Theory is the development of a genetic decomposition, i.e., description of how an individual might develop an understanding of a mathematical concept, our genetic decomposition is omitted in this paper due to space limitations.

Results

While performing a hypothesis test, it is necessary for an individual to formulate the hypotheses about a population parameter, evaluate the test statistic, find the *p*-value, compare the *p*-value to the significance level, form a decision about the null hypothesis, and form a conclusion about the claim. Through these objectives, students construct mental structures called **hypotheses**, **test statistic**, **p-value**, **decision**, and **conclusion**, each of which can be conceived as an Action, Process, or Object. In this section, we provide examples of how the mental structures of **hypotheses**, **test statistic**, **p-value**, and **decision** emerged in the reasoning of one particular student, Steve. As we will show, these constructions emerged as Processes or Objects in Steve's reasoning. We use bold font when referring to the primary mental structures that make up our genetic decomposition, to distinguish them from other uses of these terms. For simplicity, we do not use a different font to distinguish between the different levels corresponding to a concept. Note that we are not seeking to classify Steve in terms of his understanding, but instead, present evidence we found of his reasoning. Due to space limitations, we omit discussing **conclusion**.

Hypotheses

The mental structure, **hypotheses**, can be conceived as a transformation—an Action or Process—that acts on the claim of the hypothesis test and returns the null and alternative hypotheses. As an Object, additional transformations can be performed on **hypotheses**. Steve exhibited both a Process conception and Object conception of hypotheses.

To illustrate Steve's reasoning of **hypotheses** as a Process, the following excerpt is considered from Question 1 of the instrument.

> Um, well, when you're doing null and alternative you always focus on the claim they give you. Um, so 70%, and just to make things easier, uh we do the null is equal to .7, and then the alternative would be whatever you're asking, in this case you're asking, is it 70%. So you use not equal to 70%.

Steve acknowledged, in general terms, that the claim is used to formulate the hypotheses. We consider this to be evidence of a Process conception of hypotheses.

To illustrate Steve's reasoning of **hypotheses** as an Object, the following excerpt is considered from Question 2 of the instrument.

> OK. I just did the same thing I did with proportion, and I said the null is equal to um 210, in this case, and uh the alternative is greater than 210. But the only reason I said that is because um in this bottom line of the question says, test the claim that the sample is from a population um with a mean greater than 210.

Steve used the phrase, "in this case," to indicate that in his mind he distinguished his procedure for Question 2 from his procedure for Question 1. Despite the fact that the questions on the instrument pertained to two different contexts, Steve said, "I just did the same thing I did with proportion." In order to be able to describe his procedures as the same, while also distinguishing between them in the different situations in which they arose, he had to have compared them, which is evidence of an Object conception of hypotheses.

Test Statistic

The mental structure, **test statistic**, can be conceived as a transformation—an Action or Process—that acts on various population parameters and sample statistics and returns a standardized value, namely the test statistic, which is the number of standard deviations a sample statistic is away from the distribution's center, or expected value. As an Object, additional transformations can be performed on **test statistic**. Steve exhibited both a Process conception and Object conception of test statistic.

To illustrate Steve's reasoning of **test statistic** as both a Process and an Object, the following excerpt is taken from Steve's discussion of Question 1, in which he described what accounted for an extreme value of the test statistic.

> But going back on it, it makes sense, you know, if you've got a *p*-hat that, that's very very different from your, from your *p*, you know, 78 is a whole 8% off of uh the 70%. And also your test statistic is very large. I'm not totally sure what a test stat is, but it reminds me of *z*-scores, and I remember when you have a *z*-score that gets above 3, it starts to get pretty, pretty crazy. So 5 is huge, which is also the reason that you're getting a bunch of zeros or very close to 1.

Steve appeared to have encapsulated into an Object the Process of calculating a *z*-score for proportions, in order to consider how it resulted in an extreme value of the test statistic. He explained that a large value of the test statistic resulted from having a value of the sample statistic that is very different from the value of the population parameter in the null hypothesis. APOS Theory acknowledges, in general, that it is necessary to de-encapsulate an Object back into a Process, which appears to be the case with Steve. That is, he de-encapsulated his test statistic Object back into a Process to consider the difference between $\hat{p}$ and *p*. We should note that based on Steve's statement, "I'm not totally sure what a test stat is, but it reminds me of *z*-scores," he appeared to have constructed isolated Processes for each test statistic, which he needed to further coordinate in order to construct a single test statistic Process.

P-value

The mental structure, **p-value**, can be conceived as a transformation—an Action or Process—that acts on the test statistic and returns a probability—a number between 0 and 1. As an Object, additional transformations can be performed on **p-value**. Steve exhibited both a Process conception and Object conception of p-value.

To illustrate Steve's reasoning of **p-value** as both a Process and an Object, we consider the following excerpt from Steve's discussion of the *p*-value for Question 1, in which he explained various procedures for calculating the *p*-value, depending on the situation.

> *Steve:* Well, whenever you're finding a *p*-value you're doing a .DIST function, and when you're doing proportions, it's NORM, and when you're doing means, it's T. So in this case we used NORM.S.DIST cause I think the other formula is silly. But uh since it's a two-tailed test I couldn't just stop there. I had to 1 minus that and then double it.
>
> *Interviewer:* OK, OK. And you did the 1 minus, why?
>
> *Steve:* Um because if you don't do 1 minus, it ends up being something very very close to 1. So a bunch of .9999…, and you can't double that. Whenever I got stumbled, I was like, oh wait, do I, uh, do I double the 1 minus or it by itself. Well, you can't go over 1. It has to be between 0 and 1.

Steve explained, in general, that an Excel .DIST function is used to calculate a *p*-value, and he said the result "has to be between 0 and 1." Steve's description in general terms of the transformation on the test statistic that resulted in the *p*-value and recognition of the p-value as a probability is evidence of a Process conception of *p*-value. Furthermore, Steve described situations in which you would use NORM.S.DIST versus T.DIST. Although Steve was not completely correct in stating that you always use T.DIST in the context of means, he clearly compared different procedures for calculating the *p*-value and considered situations in which these procedures would arise. Thus, we consider this to be evidence of an Object conception of *p*-value.

Decision

In hypothesis testing, we make a decision about whether or not to reject the null hypothesis by comparing the *p*-value to the significance level, which, in this course, was a predefined upper bound for the *p*-value. In particular, if the *p*-value is less than or equal to the significance level, we reject the null hypothesis. The mental structure, **decision**, can be conceived as a transformation—an Action or Process—that compares the *p*-value to the significance level and returns the decision about whether to reject the null hypothesis. In particular, **decision** compares the *p*-value and significance level as areas or probabilities. As an Object, additional transformations can be performed on **decision**. Steve exhibited a Process conception of decision.

To illustrate Steve's reasoning of **decision** Process, we first consider the following excerpt from Steve where he demonstrated that he compared the *p*-value and significance level as areas.

> Oh wait! Wasn't the *p*-value supposed to be from the edge? So wasn't the *p*-value supposed to be like this … [*draws on paper*] … the stuff on the outside? I remember now. It was um . . . I don't see how that relates to those, but I know it relates to the significance level 'cause your .05 is going to be outside of that.

Steve explained how he was able to graphically represent the *p*-value (see Figure 1). Finding that the *p*-value is less than the significance level, he drew the region whose area is the *p*-value inside the region whose area is the significance level, evidence that he compared the *p*-value and significance level as areas. To clarify, when Steve said, ".05 is going to be outside of that," we interpret it to mean that the rejection region is not strictly contained in the region whose area is the *p*-value. In addition to considering this to be evidence of a component of a **decision** Process, we also consider this as further evidence of a **p-value** Object.

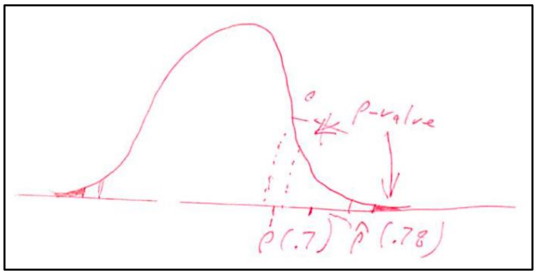

Figure 1: Steve's graph of the p-value for Question 1.

The previous excerpt established that Steve was able to compare the *p*-value and significance levels as areas, which we consider to be a necessary characterization of a **decision** Process. To further illustrate Steve's reasoning, we consider the following excerpt about whether or not to reject the null hypothesis for Question 1. Note that part of this excerpt was discussed previously in the section on test statistic.

> *Interviewer:* OK, so, and how did you arrive at your conclusion? What did you arrive at?
> *Steve:* I just remembered anytime the *p*-value is less than the, uh, significance level you reject the null, uh, I think [*laughs*]. But going back on it, it makes sense, you know, if

> you've got a *p*-hat that, that's very very different from your, from your *p*, you know, 78 is a whole 8% off of uh the 70%. And also your test statistic is very large. I'm not totally sure what a test stat is, but it reminds me of *z*-scores, and I remember when you have a *z*-score that gets above 3, it starts to get pretty, pretty crazy. So 5 is huge, which is also the reason that you're getting a bunch of zeros or very close to 1 […] So it's interesting, we always go all the way out to the *p*-value, but you can pretty much tell from your test statistic if it's correct or not.

Initially, Steve rejected the null hypothesis based on a memorized rule, suggestive of a **decision** Action. However, he reflected on this Action and related an extreme test statistic to a small *p*-value. As a result, Steve explained that depending on the magnitude of the test statistic, you could potentially form a decision about the null hypothesis without comparing the *p*-value to the significance level. The ability to describe the result of a transformation without needing to perform all of its steps is evidence of a Process conception.

Discussion and Concluding Remarks

Since the number of students enrolling in introductory statistics courses each year is continually increasing, it is important to explore students' reasoning of hypothesis testing (GAISE College Report ASA Revision Committee, 2016; Krishnan & Idris, 2015). This report, part of a larger study, focused on examples of how the mental structures of **hypotheses**, **test statistic**, **p-value**, and **decision** emerged in the reasoning of one particular student, Steve. Steve's constructions of the mental structures emerged as Processes or Objects in his reasoning. Steve exhibited a Process conception of hypotheses by acknowledging that, in general, the claim is used to formulate the hypotheses. In another situation, Steve exhibited an Object conception of hypotheses by being able to compare procedures for formulating hypotheses between two different problems. Steve illustrated test statistic as both a Process and an Object by describing what accounts for an extreme value of the test statistic in a situation. Steve exhibited an Object conception of *p*-value by being able to explain and compare various procedures for calculating the *p*-value, depending on the situation. Lastly, we found evidence that Steve illustrated a Process conception of decision by being able to describe the results of his decision without going through the steps of comparing the *p*-value to the significance level. In this case, he related a large test statistic to a small *p*-value.

Our results suggest that concepts involved in hypothesis testing are related through the construction of higher order transformations, operating on Processes that have been encapsulated into an Object. It has been widely recognized in APOS Theory literature that encapsulation of a Process into an Object is difficult to achieve, a possible explanation for why hypothesis testing is such a challenging topic for students. However, we found evidence of these constructions of higher order transformations in Steve's rich descriptions of the concepts.

With textbooks and instructors frequently introducing the topic by giving a step-by-step script to follow, what Link (2002) describes as a six-part procedure, construction of higher order transformations becomes even more difficult as this instruction leads students to look for keywords and phrases as guides when solving hypothesis testing problems. Based on the results, it is important when teaching to develop questions for students that motivate them to think and explain beyond a procedural approach. Creating activities with guiding questions will encourage students to think such as Steve, and to develop deeper knowledge of hypothesis testing.

References

Arnon, I., Cotrrill, J., Dubinsky, E., Oktac, A., Fuentes, S. R., Trigueros, M., & Weller, K. (2014). *APOS theory: A framework for research and curriculum development in mathematics education.* New York, NY: Springer.

Batanero, C., Godino, J. D., Vallecillos, A., Green. D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527–547.

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal, 14*(1), 60-89.

GAISE College Report. ASA Revision Committee. (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report.* http://www.amstat.org/education/gaise.

Krishnan, S. & Idris, N. (2015). An overview of students' learning problems in hypothesis testing. *Jurnal Pendidikan Malaysia, 40*(2), 193-196.

Link, W. C. (2002). *An examination of student mistakes in setting up hypothesis testing problems.* Louisiana-Mississippi Section of the Mathematical Association of America.

Lui, Y. & Thompson, P.W. (2005). Teachers' understanding of hypothesis testing. In S. Wilson (Ed.), *Proceedings of the twenty-seventh annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education*. Roanoke, VA: Virginia Polytechnic Institute and State University.

Schuyten, G. (1990). Statistical thinking in psychology and education. *ICOTS, 3*, 486-489.

Smith, T.M. (2008). *An investigation into student understanding of statistical hypothesis testing.* Doctoral Dissertation. University of Maryland.

Vallecillos, A. (2000). Understanding the logic of hypothesis testing amongst university students. *JMD, 21*, 101-123.

Williams, A.M. (1997). Students' understanding of hypothesis testing: The case of the significance concept. *MERGA, 20*, 585-591.