Tech Know Learn (2011) 16:199–220
DOI 10.1007/s10758-011-9187-9

Computing the Average Square: An Agent-Based Introduction to Aspects of Current Psychometric Practice

Walter M. Stroup · Thomas Hills · Guadalupe Carmona

Published online: 7 January 2012

Abstract This paper summarizes an approach to helping future educators to engage with key issues related to the application of measurement-related statistics to learning and teaching, especially in the contexts of science, mathematics, technology and engineering (STEM) education. The approach we outline has two major elements. First, students are asked to compute an "average square." Second, students work with an agent-based simulation that helps them to understand how aspects of the central limit theorem might be integrated into a much larger conversation about the appropriateness, or validity, of current psychometric practices. We are particularly interested in how such practices and interpretive frameworks inform the construction of high-stakes tests. In nearly all current high-stakes test development, tests are thought of as being built-up from individual items, each of which has known statistical properties. The activity sequence outlined in this paper helps future educators to understand the implications of this practice, and the sometimes problematic assumptions it entails. This instructional sequence has been used extensively as part of a core course in a university-based certification program in the United States (UTeach) recognized for its innovative approaches to developing a new generation of secondary STEM educators.

Keywords Statistics · Learning theory · Central limit theorem · Simulation · Conceptual change · Psychometrics · Item response theory

Over the past 20 years the idea of advancing scientifically based practices and policies in education has been marked by a renewed emphasis on the use of statistical inference, especially in relation to outcomes on high-stakes summative tests. This renewed emphasis is most evident in the United States. But there are similar efforts underway around the

W. M. Stroup (✉) · G. Carmona
Science and Math Education, The University of Texas at Austin, 1 University Station, D5705, Austin, TX 78712-0382, USA
e-mail: wstroup@mail.utexas.edu

T. Hills
University of Basel, Basel, Switzerland

world, with the support of international development organizations such as the Organization for Economic Cooperation and Development (OECD). During this same period computational tools and analytic approaches supported by computation have served to accelerate an ongoing shift in how high-stakes tests are developed. Most current test development centers on building-up a test as an aggregate of individual items, each of which has its own specific psychometric properties.

Test developers now can "pick and choose" from items intended to characterize learning outcomes or criteria and then, from the known profiles of how these individual items have functioned in the past, build expectations about how a test made up of these items will distribute student results. Actual results can then be viewed through the lenses provided by these expected results to make claims about student growth, content-specific learning, the quality of instruction, or even a host of broad educational policies pursued by schools, districts, regional governments or nations.

Given both the global reach as well as the high-stakes significance such item-based approaches have now attained, we find ourselves now siding with those who would want educators, researchers and, perhaps, all stakeholders in our formal educational systems to better understand the warrants and limitations of what had previously been treated as an almost esoteric kind of knowledge. In the past, statistical practices associated with item-based test development were to be understood by only a very select group of psychometricians. Most of the rest of humanity, it sometimes seemed, was expected to simply defer to their interpretations of the state of schooling, teaching and student learning.

As part of our efforts to address this asymmetry in interpretive warrant, we present an account of a series of activities we've used for many years in our undergraduate teaching to help students better understand, embrace, or possibly critique the many (often only tacit or implicit) assumptions that are made in these existing systems of accountability. Specifically, we present a task, or prompt, requiring students to compute an "average square" followed by a series of discussions of computer simulated results related to exploring implications of the central limit theorem. We use this sequence as a way into having our students—all of whom are undergraduate science, technology, engineering and mathematics (STEM) majors who are simultaneously completing a teaching certification program—explore connections between measurement ideas as they might typically encounter them in their undergraduate majors with how such ideas are extended to characterizing learning, teaching or ability in current educational practice.

While the use of agent-based simulations to explore ideas related to the use of formalisms like mean, hypothesis testing, and/or measures of significance is certainly not new (c.f. Wilensky 1993, 1995), we do believe our approach to some of these same ideas is novel. The novelty comes from foregrounding what our students have found to be somewhat surprising implications of the central limit theorem (CLT).

Prior agent-based introductions to statistical inference tend to build from a sense of the word "error" that implicates its entomological roots in the Latin word for "wander". Such approaches might typically use simulations where collections of agents [e.g., "turtles" created in the NetLogo programming language (Wilensky 1999)] start in one location on the computer screen and then "wander" out from this starting place based on fixed probabilities for taking a forward or backward step of a given size.

After running an agent-based model simulating this kind of 'wandering', students might readily observe how a binomial distribution with a central value—assuming symmetric probabilities and represented in terms of movement from an original location—emerges. The accumulation of individual forward and backward steps for the agents can be treated as a kind of model of the accumulation of errors associated with individual measurement

trials. The subsequent computing of a mean and standard deviation for the simulated distribution would reinforce a sense that there is, as is typically assumed in the natural sciences or engineering, something like the 'true' value which, for the simulation, could be seen as analogous to the 'original' or starting location of the agents before simulation is run (that is, before errors accumulate).

While such an approach is indeed likely to be useful for explorations of the binomial distributions and closely associated statistical formalisms, our experience has been that our STEM students can tend to over-generalized expectations of transparency and unproblematic utility for these approaches when they broach topics in learning and teaching STEM disciplines. This overgeneralization can then limit, or sometimes undermine, an appreciation of what it might mean to make empirically validated claims about learning, teaching, or the full range of possible trajectories for improving STEM education. Subsequent considerations of the logic or structure of what it means to know, learn, or teach STEM disciplines can then come to be organized in terms of metrics and expectations at least "once-removed" from attending to their students' actual understandings, expressions, or ideas about how the world works.

More than just providing an alternative way "into" the formalisms used in inferential statistics, the approach we summarize herein has helped us to address this tendency toward overgeneralization. The approach fits well with having our students—most with strong, if still developing, STEM backgrounds—come to appreciate some of the challenges, limitations, and nuances related to applying standard measurement theory to specific instances of knowing and learning. Our goal is to support students in engaging a set of pervasive issues about the fit between standard measurement theory, and the often contrasting range of possible representations of learning and teaching commonplace both in the STEM education research literature and in most teachers' day-to-day accounts of school-based practices (e.g., experienced teachers' accounts of who seems to understand what topics, how they seemed to understand these topics, and what might be pointed to as evidence for these evolving understandings).

To do this we wanted to go well beyond our own previous efforts where, in all candor, we found ourselves being far too willing to map prior student conceptions, say of "precision" and "accuracy", from contexts involving physical measurement (with which most of our students are familiar), onto what are often considered the near psychometric analogs, like "reliability" and "validity". What we outline herein emerged from our own critique of the kinds of interpretations we might have unwittingly advanced in our previous, barebones, "good enough" introduction to basic statistical ideas and interpretations that most educators might be expected to "know".

The dramatic ascendance of psychometrically defined accounts of what it would mean for educational practices to be "scientifically based" played a significant role in pushing us to do much better than we had in past. As our efforts progressed, we found ourselves moving more and more toward helping students see the possible *contrasts* between standard measurement-related practices with which they might be familiar, especially from the relatively strong backgrounds many of them had in natural sciences, and what it might mean to characterize the state of learners' evolving understandings of STEM-related topics.

What follows, then, is our account of the elements of an alternative approach. A measurement-like or seemingly Gaussian distribution is produced in this sequence of activities but it is produced in a way that is quite distinct from any link to a model of wandering out from an initial starting place or 'true' value. As part of this approach we use a computer simulation to illustrate how this normal distribution can emerge in ways

consistent with the central limit theorem, yet not at all consistent with the tacit sense many of our STEM majors may want to impose for how such distributions relate to accumulated error around a "true", or measured, value on results from educational settings.

Despite what the students do see as the somewhat odd origins of the normal distribution generated from their responses to the average square activity, the distribution that emerges from the central limit theorem simulation can be treated as being mathematically equivalent to that which would emerge from a wandering-based simulation.

In recognition of this convergence, the focus of this paper is on the (sometimes only tacit) *assumptions that are typically made prior to computing various measures of central tendency*. For those who might be interested, we do include a link to resources and extensions in the use of the simulation that have proven useful when we've taught courses where the goal was to more thoroughly address many of the standard topics typically covered in a standard introductory statistics course (Hills and Stroup 2008). With this paper, however, our focus is much more on a set of foundational issues STEM educators (and others) might need to understand in an era when high stakes testing has come to be so prominent.

To complete this initial overview, we should also note how the structure of this paper relates to the actual pedagogy we use in our classes. In our classes our approach is intended to be more dialogic and is typically organized in terms of addressing issues as they might emerge. To improve readability, however, we've attempted to highlight some of the ideas and issues that often do occur but in a more thematic manner than might actually be experienced in any given teaching episode. Additionally, each component is developed somewhat more completely than might be fully realized in any particular classroom discussion. The warrant for this redaction comes from our hope that the readers might better be able to think through some of the issues either on their own, or as part of preparing for the kinds of discussions that might arise in their classes.

Contrasting somewhat with the trajectory we have used in an introductory statistics course, the version of the activity sequence outlined below is closest to how it is presented in a course titled *Knowing and Learning in Science, Technology, Engineering and Mathematics*. This specific course is a core requirement in the UTeach Natural Sciences and Engineering program that originated at the University of Texas at Austin in the United States. This combined STEM-major and teacher certification program, including versions of the *Knowing and Learning* course, is now being scaled to more than twenty research universities in the United States.

For readers interested in better understanding how this activity sequence is related to the broader purposes of this course, we have included some very brief discussions of what can be treated as closely related cognitive theory. These should be taken only as cursory suggestions or sets of pointers to the kinds of links that might be made between issues raised by the activities and a more in-depth engagement with STEM-related learning research. Similarly, a brief historical sketch with some background references related to the development of statistical inferences is included to help the reader frame aspects of average square activity and related discussions within the broader introductory trajectory of the *Knowing and Learning* course.

1 A Brief History of Statistics and Related Measurement Practices in Natural Science

As an historical matter, it is unclear when, exactly, the statistics we now associate with measurement and inference came into being. Assignment of credible points of origin seems

to depend, in part, on what components of statistical analyses, or elements of probability theory, one sees as decisive. If, for example, statistical inference is to be distinguished from syllogistic proof, then we can point to the distinctions the ancient Greeks made between "chance" and "necessary" causes. If specific computational procedures are to be seen as definitive, origins can be found in the more recent, and sometimes-colorful, history surrounding the development of these procedures. The "t test", for example, is said to have developed in the 19th century from attempts to improve the quality of beer production. Finally, the interactions of developments in statistical reasoning with developments in formal probability theory serve both to complement, and to complicate, any detailed account of their interwoven histories.

Through all these difficulties, what we can say with at least some confidence is that the close association of specific statistical procedures with acts of measurement extends back to some of the earliest days of empirical scientific practice. These include both reports of means and measures of error. When asked, "What does a computed mean represent?" and "Why would we need measures of error?" our students' responses often fit reasonably well with some elements of the earliest application of statistical ideas, especially as associated with the development of astronomy.

Early astronomers, we are told, needed systems for discussing overall findings as well as variations occurring in measurements of specific celestial objects (Heidelberger 1987). To this end, a notion of what eventually became the "standard deviation" was developed to work alongside the computing of a mean. If the standard deviation for a particular set of measurements was relatively large, then other astronomers would know that a good deal of fluctuation had occurred in the measured observations. A low standard deviation meant that the observations were relatively consistent. The standard deviation was formalized in the nineteenth century by Karl Friedrich Gauss (Swijtink 1987), who gave it the numerical interpretation that it has today (i.e., allowing that the results are consistent with the central limit theorem, within one standard deviation of the mean one can expect to find nearly 68% of the trials in a normal distribution).

In the realm of celestial observations—where the focus is on objects like planets, comets, or satellites—most of us would see it as entirely reasonable to assume that there is some one thing being examined (assuming a certain level of competence) and that this same one thing continues to exist while observations are made. The methodological movement in such contexts is *from* a real object, like a planet, *to* the use of statistics.

In other contexts, however, similar statistical language may be used even when it is much less clear what is being referred to or denoted. Indeed, in some cases the trajectory of the analysis seems to be largely inverted. Rather than moving *from* a real object being observed (e.g., a planet) *to* the use of a certain kind of statistics, the analyses move *from* the use of a certain kind of statistics *to* the assumption that there is something *there* that is being measured. In such cases, what we call a kind of *operationally defined* or "*heuristic*" *realism* takes hold (Stroup 1994, 1996; Stroup and Wilensky 2000). This sense of heuristic realism is simply meant to refer to those cases where statistical artifacts (e.g., a computed average) seem to function as, or *become*, a justification for an assignment of significance, or object-like status, to specific values or depictions—even, as we'll discuss below, when such depictions may *differ in kind* from what is actually observed. Having students wrestle with *if, when, or how* this kind of inferential inversion might make sense, or be seen as reasonable, is part of what motivated the creation of the average square task and subsequent analyses we outline.

To help clarify what is at stake, including the kinds of judgments that need to be made by the investigator, we have found it useful to consider a few specific examples. The

examples are ordered in a way intended to move from a case students see as relatively unproblematic to cases the students might see as increasingly problematic. Each of these examples is an instance of where a value can be computed and where this computed value might then become a candidate for being treated as if it referred to something real or object-like:

(1) the average family size in the United States,
(2) the average ant size in a multi-caste colony (where roles or "castes" within a colony of ants can have distinct sizes, resulting in a distribution of sizes that is multi-modal)
(3) the average of ways of thinking.

We'd suggest that in the series (1) to (3), it becomes less and less obvious, as we move from family size to ways of thinking, exactly what the average is representing. All three can raise questions related to if, when, and how such averages might be credible, correct, or useful. We attempt to highlight the issues by asking questions like: "What is being referred to by the averages in each of these cases?" and, "Should averaging always, or automatically, carry with it measurement-related connotations and implications?"

After highlighting a few elements in the historical development of statistical inference and after considering these three (or similar) cases of where a computed average is taken as referring to something object-like about a set of measured values, our sequence of activities then turns to the core "average square" activity to further explore specific ideas, or concerns, relating to the application of measurement theory to acts of knowing, learning or teaching in STEM-related disciplines.

2 Computing the "Average Square"

The prompt we use with students—asking them to compute the "average square" for two squares having sides of varying length—was initially proposed by Uri Wilensky and has served elsewhere as a kind of thought experiment related to averaging (Stroup and Wilensky 2000). Our use herein, however, is quite distinct in that we explore the utility of asking students to actually compute a specific value for this prompt.

The prompt has a number of useful features. Like other instances of where the same data, input, or prompt can be seen or understood in very different ways—we can think of the varying interpretations of ambiguous images along the lines of Wittgenstein's famous duck-rabbit sketch (1953/2001, Part II, §xi)—the average square task readily generates multiple, comparably credible, responses. In addition, the task has the advantage of requiring the students to generate, not simply select from among, possible numerical values. Each of these values is produced in relation to distinct interpretations that can be considered, broadly speaking, analogous to our either seeing "the rabbit" or seeing "the duck" in Wittgenstein's ambiguous sketch.

The fact that groups of distinct values are produced by comparably plausible, and yet quite distinct, ways of reasoning about the average-square task allows us to link much more directly to issues surrounding the psychometric interpretations of results from items on high stakes assessments. At least initially, the association of numerical values with distinct forms of reasoning is not unlike the assignment of numerical values to dichotomous results (usually "0" and "1") on individual, high-stakes, test items. Again, for the average square task the students actually generate the values associated with their forms of reasoning about the task. For large-scale test development, however, the assignment of numerical value is much more the responsibility of the publisher and can often be driven as much by

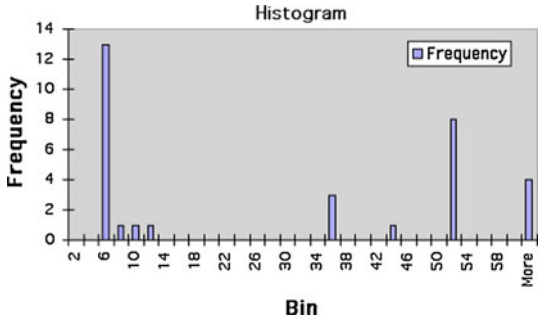

Fig. 1 Distribution of computed values for the average square

psychometric considerations as by any specific engagement with students' scientific or mathematical reasoning about the task. Issues associated with assigning metric, or fully quantitative, interpretations to what might more naturally be treated as distinct forms of reasoning, are what the average-square activity is intended to help make visible to our students.

In practice, on their way into class, we ask our students use a simple web-based form[1] to submit responses to the following prompt:

- Consider two squares: Square A with a side of 2 units and Square B with a side of 10 units. Create a way of computing the average square for these two squares. Enter your value (a number).
- Explain how you computed this number for your answer (text).

There are a significant number of ways of computing an average. For instance, responses from students can include computing the average side, the average area, the square-root of the average area (to 'correct' for dimensionality), the average perimeter, or the average diagonal. Some students have even computed the geometric mean.

Being told they need to share their results and explanations can heighten concern about there being "more than one" possible answer. Students can sometimes hesitate in submitting their responses precisely because they can create alternative responses. Once we emphasize that the question is not intended to have one "right" answer, the students tend to be more forthcoming and will submit their numerical responses and accompanying accounts of how they arrived at their respective values.

3 Illustrative Class Results

The results shown in Fig. 1 and Table 1 come from a group of our undergraduates who submitted their numerical responses and explanations to the average square question. When displayed, students immediately notice that the graph shown in Fig. 1 is multi-modal.

In the subsequent discussions, the students often appreciate that each of these "modes" is associated with particular forms of reasoning about the average-square task. They recognize that the distribution is structured by distinct kinds of reasoning such that if the students look at the explanations associated with a narrow range of values (one of the

[1] Of course other similar network tools can be used. We've also found the activity works well with students making a white-board sized histogram using adhesive notepads (e.g. Post-It™ Notes) with the value written on the front-facing side and the computing method written on the back.

Table 1 Student explanations of computed values

Explain how you computed this number for your answer
* I added 2–10 and divided by two. This would produce the average size of a side of the average square. If you wanted the average area I would take (4 + 100)/2. However I assumed that you wanted the average square in terms of an average size side unit instead of an average total area since you mentioned sides more than total area
* I found the square for each square then just averaged these two numbers
* I took the number of units and added them together to get the number 12. Since there are two squares for the twelve units I divided the number by 2 to give me 6. This answer would be a common number for each square to have based on the number of present units
* 2^2 + 10^2 = 104 104/2 = 52 (average of the area of both squares)
* Well, 52 is an average of the areas of the 2 squares, and the square root of 52 is the side length of this square
* A + B = 12 then I divided by 2
* I added side units and divided by the number of squares in the set
* If you just talking about the side length, I just added the two sides together and divided by two. But if you want the average area or volume or something that will take a little more work. What do you mean by average square?
* I calculated the square value for each square shape first. Then added them together and divided the sum by two
* The sides of Square A are 2 units. The sides of Square B are 10 units. If you take the average of the lengths you get 6. That is the length of the average square
* I squared each term (2^2 = 4, 10^2 = 100). From there I added the numbers (=104). Then I took the square root (104^(1/2)) = 10.2
* I took the area of the two squares and got the average and got 52
* Average square = (A^2 + B^2)/2
* I took the average number between 2 and 10, which is 6, and squared it, which is 36
* I added the side of Square A to the side of square B (=12) then I divided by two to get the average and then I squared the 6 to get 36
* I took 2 + 10 = 12 and divided that by 2 and got 6. I think that is how to take the average
* 2 + 10 = 12/2 = 6^2 = 36 100 + 4 = 104/2 = 52 36 + 52 = 88/2 = 44
* 6 is the value between 2 and 10 and I simply squared that value
* Added 2 and 10, then divided by 2
* I squared 2 (multiplied 2 times 2) and I got 4. Then I squared 10 (10 times 10) and I got 100. When you add the two up (4 + 100) then you get 104. Half of 100 is 50 and half of 4 is 2, so 50 + 2 = 52 (which is half of 104)
* I took the average of 2 and 10
* It's an approximate value for the square root of 104/2. I computed the areas for the 2 squares and found the average of them, and then I took the square root to find the length of that average area's side. I was also wanted to compute the geometric mean of 2 and 10, but I don't know if that related to my thought of squareness as a measure of area
* I added the two numbers and divided by two (amount of squares involved). This gave me an average value
* I added the two sides and took the average. It is 6, so the side of the average square is 6
* I took the square of each square. Then add each and divide that number by two
* [(2^3) + (10^3)]/2
* I found the area of both squares, which equals 104. Then I divided that answer by two to get the average of both areas
* I squared both of them, divided by two, and took the square root of the answer
* 2 squared plus 10 squared =104 divided by 2 = 52 for the avg. square
* 2 + 10 = 12 and then divide it by 2

modes), the explanations are similar (see Table 1). To highlight this sense, we've found it is sometimes useful to ask students to guess what procedures might have been used in columns of responses in the histogram that *don't* include their own responses. Then we can ask someone whose response is included in one of these other modes if the account is accurate. Usually it is.

We can highlight that the prompt, much like the duck-rabbit sketch used by Wittgenstein, has some claim to starting off as being 'objectively' the same for all of participants. Just as the duck or rabbit interpretations both come from looking at the same line drawing, the distinct ways of computing the average were produced in response to the same average-square prompt. What becomes particularly important and relevant—certainly for our students' future lives as educators—are the distinct, often consistent, and usually shared interpretations that structure the varied responses to the given task. Similar values (modes on the histogram) are almost always structured by similar modes of thinking about the task, and these forms of reasoning can be analyzed and discussed (see Table 1). Alternative responses to STEM-related tasks are usually not simply haphazard guesses but are actively structured by forms of reasoning that can be explicated, explored, extended or changed over time.

We then can ask: What might happen to the distribution if someone in the class was successful in making a reasonably compelling case for *one* of these modes of reasoning about the task? Students will point out that if this is the case, the populations associated with some of the other modes would diminish and relocate, discontinuously, to the mode associated with what is now seen as a more compelling way of approaching the task. Participants would "shift", or relocate, their modes of reasoning, or even develop new modes of reasoning, in ways not arrived at by progressive incremental movement (monotonic sliding) from one understanding to another (Stroup 1994, 1996; Stroup and Wilensky 2000).

These *shifts in kinds of reasoning* can contrast with an assumption in standard measurement theory that one kind of thing or object is being attended to in a way that, returning to the prior discussion, might function analogously to a planet. When measuring the location of a planet it makes sense to treat the measurements as being about the same object. With forms of reasoning, however, it is much less clear in what senses we might be talking about the same thing (even if they are all responses to the same prompt).

As we will return to later in this paper, attending to these kinds of shifts in reasoning has a long history within cognitive learning theory (c.f., Piaget 1970; Bruner et al. 1956) and serves as the basis for an extensive literature on conceptual change (c.f., Posner et al. 1982). As is true for the average square question, understandings can be projected onto a particular axis (e.g. an axis of values computed for the average square in Fig. 1) but the multiple understandings themselves might better be thought of as being located in a much more complex conceptual *n*-space represented by specific patterns of interaction with the information given and what the students understand (Stroup 1994, 1996; Hills and Stroup 2004).

For much of current cognitive theory, shifts in kinds of understandings are what *characterize* what we mean by learning (e.g., assimilation for Piaget). Teaching, it follows, is about supporting and extending these shifts in kind. Indeed, elsewhere it has been argued that these multimodal distributions with relocation form the basis for developing what might be called a "cognitive statistics" (Stroup 1994, 1996; Stroup and Wilensky 2000). A more thorough engagement with the specific features of cognitive statistics takes us beyond the scope of this paper and so we now return to the specifics of connecting this activity to issues surrounding the application of measurement theory to forms of reasoning.

4 Initial Reasoning About the Average

To begin to move the conversation back to exploring the warrant of standard psychometrics, we now ask the students: How might we characterize the results from the class as a whole? What overall statistic or statistical representation might be useful at this point?

Computing an average or mean is so commonplace in many measurement settings from natural science or engineering that it often comes to students' minds as an appropriate "next step." This tendency is especially strong, we have found, among students who either have taken a formal statistics classes or have some background using standard measurement theory and/or error analysis in lab settings. Separate from the issue of whether computing an average is useful or appropriate, all our students understand that it is certainly possible to compute an average (mean) value for the data in a multimodal distribution.

5 The Reasonableness of Computing the Average

When asked to then reflect on the reasonableness of computing the average as a representation of the results for the class as a whole for the average square task, our experience has been that some students will raise the question, "Why would you want to do that?" and go on to point out, "No one is there."

What these students seem to notice is that for a distribution like that shown in Fig. 1, the average value for the data is approximately 34 and, unlike the existing modes already present in the data (see Fig. 1, Table 1), the value for the average of all the responses *has no form of reasoning associated with it*. The value isn't seen to point to, or even get particularly close to,[2] any of the actual forms of thinking present in the population of results. Instead, the very distinct *forms of reasoning* that we actually observe in the data are seemingly misrepresented by computing a value that doesn't correspond to, or refer to, anyone's thinking.

In considering alternatives to the average for representing the results for the class as whole, some students will argue that choosing a value associated with one of the particular modes of reasoning, even if it represents only a subset of the results, is preferable to computing a mean value that fails to represent anyone's thinking. As a representation of the class results, sentences of the form, "A plurality of us thought about the task this way…," are seen as preferable to sentences of the form, "The mean value for the class is 34."

This initial sensitivity to the phenomenology of the data exhibited by at least some of the students does comprise, in our view, a nascent sense of the kind of insight we associate with higher, or at least more nuanced, levels of statistical reasoning. Students can begin to address a common "misconception" about computed averages: That an average represents the most common result, "the value that occurs more often than the others" (Garfield 2002). No one actually computed the value associated with the average of 34, and so the mean certainly does not represent the most common result. In the context of the average square activity, the distinction between an average value and the most common value is particularly salient. In a similar way, the computed average is not necessarily closer to—or a better representation of—the true value of whatever quantity it is we are investigating.

[2] There is not a sense in which this shortcoming in the use of this average might be appreciably improved by having a larger sample or by carrying out many more trials.

To further explore the issues surrounding the use of an average to represent a population, we can raise related examples like: "No one has seen the average family with, say, 2.4 children … so in what sense might we be justified in using the average to talk about family size?" Students may start to note that the validity of a statistical average requires a question that knowing the average somehow answers. For example, do South Asian families have more children *on average* than South American families? Do hawks have longer bills *on average* than sparrows? An average, by itself, has no "real" significance. On its own, it is not *the* most "objective" or "scientific" way of representing data. Instead, computing an average needs to be seen as situated relative to a particular line of inquiry.

In a similar vein, other subtle issues can be broached in relation to computing the average. For example, it is possible that the average family size in a city of extremes is higher than in another, more homogeneous city, but that the typical (mode) family size in Extreme City may actually be smaller than in Typical City. Or we can begin to address the tendency of our students to over generalize the expectation that the mean, median and mode will converge.

We also have found it useful to revisit contexts where multi-modal distributions, like those generated in response to the average square question, occur in a wide variety of biological and social contexts. As was introduced earlier, in biology the measurement of ant sizes in a typical ant colony reveals that these sizes are distributed in distinct modes associated with different castes (Holldobler and Wilson 1990). Different castes have different jobs in an ant colony and the number of castes is often representative of the complexity and capability of the colony. Similarly, if one measures hair length in a human population, the modes are commonly distinguished by gender (females generally have longer hair than males, and this can result in a bimodal distribution). Family sizes come in distinct integer quantities.

Although alternative representations or characterizations of these distributions may be more complicated than simply computing an average, students may recognize the need to link specific representations to specific purposes and contexts. Student can be asked, "How can we attempt to characterize, or say something meaningfully 'representative' about, a multi-modal distribution?" For example, in order for a researcher to be able to make a practical judgment whether he/she had a distribution that was like another multi-modal distribution reported in the literature, what would he/she need to know? What would we need to tell her or him? Or, closely related, how could a fellow researcher decide if a given distribution was distinctive enough to be considered unique or exceptional?

As part of a further exploration of these and related questions, we now move to the use of the central limit theorem simulation environment. The purpose of working in this environment is to deepen—through exploration and possible extension—our students' appreciation of the power and subtlety of the central limit theorem especially as associated with the application of inferential statistics.

6 The Central Limit Theorem Simulation of the Average Square

Among the questions we keep returning to throughout these activities are these: "When might the interpretation of an instance of averaging as an act of measurement be warranted?" and "When, especially when it comes to characterizing forms of human understanding and insight, such as they are likely to encounter in their classrooms, might such an interpretation be unwarranted or even misleading?" We developed this simulation to help explore the senses in which implications related to the central limit theorem might be seen

to provide a certain kind of warrant for some existing practices. Implications following from the CLT might be seen to wrap, or plausibly surround, a computed average with the trappings of a normal distribution. Using the simulation we can begin to explore how is it that we might find ourselves sliding from computing an average square or, in what follows, an average way of computing the average square, to then developing a sense that such acts are (or should be) considered measurements.

Formally, the central limit theorem implies that if one takes multiple samples, each of size, *N,* from any population and finds the mean of each of these samples, then *the distribution of these sample means* will be *approximately normal* with the same average as that of the underlying ("true") population (c.f., Pelosi and Sandifer 2003). In other words, if we allow our simulation to take a reasonably sized sample of the "average square" solutions and average these results, and then have it take another sample and averages those results, and we continue doing this and then plot the respective averages, this plot will eventually form a normal distribution.

This emergence of an approximately normal distribution, it should be noted, will happen *even if the initial population distribution is multi-modal*. Indeed the CLT simulation (Hills and Stroup 2008) was developed to illustrate this implication using our multi-modal average square results (Fig. 1). The question then becomes, what does this normal distribution represent?

Because each computed average for sets of data from classes completing the average square task is likely to point to a value not selected by any student, for many such samples it is similarly unlikely that the emergent normal distribution will be centered at a value that represents the reasoning of *any* of the students in the original population. This is not an issue of an average being within some "margin of error" relative to particular values, but instead points to the likely event that the normal distribution will emerge at a location significantly removed from any overlap with actual values computed by students.

This reservation notwithstanding, in other contexts where numerical data is generated the emergence of a normal distribution might support the use of inferential statistics. Most fields that use quantitative information, including psychology, biology, physics, sociology, education, business, medicine, economics, geology, and engineering, use these ideas to surround the computing of an average with a consistent interpretive framework. Our point is only that some contexts might more credibly be viewed as instances of measurement than others. This credibility need not be ascribed in advance, based solely on the shared use of technical terms, well-established metrics or computational procedures, but might better be judged in terms of consistency and utility in serving particular lines of inquiry.

To more clearly illustrate aspects of how the central limit theorem can be applied to the average square responses, we have chosen to focus on just two of the most frequently occurring values (6 and 52) from the average square activity in our central limit theorem model (written in NetLogo). The model takes as inputs the relative portion of these two values (the ratio of 6's to 52) and generates an overall population of independent agents consistent with this relative ratio. The size of the population of simulated students is determined by another variable input value. This population value can be taken to represent all possible, or potential, students who, out over a number of semesters or years, might take our class (see the shaded circles in Fig. 2).

In our simulation, the individual agents representing the two most common modes of thinking about the average square task are discriminated by both color and size. These features, however, play no role in the subsequent sampling processes. Figure 2 shows the population both as circles and as represented in a histogram.

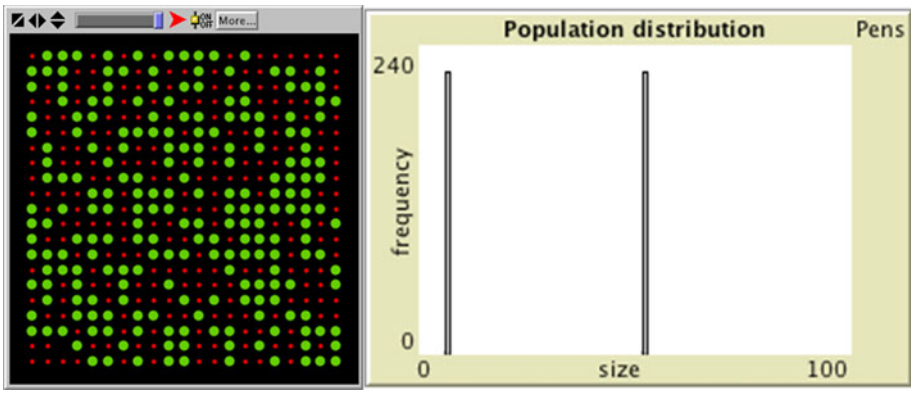

Fig. 2 Histogram and graphic of groups with two different computed values for the average square (two different *shades* and sizes shown)

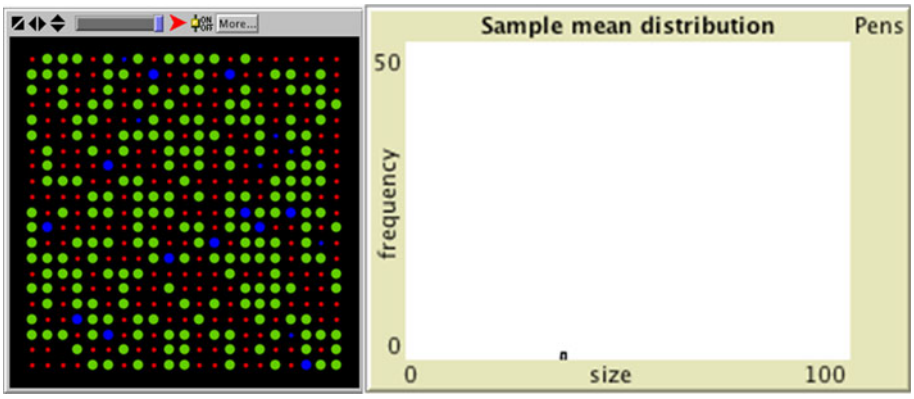

Fig. 3 Average value for a sampling (*dark shaded* agents). The average value for this sample is plotted on the histogram on the *right*

Of course for the purposes of exploring the emergence of a normal distribution, the relative proportions one chooses and/or whether one chooses to alter the code to add other modes to the mix of values in the population as a whole is somewhat arbitrary.

A particular group of students (i.e., a class or set of classes) in a given semester would represent a sampling from the total population of students who might, over time, take the course. This sampling from the total population is shown by the agents changing color (Fig. 3).

Additional samplings can be made (with replacement[3]) and the average of the average-square value is computed for each of these samplings. Results for a second sampling (using data for another simulated class or set of classes) is show in Fig. 4.

[3] We set aside for now issues related to replacement. Humorously, one of our student suggested this aspect of the model might represent students who would freely elect to retake the course, "just for the fun of it," in subsequent semesters.

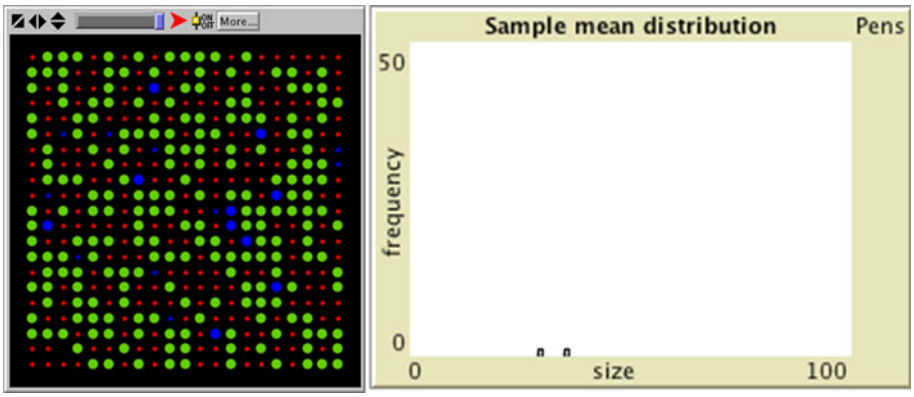

Fig. 4 A second sample (*left*) and average value for this sample (*right*) is shown

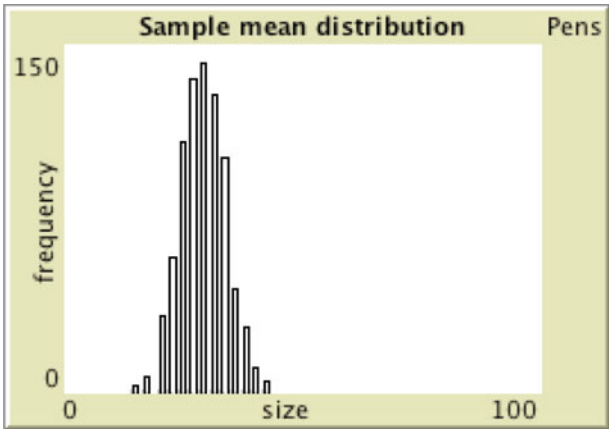

Fig. 5 Repeated sampling reveals a normal distribution. This figure represents 800 samples of size 20

Note that the value for the second average also occurs between the modes of observed forms of thinking (6 and 52). With repeated samplings the histogram develops and a pattern emerges in a way that is often surprising to students.

As is predicted by the central limit theorem, over many samplings a standard statistical "bump" will emerge and this distribution can be taken as approximating the normal distribution (Fig. 5). The general shape of this distribution for our multimodal populations is indistinguishable from the shape of a distribution that might arise in sampling a symmetric, normally distributed, population. The presence or absence of modes in the original dataset is "lost" in a way consistent with the CLT, but that can still seem remarkable to those more accustomed to sampling from populations they assume are normally distributed or who may never have used this kind of simulation with multi-modal datasets.

This emergent statistical artifact, because it looks "just like" what one might expect from a series of measurements, can begin to be interpreted as if some "thing" is being measured. If seeing is believing, now students can "see" and talk about a (seemingly) measured (average) value, surrounded by a normal distribution like what might be expected for a set of standard measurements in a lab. Of course this value has the unusual

property of not actually existing in (or, in many cases, not even being close to values found in) the data.[4]

This lack of a referent (in this case, an identifiable form of reasoning about the task) doesn't diminish the possible use of the central limit theorem to help impose a consistent interpretive scheme around this average value. Whatever shortcomings there might be to modeling human reasoning with averages or to using similar operationally, or computationally, defined metrics, these cannot be "blamed" on the central limit theorem. Formal interpretive schemes, like those associated with the central limit theorem, can be brought to bare on data that does not require, or that might not even support, the responsible use of these interpretive schemes. A judgment needs to be made regarding the fit between phenomena and formalism.

What understanding aspects of formal statistical reasoning can do is advance a more meaningful and reflective perspective on when to apply statistics ideas and constructs and what kinds of phenomena statistics might describe (Reid and Petocz 2002). Certainly inferential statistics, based on implications of the central limit theorem, are valuable in many kinds of comparative studies. This we do believe. We also believe students can better appreciate when and how inferential statistics can "make sense" as a tool of inquiry, if they are given an opportunity to explore for themselves how aspects of central limit theorem do form the basis for many of the core constructs of inferential statistics.

We now move on to the kinds of follow-up discussions and literature overview we've found helpful in supporting our students' engagement with the relationships between standard psychometrics and STEM-related teaching and learning.

7 Computing the Average Square and Some Examples from Recent Cognitive Theory

Most cognitive theory rests on the assumption that we can build models of mind. At stake in learning research is this: What kinds of models of mind will typify a good "fit" with the phenomenology of human insight? Are we to presume, in the end, that representations of cognitive activity in terms of computed averages are to stand as the best, most objective, or most significant rendering of cognition? Or is there a significant sense in which a move to think about reasoning in terms of averages so fundamentally misrepresents the phenomena at hand that the inclination to compute averages should be seen as deeply suspect in our efforts to generate empirically based models of mind? The issue is not simply whether averaging *subtracts* from or reduces the amount of plausibly salient information about our thinking that is available for analysis and response in our teaching. We can also ask whether averaging *adds* the expectation that a computed average points to something meaningful or relevant, and then surrounds this expectation with a set of associated interpretive metrics that may, in the end, fundamentally misrepresent the phenomena itself—e.g., the modes of reasoning used to support the computing of the average square.

Similar to the discussions we use in our classes, we now briefly explore these issues and questions relative to what might be seen as two major "schools" of cognitive research and the kinds of data each school sees as significant in informing both research and educational activity. As we have said elsewhere, these are only our suggestions for where a conversation might go relative to two important frameworks for analyzing learning, and should not be construed as a prescriptive sequence. The two schools we highlight—selected in part

[4] In the history of intelligence research, the ascription of significance to the emergence of distinct "bell curves" has given rise to some rather incendiary racist claims (c.f., Herrnstein and Murray (1994).

because of their established standing in analyzing teaching and learning in school settings—are the novice-expert framework that has come to dominate traditions associated with the still nascent learning sciences and the somewhat longer-standing positions associated with constructivism that are seen to have their roots in the writings of Piaget, Vygotsky or, prior to this, the positions associated with pragmatism as a philosophical movement (not to be confused with philosophical utilitarianism).

If we consider the novice-expert theories of knowing and learning (e.g., Bransford et al. 2000, especially Chap. 2), we can begin to see the ways in which it is likely that standard statistical procedures will fail to meaningfully render the phenomena under consideration. Novice-expert theories of cognition suggest that individuals map knowledge through conceptual associations. The difference between a novice and an expert can be largely understood in terms of the structure of the conceptual architecture. This architecture itself can be represented using concept maps (conceptual nodes with connectors). These maps are taken to be meaningful representations of either expert or novice understandings and should be engaged as such. Moreover, an engagement with these maps can be used to inform educational practice (Bransford et al. 2000; Novak and Gowin 1984).

As models of cognition, these concept maps are not readily adaptable to standard statistical protocols. While efforts to quantify the associations found in concept maps do exist in the research literature (Stuart 1985; Nicoll et al. 2001), creating such metrics with the intent to make standard statistical comparisons runs the risk of significantly distorting the phenomena under consideration (much like the computing of the average of the average square). Students who exhibited skepticism in the earlier discussions regarding the urge to assign significance to the computing of the average of the average square can be expected to be just as likely to express skepticism here.

The use of concept maps to analyze human insight is situated relative to these maps' ability to help answer specific questions about the nature and development of ability within the novice-expert framework. The point is not whether or not a concept map is a correct or adequate representation of mind. This is a separate question to be considered on its own terms relative to assessing the warrant of the novice-expert stance itself. What is being questioned is the urge to create an average-based quantitative representation of the insights expressed using concept maps. Like the understandings they represent, the significance of concept maps may be misrepresented by an effort to create and compare computed averages related to these maps. We suspect that any notion of average concept-map-ness for either novice or expert understandings is likely to be misleading in many of the ways that computing an average for the average square is potentially misleading.

The phenomenology and significance of concept maps may be missed or distorted by the computing of an average with its attendant implication that there must be a referent, or something of significance, to which the average refers. The multi-dimensionality of different kinds of reasoning that are likely to be present for a given person or group of students in a classroom is collapsed to a value that is an artifact of the mathematical machinery brought to the task of averaging. Relative to a particular area of investigation, this value is not likely to represent the relational significance of the nodes and connectors as they are found in the data of the concept maps themselves.

Similar concerns can be discussed with students relative to the kinds of data taken as meaningful in yet another major branch of cognitive research. Constructivist researchers and educators attend to the changes in the manifest forms of reasoning that occur both during intellectual development *writ* large, and during episodes within specific learning contexts (c.f., Piaget 1967; Duckworth 1987; Bandura 1986; Vygotsky 1962). While constructivist research can be considered quite distinct in its histories and motivations from

The accompanying figure shows a ball thrown vertically upwards from point A. The ball reaches a point higher than C. B is a point halfway between A and C (i.e., AB = BC). Ignoring air resistance,

On its way up, what force(s) act on the ball?

(a) Its weight, vertically downward.
(b) A force that maintains the motion, vertically upward.
(c) The downward weight and a constant upward force.
(d) The downward weight and a decreasing upward force.
(e) An upward force, first acting alone on the ball from point A to a certain higher point, beyond which the downward weight starts acting on the ball.

(Halloun & Hestenes, 1985)

Fig. 6 A conceptual question related to Newton's Laws in physics

that of the novice-expert frameworks (with the former having much more overt connections to pragmatist philosophy and the latter having more direct ties to the development of expert systems within artificial intelligence research), the limitations in the application of standard measurement theory are comparably apparent relative to the data collected from constructivist research and practice. Analyses of conceptual change (students can be referred back to how portions of the population associated with particular modes of computing the average square might be convinced to change their thinking, and thus shift to another mode of computing the average square), from within the larger constructivist framework, illustrate how problematic the urge to think in terms of an average understanding can become.

To connect these ideas to the specific challenges associated with having learners develop understandings of important ideas within the STEM disciplines, the same group of students who answered the "average square" question were later asked to respond to a standard physics question related to freefall (see Fig. 6). The students were asked what forces act on a ball as it is thrown vertically in the air.

Rather than "average" their answers, the constructivist might attempt to discern what the individual answers represent in terms of patterns of reasoning.

For the histogram shown in Fig. 7, each of the bins is taken to represent responses that might be analogous to the distinctive modes associated with computing the average square. Each mode or bin might represent, or be populated by, students sharing similar ways of thinking about freefall. The relative heights of bars in the graph would, like was the case for the results from the average square activity, give us some insight into the relative presence of the respective modes of thinking in a given group of students.

These results for the freefall question are remarkably similar to those for other classes (groups) of university-level students (Halloun and Hestenes 1985) in that the majority of students respond in ways consistent with widely held alternative conceptions (Halloun and Hestenes used the then more common term "misconceptions") about freefall. Nearly all of the students in this particular class had already taken introductory physics at some point in their academic careers and were (much like the students discussed by Halloun and Hestenes), no doubt, told (and tested) about Newton's laws as applied to freefall. This prior instruction notwithstanding, a range of patterns of reasoning existed among our students.

Like the results for the average square, if a convincing case could be made for one of the modes of reasoning—answer "a" would be the standard response expected in an introductory physics course—then some number of students would be expected to shift, in a discontinuous way, from one bin (mode) of thinking to another. Conceptual learning, under

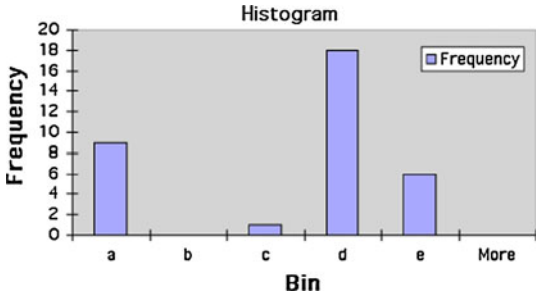

Fig. 7 Class results for the freefall question

this analysis, is related to a kind of restructuring of one's reasoning. Attending to and supporting these shifts is a central focus of constructivist pedagogies. Good teaching, within this framework, centers on a sustained engagement with these different modes of thinking and on the potential of all students to change their minds (c.f., diSessa 2000) about how freefall works.

Relative to the application of standard statistical methodologies, it has been our experience that students are able to discuss the possible parallels: Their results on the average square question are like their results on this freefall assessment item. Similar to the results from the average square activity, the responses on this item represent relatively distinct forms of reasoning about the question posed.

We could even go so far as to discuss, and possibly agree to, a simple metric of relative nearness to the right answer (e.g., "a" would be highest, possibly "c" would be next, and so on).[5] We could then use this metric to project the student results onto an axis of "freefall performance" (not unlike what does happen with multi-item testing focused on specific "objectives" or topics). The mathematical machinery of averaging could then be set loose on these results and an average understanding for freefall computed (based only on the relative distribution found in Fig. 7).

Our students readily appreciate how this value would be the analog of computing the average of the average square. This average doesn't represent anyone's understanding and misrepresents, in a fundamental way, the phenomenology of there being multiple *forms of reasoning* present in the classroom. Averaging can create a value (a heuristic "reality") that may have no significance other than that it can be computed.

Although sometimes harder to see, these issues also extend to multiple item (task- or topic-specific) assessments. As most formal testing has moved away from establishing norms based on implementations of the test as a whole and has moved toward establishing expected norms for a test based on statistical profiles of individual items, an operationally defined reality (e.g., high-stakes test score) of great significance for both teachers and their students is similarly created. The possibility that for a multiple-item assessment some

[5] Such relative ordering of responses "better … characterized" as reflecting "discrete classes or types of understanding" is discussed in a book titled *Knowing What Students Know* published by the National Research Council (NAP 2001, p. 126) as an instance of an ordered latent class model (p. 127). Using this kind of model, each class or mode of response "can be viewed as a point on a continuum" of performance (p. 127) or as a "quantitative parameter" reflecting a relative "degree of proficiency" (p. 151). At an even more basic level, simple dichotomous coding ("1" for a correct response and "0" for any of the wrong responses) also does exactly this (and most item response theory also assigns fine-grain significance to expected values between 0 and 1 for students of varying abilities). Mitchell (2009) discusses this move to assign such quantitative significance as the "psychometricians' fallacy" that he says "occupies a central place in the paradigm of psychometrics" and consists of the tendency to conclude "that an attribute is quantitative from the premise that it is ordinal."

number of students might actually receive the average score shouldn't be seen to diminish the senses in which this average may not, by itself, stand for the complex modes of thinking that structure the student performances. It is very likely, for example, that two students receiving the same averaged score on a high-stakes physics test could have very different ways of thinking about physics. Moreover, much of the information that researchers or teachers would need to engage, and then to further develop, students' reasoning would not be represented by this average. Multiple-item assessments, including criterion referenced tests that use current item response theory, only make it harder to evaluate whether the average has any significance (i.e., says anything meaningful about the modes of reasoning) beyond creating computable artifacts whose stability and consistency may owe more to the procedures surrounding item selection and calibration than to any imputed features of students' understanding of particular domains (Stroup 2009a, b, Pham 2009).

Given that cognitive learning research might look to center more precisely on forms of reasoning that learners *do* bring to tasks, useful assessments would involve tasks that would best serve to make the multiple modes of thinking found in a classroom visible to educators and researchers alike. Deeply suspect would be any urge to then collapse, or substantively mischaracterize, these distinctions in ways associated with computing an average and applying the machinery of standard statistical analysis. Rather than reduce the complexity and contrasts found in the data of actual performance to a single value, a "cognitive" representation (or "statistics") would attempt to follow the patterns and shifts in reasoning (Stroup 1996; Stroup and Wilensky 2000; Hills and Stroup 2004). For the freefall example, assessment should help make visible the many ways students construct an understanding of momentum and its relation to force (c.f., diSessa 1993).

In framing our classroom discussions, students can struggle with what role this multiplicity of understandings might have in organizing teaching and learning. For example, is education more likely to be improved when teachers act upon the forms of reasoning students exhibit rather than reduce this multiplicity to a single test value? As is suggested by our analyses of results from individual tasks, similar concerns arise for the interpretation of average results on individual test items even when such items are ostensibly meant to address specific learning objectives. Can we create non-dichotomous assessment items that are capable of identifying different forms of reasoning (c.f. Stroup 2009b for examples of non-dichotomous multiple choice items that could be used at scale and would center much more directly on issues related to learners' evolving understandings)? How can educational research and school-based practices best respond to the challenges of accountability and the need for standards in education? Are we necessarily bound to averaging-based approaches to assessment in order for systems of formal education to be held "accountable"? Is a simple metric model of mind reasonable? And, indeed, should our answers to this last question be represented by an average?

There are a host of questions that can be raised and tailored to suit the interests of specific classes. We have found the best questions with students tend to be those that juxtapose the power of statistical thinking in certain domains (biology, chemistry, or even studies of beer production) with problems or contexts that may not easily or accurately lend themselves to standard statistical methodologies.

8 Conclusion

We have presented a series of measurement-related discussion activities specifically designed to explore the multimodal and discontinuous nature of important aspects of

human reasoning, as well as some top-level aspects and assumptions related to statistical inference that can be linked to understanding implications of the central limit theorem. Our hope in presenting this account is that others might find both the specific activities and the surrounding discussions to be useful in their teaching. We have found ourselves convinced that in an era when high-stakes testing—based on a relatively narrow range of psychometric practices—is tending to become increasingly pervasive at all levels of formal education (starting, for example, at third grade in the State of Texas in the United States), that our university students actively preparing for the possibility of becoming classroom-based educators should be provided with an opportunity to reflect on the assumptions statistical procedures imply about the underlying phenomenology of learning and teaching.

At the very least the sequence outlined in this paper should support a general awareness in our students of how the deployment of particular statistical procedures and interpretive schemes must be situated relative to the kinds of questions being investigated. This awareness should be as useful to them in their work within STEM disciplines as it might be in their work as STEM educators. Certainly our students should understand the need to look at the underlying population distributions and grapple with how the computing of an average can represent, or misrepresent, the distribution in significant ways. More than this, however, these explorations may serve to make clear the need for a wide range of methodologies to be brought to the task of teaching well. These methodologies can include, of course, qualitative methodologies but also new forms of statistics centered on shifts in the "modes", and related kinds, of reasoning.

The substance of this account also might serve to advance a larger conversation within the research community coalescing around the title of "learning sciences." We do believe a cognitive statistics can and should emerge from what we know of knowing, learning and teaching in STEM-related fields. Methodology doesn't stand apart or "above" a context, but is always situated relative a particular line of inquiry or of purposeful sense-making. Limiting notions of research rigor to averaging-based analyses may bias our findings in ways that may significantly misrepresent the learning-related phenomena at hand.

At the very least, questions surrounding what practices might allow us to think of what we do in the name of STEM education as "scientific" or credibly supporting the goal of making education more "accountable" to a wide range of stakeholders in the educational system, need to be asked and engaged by our students. This, we believe, is true whether or not our students become career–long educators or choose other career paths and end up functioning as informed citizens with what we hope will be a more-than-passing interest in improving on-going educational practice.

Acknowledgments The authors wish to express their appreciation to Uri Wilensky, Andy diSessa, and Bruce Sherin for their longstanding engagement and support in the preparation of this paper, and to John Henry Newman for reminding and reassuring us that some ideas and stances need to be a long time in the world—and may indeed wait for far better advocates than ourselves—as they become "deep, and broad, and full." Funding from the National Science Foundation: Grant # 09093 entitled CAREER: Learning Entropy and Energy Project (W. Stroup, Principal Investigator) helped support this work. The views expressed herein are those of the authors and do not necessarily reflect those of the funding institutions.

References

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn*. Washington: National Academy Press.

Bruner, J. S., Goodnow, J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10*(2–3), 105–225.
diSessa, A. (2000). *Changing minds: Computers, learning, and literacy*. New York: MIT Press.
Duckworth, E. (1987). *The having of wonderful ideas*. New York: Teachers College Press.
Garfield, J. (2002). The challenge of developing statistical reasoning. *The Journal of Statistics Education* [Online] *10*(3). http://www.amstat.org/publications/jse/v10n3/garfield.html.
Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53*, 1043–1055.
Heidelberger, M. (1987). Fechner's indeterminism: From freedom to laws of chance. In L. Kruger, L. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution* (Vol. 2). Cambridge, MA: MIT Press.
Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
Hills, T. (2003). Central limit theorem simulation. http://generative.edb.utexas.edu/projects/esmi/esmiapplets/applets21/centrallimittheorem.html.
Hills, T., & Stroup, W. (2004). Cognitive exploration and search behavior in the development of endogenous representations. Presentation and paper presented to the *Annual meeting of the American Educational Research Association*. San Diego, CA.
Hills, T., & Stroup, W. (2008). Exploring the central limit theorem further. http://generative.edb.utexas.edu/clt/Exploring_the_CLT_further_v01.
Holldobler, B., & Wilson, E. O. (1990). *The ants*. Cambridge, MA: The Belknap Press of Harvard University Press.
Mitchel, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology, 62*, 41–55.
Nicoll, G., Francisco, J., & Nakhleh, M. A. (2001). Three-tier system for assessing concept map links: A methodological study. *International Journal of Science Education, 23*, 863–875.
Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge: Cambridge University Press.
Pelosi, M. K., & Sandifer, T. M. (2003). *Elementary statistics: From discovery to decision*. Hoboken, NJ: Wiley.
Pham, V. (2009). *Computer modeling of the instructionally insensitive nature of the Texas assessment of knowledge and skills* (*TAKS*) *exam*. Unpublished Dissertation, The University of Texas at Austin.
Piaget, J. (1967). *Biology and knowledge*. Chicago: University of Chicago Press.
Piaget, J. (1970). *The Child's conception of movement and speed*. New York: Basic Books, Inc. (Original work published in 1946).
Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*, 211–227.
Reid, A., & Petocz, P. (2002). Students' conceptions of statistics: A phenomenographic study. *Journal of Statistics Education, 10*(2). Online at: www.amstat.org/publications/jse/v10n2/reid.html.
Stroup, W. M. (1994). *What the development of non-universal understanding looks like: An investigation of results from a series of qualitative calculus assessments*. Technical report no. TR94-1. Cambridge, MA: Harvard University, Educational Technology Center.
Stroup, W. M. (1996). *Embodying a nominalist constructivism: Making graphical sense of learning the calculus of how much and how fast*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.
Stroup, W. (2009a). What Bernie Madoff can teach us about accountability in education. *Education Weekly, 28*, 22–23.
Stroup, W. (2009b). *What it means for mathematics tests to be insensitive to instruction*. Plenary Address delivered at the Thirty-First Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education. Available at: participatorylearning.org.
Stroup, W. M., & Wilensky, U. (2000). Assessing learning as emergent phenomena: Moving constructivist statistics beyond the bell curve. In A. E. Kelly & R. Lesh (Eds.), *Handbook of methods for research in learning and teaching science and mathematics* (pp. 877–911). Dordrecht: Kluwer.
Stuart, H. A. (1985). Should concept maps be scored numerically? *European Journal of Science Education, 7*, 73–81.
Swijtink, D. (1987). The objectification of observation: Measurement and statistical methods in the nineteenth century. In L. Kruger, L. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution*, (Vol. 2). Cambridge, MA: MIT Press.
Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.
Wilensky, U. (1993). *Connected mathematics: Building concrete relationships with mathematical knowledge*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Wilensky, U. (1995). Paradox programming and learning probability: A case study in a connected mathematics framework. *Journal of Mathematical Behavior, 14*(2), 231–280.
Wilensky, U. (1999). *NetLogo*. http://ccl.northwestern.edu/netlogo. *Center for connected learning and computer-based modeling*. Evanston, IL: Northwestern University.
Wittgenstein, L. (1953/2001). *Philosophical investigations*. Oxford: Blackwell, Part II, §xi.