Development of Diagnostic Assessments in Probability for Middle Graders

Hollylynne Lee, NC State University (lead author and presenter)
Laine Bradshaw, University of Georgia
Lisa Famularo and Jessica Masters, Research Matters, Inc.
Roger Azevedo, University of Central Florida
Sheri Johnson and Madeline Schellman, University of Georgia
Emily Elrod and Hamid Sanei, NC State University

Research Report presented at National Council of Teachers of Mathematics Research Conference
April 2019

In the United States, formative assessment has historically been thought of as quizzes and tests. In reality, truly *formative* assessment is a *process* used by teachers and students during instruction that provides feedback used to adjust ongoing teaching and learning (Black & Wiliam, 1998; Heritage, 2010; Sadler, 1989). The *DICE* project aims to address this reality by developing a freely-available, web-based assessment system that efficiently provides teachers with timely, accurate, and actionable feedback about student cognition in probabilistic reasoning. Our primary objective is to support learning of foundational concepts in middle-grades statistics and probability by developing and validating a computer-adaptive, diagnostic concept inventory. A *concept inventory* (Hestenes, Wells, & Swackhamer, 1992) is a test created specifically to identify examinees who exhibit robust *conceptions* and *misconceptions* when reasoning. We define a misconception as a conception that is incongruous with expert or scientific understanding.

Our assessment focuses on probabilistic reasoning, a critical middle grades' mathematics concept foundational for developing descriptive and inferential statistical reasoning (e.g., Borovcnik, 2006; Shaughnessy, 2003). Few concepts that appear in mathematical content standards have such wide-reaching impacts in students' lives as probabilistic reasoning. Probabilistic reasoning is a fundamental component of statistical literacy as a trait necessary for thriving in one's citizenship, workplace, and personal life (Franklin et al., 2007). Moreover, the importance of probabilistic reasoning is highlighted in new educational standards (including Common Core), which placed the underlying

ability to reason about probability as a critical prerequisite skill for learning outcomes in the statistics strand of mathematics at middle and high school levels (CCSSI, 2010).

Misconceptions in reasoning about statistical probability and chance are widely documented in the decision-making and statistics education literature (e.g., Borovcnik & Kapadia, 2014; Kahneman & Tversky, 1972; Konold, 1995). Due to the plethora of misconceptions in probabilistic reasoning and the difficulty with probabilistic reasoning that many teachers share (Stohl, 2005), there is a strong need to create assessments that can efficiently identify students' cognitive profiles of probabilistic conceptions, including misconceptions, and effectively communicate that information to teachers.

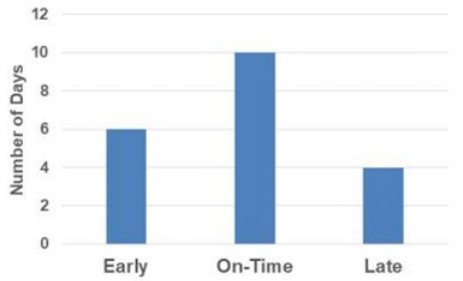
Theoretical Perspectives to Guide Research

We use the term *misconception* broadly to reflect that most misconceptions are a mix of both flawed and productive thinking (Schoenfeld, Smith, & Arcavi, 1993; Smith, diSessa, & Roschelle, 1994; Swan, 2001). The term is somewhat contentious because it can be interpreted to have a negative connotation and be associated with an outdated "fix and replace" instructional approach. Our more broad use of the term reflects that having a misconception can reflect a degree of sophistication and independence in reasoning, which are positive student characteristics. In addition, a misconception is often logically formed and may be useful for progressing towards an accurate and robust understanding of a given concept (Smith, diSessa, & Roschelle, 1994). Some prefer other terms, such as *alternative conception* (e.g., Sadler et al., 2010). At the beginning of the project, we identified 6 misconceptions that were prevalent in literature and developed an initial set of items targeted to diagnose whether students exhibited these misconceptions (sample item in Figure 1). Initial targeted misconceptions were:

1. Probabilities give the exact proportion of outcomes that will occur.
2. (a) More frequent outcomes in sample space increases probabilities of outcomes, regardless of other outcomes. (b) Conversely, less frequent outcomes in sample space decreases probabilities of outcomes, regardless of other outcomes.

3. (a) Higher sample sizes increase probabilities of outcomes, especially outcomes that vary significantly from expected. (b) Conversely, smaller sample sizes decrease probabilities of outcomes, especially outcomes that vary significantly from expected.
4. Later random events compensate for earlier ones (when you ignore independence)
5. Illusion of linearity makes sample size irrelevant.
6. All outcomes are equally likely (without considering that some are much more likely than others).

The school bus is on-time if it arrives at Thomas' house between 7:05 am and 7:10 am. Thomas made the chart below to show the number of days last month the bus was early, on-time, or late.



Thomas is trying to figure out the chance that the school bus will be late tomorrow. Thomas claims, "If the bus runs its normal routine tomorrow, then the chance that it will be late is 1 out of 3." Do you think Thomas's claim is correct?

A. Yes, being late is one of the 3 outcomes that can occur. **M6 equiprobability**
B. No, it is certain that the bus will be on-time tomorrow. **M1 probexact**
C. No, the probability the bus will be early or on time tomorrow is greater than the probability the bus will be late **Correct**

Figure 1. Original item developed to assess misconceptions (indicated in **bold**)

The purpose of this report is to provide an illustration of how our project, through an iterative process, has moved from this simplistic view of misconceptions, to a more robust way of writing items that map onto a more complex conception space of 18 different nuanced conceptions. The iterative process included item development, expert item review, cognitive lab interviews with students, analysis of students' reasoning, and revisiting results reported by others in literature. Our research question is: *How can we best model students' probabilistic reasoning through developing a robust description of a space of conceptions and assessments items that validly and reliably indicate students' reasoning.*

Methods & Data Sources

The initial phase of our study focuses on the iterative development of the diagnostic items (75 items) and collecting validity evidence to confidently map each possible response in an item to specific conceptions that model students' probabilistic reasoning in valid and reliable ways. We use a two-part process to gather evidence of content validity.

After 4-5 members of the research team have iteratively drafted and revised items, expert advisors review items and identify: (a) systematic influences on the item response outside of the target construct an item is designed to assess, (b) ambiguities in wording or context that would confuse students or obfuscate the item's intent, (c) item content or context features that introduce bias for or are culturally insensitive to a subgroup of students, (d) inappropriate levels of item difficulty for the target population, and e) the mapping of item response choices to particular conceptions. While on the surface this review may seem a rather traditional method for a rather innovative assessment, the difference in the expert review for our assessment versus traditional assessments is the grain size of cognition that we asked experts to evaluate. Typically experts are asked if the *item* aligns with a target construct, while we asked experts to also attend to whether all *response options* map to multiple target conceptions.

The second part of our process gathers validity evidence through students' responses to items. We used cognitive labs to gather empirical content validity evidence that the items measure what we intend (e.g., Boorsboom & Mellenberg, 2007). The goal of the labs was to determine the extent to which incorrect options indicate the presence of targeted alternative or misconceptions, and correct options indicate correct reasoning. Students were asked to think aloud as they reason through a set of 8-10 items in the online platform using a laptop. We asked students to (a) reinterpret the question in their own words, (b) verbalize their thinking as they evaluate the possible item options, and (c) describe their final answer selection, including why other options were not selected. Each lab is facilitated by two project team members with extensive experience using these methods with K-12 students, with sessions audio recorded and researchers documenting student actions (e.g., body language, use of technology-enhanced item features) to supplement the recordings.

Thus far, 40 items have been examined through interviews conducted with 25 middle school students (6 6th graders, 12 7th graders, 7 8th graders) at two different locations in the US. Thirteen students were female, and 12 male. The students represent diverse racial backgrounds (15 white, 4 black or African American, 2 Hispanic or Latino, 1 Asian, 1 Moroccan, 1 multi-racial, and 1 who did not specify). Most (17) indicated English spoken regularly at home, while 6 indicated another language was spoken at home at least half of the time one did not indicate). (Note: during 2018-19, 30-40 additional students will participate in interviews, and their results will be compiled with the existing 25).

All interviews were transcribed. Students' responses were coded to determine the extent to which the option selected by a student was reflective of their understanding. We categorized explanations as those that indicated students (a) selected an option indicative of a misconception by demonstrating the intended misconception (true negative), (b) selected an option indicative of a misconception without demonstrating the intended misconception (false negative), (c) selected the correct option and did not demonstrate the intended misconception(s) (true positive), (d) selected the correct option but demonstrated some other alternative or misconception(s) (false positive). For example, for the item in Figure 1, here are three examples of students' reasoning:

> **True negative of M6: Equiprobability "**I think that... <pauses> **I think that it's A.** Because of the probability...because, out of those 3 that it can be early, on time, or late, one out of those 3...there's one of the three chances that it will be late tomorrow."
>
> **True positive of Correct response** "I think I would **go with C**, because even though I'm not like, I wouldn't agree with B because it's not certain that the bus would be on time. But I'm pretty sure they'll be either on time or early, because those are both more (pause) than late [in the graph]."
>
> **False positive of Correct response:** "Well I don't think it's right, but let's read the choices just in case. <Reads choice A> Well it is, but it's been late less than being early or on time so I don't think so. <Reads choice B> It could still be early or late, so that's not my reasoning, but that's my answer [No]. And then C, <Reads choice C> yeah that's what that's pretty much what I'm thinking and that makes the most sense and the other answers I both don't agree with so **I'm gonna go with C.** [the student did not elaborate their reasoning for C, and we interpret "pretty much what I am thinking" and "makes most sense" as an indication that he felt the claim was incorrect, but did not have a strong correct reason for supporting his choice.]

Qualitative analysis was further used to evaluate whether there was confusion about an item or item context, whether students demonstrating understanding converge on the correct response, whether

students that exhibit misconception reasoning converge on a misconception response, and to identify if students were using other ways of reasoning that were not represented in our conception space. Once analysis was completed for all 40 items, we also looked across items to see if there were common ways of reasoning across items that we could better represent with descriptions of conceptions, and how that may impact revisions of items and writing of new items.

Results

The cognitive interviews revealed that there were many other nuanced ways that students were reasoning that did not map directly on to any of the original 6 misconceptions. We also were not well describing conceptions and ways of reasoning that led to correct item responses, with some of that reasoning containing incomplete conceptions of probability ideas. By laying out a larger conception space and revising items so that we could be more confident of how a students' reasoning is leading them to certain responses on an assessment item, we hope to build more robust models of students' thinking that can provide teachers with important diagnostic information that can inform their instruction.

The project is now targeting assessment items with **three common situations**. Within these situations, there are ways of reasoning that can lead to incorrect responses, and others that lead to correct responses. Within each cluster, there are six conceptions described, based on typical ways researchers have observed students' reasoning, both in literature and in our cognitive interviews. The conceptions are clustered based on situations where students are:

1. Reasoning from known sample spaces to estimate and compare probability of possible events
2. Reasoning with empirical data to make sense of probability of events
3. Reasoning about variation from expected when given a theoretical probability

Here, we provide a short list of the six conceptions in Cluster 2. These can be elaborated in the presentation and paper presented at the conference. Conceptions 2a and 2c are re-expressed versions of two misconceptions we initially intended to target. However, we noticed that across items students interacted with in the cognitive interviews, we needed a way to describe a correct way of reasoning (see 2di, 2e, 2f in Table 1) that we had represented in our items. Additionally, there were several other nuanced

ways of reasoning when students were asked to consider situations where data was provided, and we either gave them a theoretical probability to compare the data with or we expected them to infer a probability based on data given.

Table 1: Conceptions for Situation 2 Reasoning with empirical data to make sense of probability of events

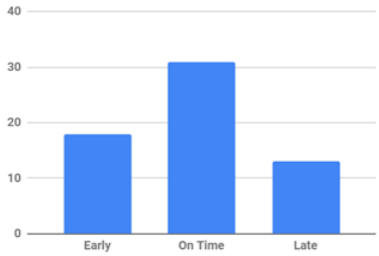
Conceptions	**Working definition**
2a Independence ignored **(Prior Misconception 4)**	If a theoretical probability is known or reasonably assumed and a process generates independent random trials, the student assumes probability of an event can be influenced by a pattern of outcomes of recent independent trials. This can manifest itself in reasoning about negative recency or positive recency.
2b. Random means unpredictable results	Students believe that the nature of randomness means that results from independent trials can not be predicted because anything can happen when things are considered "random". This is connected to a core belief about randomness as being haphazard or unpredictable.
2.c Data determine exact probability **(Prior Misconception 1)**	This conception involves reasoning that a sample of data should exactly represent a probability distribution (or population), without considering sample size or variability.
2d. Using information or data to estimate probabilities	This conception involves students' understanding of the relationship between data and the real world context, and models of probability distributions in a way that affords using data (or information about a context) to estimate probability of events. This reasoning can be **correctly applied** (2di) to use data to inform probability estimates, as well as incorrectly applied because of a student using other information given in the item situation and ignoring given theoretical probabilities (2dii).
2e Appreciates Independence	**Correct Conception.** If a theoretical probability is known or reasonably assumed, a student understands that past results from independent trials do not influence the probability of future results.
2f Long run behavior	**Correct Conception.** This reasoning can be applied correctly to assume the empirical distribution will eventually well represent the population distribution due to the law of large numbers. Students recognize there will be variability between empirical and population distributions but have a strong understanding that a large sample size is needed to see this long run behavior.

As an example, we often saw students using reasoning that indicated they believed that anything could happen with random events and thus they could not well predict anything. In some of our items, students were selecting response choices that were targeted to identify if they thought all events had equal

probabilities. However, we saw that several students were selecting those response choices and then explained they just were really not sure what was going to happen since it was random, and that 50% or "everything equal" was the closest they could come to a prediction. Other students would apply "anything could happen" when they were unsure of a claim. Thus, we added this conception to our list and purposely edited items to include response choices that could tease out if a student held that conception.

Item revisions took into account ways students interacting with individual items, as well ways of reasoning with similar items. For example, in the School Bus item in Figure 1, we noticed students would respond first by indicating whether they agreed with the claim and then were looking for statements that started with Yes or No. We also realized that with the new conception 2b and 2dii (see Table 1), we could alter this item to better reflect these ways of reasoning. Consider the revised School Bus item in Figure 2 that will be used in cognitive interviews in October 2018, where we will collect further validity evidence.

The school bus is "on time" if it arrives at Thomas's bus stop between 7:10 am and 7:15 am. Thomas kept track of the number of days in the past 3 months (62 school days) that the bus was early, on time, or late and made the chart below.



Thomas is trying to figure out the probability that the school bus will be late tomorrow. Thomas claims, "If the bus runs its normal routine tomorrow, then the probability that it will be late is 1 out of 3." Do you think Thomas's claim is correct?

A. Yes **1b: Equiprobability**

B. No

Followup if No is chosen:

Why do you think that Thomas' claim is **NOT** correct?

1. The arrival time of the bus is random, so anything can happen and you can't make a reliable prediction. **2b: RandomMeansUnpredictable**
2. It is certain that the bus will be on time tomorrow. **2c: ProbExact**
3. The probability that the bus will be either early or on time tomorrow is **greater** than the probability that the bus will be late. **CORRECT 2d(i): UpdateProbCorrectly**

Figure 2. Revised School Bus item mapping to new conceptions.

Discussion and Significance of Research

The *DICE* project addresses a specific measurement need in statistics education by developing a sound assessment to help both teachers and researchers better understand students' cognition for reasoning about probability, an especially critical and complex construct. This is significant for teachers because currently no instrument exists, despite it being a content area where students and teachers both struggle. For researchers, our process of collecting validity evidence for our items can inform other researchers as they embark on developing other assessments in mathematics education. Once finalized, our assessment system and diagnostic inventories can promote further study of students' probabilistic reasoning by serving as a research tool to collect quantitative data in systematic, large-scale studies to better understand how probabilistic reasoning develops, how it is related to other statistical reasoning skills, and how interventions impact the development of expert-like reasoning.

Acknowledgement: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170441 to the University of Georgia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-74.

Borovcnik, M. (2006). Probabilistic and statistical thinking. In M. Bosch (Ed.), *Proceedings of the fourth congress of the European Society for Research in Mathematics Education*, Sant Feliu de Guixols, Spain, 17–21 February 2005 (pp. 484–506).

Borovcnik, M., & Kapadia, R. (2014). From puzzles and paradoxes to concepts in probability. In E. J. Chernoff, & B. Sriraman (Eds), *Probabilistic thinking: presenting plural perspectives* (pp. 35-73). New York: Springer.

Borsboom, D., & Mellenberg, G. J. (2007). Test validity in cognitive test. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic test for education: Theory and applications* (pp. 85–118). Cambridge, UK: Cambridge University Press.

Common Core State Standards Initiative (CCSSI; 2010). *The common core state standards for mathematics*. Washington, D.C.: Author.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., & Scheaffer, R. (2007). Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework. *Alexandra, VA: American Statistical Association*.

Heritage, M. (2010). *Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity?* Washington, D.C.: Council of Chief State School Officers.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158.

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

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1), 1-9.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265-296.

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 National Science Standards. *Astronomy Education Review*, *8*(1), 010111.

Schoenfeld, A.H., Smith, J.P., & Arcavi, A. (1993). Learning: The microgenetic analysis of one student's evolving understanding of a complex subject-matter. In R. Glaser (Ed.), *Advances in Instructional Psychology, Vol. 4, 55-175*. Hillsdale, NJ: Erlbaum.

Shaughnessy, J. M. (2003). Research on students' understandings of probability. In J. Kilpatrick, W. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 216-226). Reston, VA: National Council Teachers of Mathematics.

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

Stohl, H. (2005). Probability in teacher education and development. In G. Jones (Ed.). *Exploring probability in schools: Challenges for teaching and learning* (pp. 345-366). New York: Springer.

Swan, M. (2001). Dealing with misconceptions. In P. Gates (Ed.) *Issues in Mathematics Teaching* (pp. 147-165). London: Routledge.