Students' reasoning about probability in the context of a raffle

Razonamiento de los estudiantes acerca de la probabilidad en el contexto de una rifa

Hollylynne S. Lee[1], Lisa Famularo[2], Jessica Masters[2], Laine Bradshaw[3]

and Hamid Sanei[1]

[1]North Carolina State University, [2]Research Matters Inc., [3]University of Georgia

United States

Abstract

Assessing students' conceptions related to independence and probability has been the focus of research in probability education for over 40 years. While we know a lot from past studies about predictable ways students may reason with well-known tasks, developing a diagnostic assessment that can be used by teachers to inform instruction demands the use of familiar and unfamiliar contexts. This paper presents the current work of a research team whose aim is to create a valid formative assessment that uses a psychometric model to confidently predict the presence of a misconception or conception across many items. The focus in this paper is on the evolution of one item and the difficulty it presents in accurately diagnosing students' conceptions of independence.

Keywords: Probability, assessment, misconceptions, students' reasoning

Resumen

La evaluación de las concepciones de los estudiantes relacionadas con la independencia y la probabilidad ha sido el foco de la investigación en la enseñanza de la probabilidad por más de 40 años. Si bien sabemos mucho de los últimos estudios sobre las maneras predecibles en que pueden razonar los estudiantes con tareas bien conocidas, desarrollar una evaluación de diagnóstico que pueda ser utilizada por los maestros para informar la instrucción exige el uso de contextos familiares y desconocidos. Este documento presenta el trabajo actual de un equipo de investigación cuyo objetivo es crear una evaluación formativa válida que utiliza un modelo psicométrico que permita predecir con confianza la presencia de una concepción errónea a través de muchos ítems. En el presente trabajo trata de la evolución de un ítem y la dificultad que presenta para el diagnóstico con precisión de las concepciones de los estudiantes sobre la independencia.

Palabras clave: Probabilidad, evaluación, concepciones erróneas, razonamiento de los estudiantes

This paper is dedicated to Carmen Batanero, who has given the field of probability education a better understanding of historical, theoretical, and practical applications of randomness and probability in education. Her research with students and teachers throughout her career has been groundbreaking and inspirational. The first author, Hollylynne Lee has learned so much through many conversations and recent collaborative writing with Carmen. She is honored to present this invited paper at the CIVEEST conference in celebration of Carmen's 70th birthday.

1. Introduction

In the United States, formative assessment is typically thought of as quizzes and tests. Truly *formative* assessment should be a *process* used by teachers and students during instruction that provides feedback used to adjust ongoing teaching and learning

Lee, H. S., Famularo, L., Masters, J. Bradshaw, L., & Sanei, H. R. (2019). Students' reasoning about probability in the context of a raffle. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística.* Disponible en www.ugr.es/local/fqm126/civeest.html

(Heritage, 2010). The *Diagnostic Inventories of Cognition in Education (DICE)* project aims to develop 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 (ages 11-14) probability by developing and validating a computer-adaptive, diagnostic concept inventory. A *concept inventory* 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 normative understanding.

The DICE assessment focuses on probabilistic reasoning, a critical mathematics concept foundational for developing descriptive and inferential statistical reasoning (e.g., Shaughnessy, 2003). Few concepts that appear in mathematical content standards have such wide-reaching impacts in students' lives as probabilistic reasoning (Batanero, Chernoff, Engel, Lee, & Sanchez, 2016). Probabilistic reasoning is a fundamental component of statistical literacy as a trait necessary for thriving in one's citizenship, workplace, and personal life (e.g., Batanero & Borovcnik, 2016; Franklin et al., 2007). Moreover, the importance of probabilistic reasoning is highlighted in educational standards around the world which place the ability to reason about probability as a critical skill. In the U.S., probability topics are typically introduced in middle school and further developed in high school (Common Core State Standards, 2010).

Misconceptions in reasoning about randomness and probability are widely documented in the decision-making and statistics education literature (e.g., Batanero, et al., 2016; Chernoff & Sriraman, 2014; Jones, Langrall, & Mooney, 2007; Kahneman & Tversky, 1972). 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 and accurately identify students' cognitive profiles of probabilistic conceptions, including misconceptions, and effectively communicate that information to teachers to inform their instruction.

The purpose of this paper is twofold. First, to describe the process used in the DICE project to design valid items targeting certain conceptions and misconceptions in probability for use in a probability concept inventory assessment. Second, to discuss in depth the difficulty in accurately assessing students' understandings of independence and probability in the context of a raffle item in a multiple choice format. The results reported here are part of an ongoing project and should be considered preliminary.

2. Theoretical perspectives to guide research

The DICE project uses 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). The term can be interpreted to have a negative connotation and 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, students' misconceptions are often logically formed and can be useful building blocks for progressing towards a more robust understanding of a given concept (Smith, diSessa, & Roschelle, 1994). At the beginning of the project, we identified six misconceptions prevalent in literature and developed an initial set of items targeted to diagnose whether students exhibited these misconceptions:

MC1. Probabilities give the exact proportion of outcomes that will occur.

MC2. (a) Higher frequency of an outcome in a sample space increases the probability of the outcome, regardless of other outcomes. (b) Conversely, a lower frequency of an outcome in a sample space decreases the probability of an outcome, regardless of other outcomes.

MC3. (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.

MC4. Later random events compensate for earlier ones (when you ignore independence)

MC5. An illusion of linearity makes sample size irrelevant.

MC6. All outcomes are equally likely.

In this paper, we will focus on misconceptions regarding independence of events (4) and equiprobability of outcomes (6).

A misconception about independence, commonly termed the *gambler's fallacy,* is an example of the *Representative Bias* (Tversky & Khaneman, 1974) and stems from a lack of understanding about independent events. Students with this misconception might assume that the probability of an event can be influenced by the pattern of outcomes of recent independent events; i.e., after a streak of events that do not represent the population, an event that will 'even out' the probability distribution is more likely. In the long run, this reasoning can be applied correctly to assume a distribution will eventually represent the population distribution due to the law of large numbers. But the misconception applies to reasoning about individual events, the probabilities of which are not impacted by the law of large numbers. For example, if a fair coin is flipped five times resulting in five heads, the student might believe that it is more likely for the next flip to be tails, as this will *even out* the previous flips. Conversely, a student might believe a head is more likely because the coin is *on a streak of heads*. Thus, students reasoning incorrectly about independence may use a positive recency or negative recency approach. In Batanero and Serrano's 1999 study, they found that students evoked understandings and misunderstandings of independence in supporting their reasoning about whether an event from a series of trials was not random (series of 40 coin flips or 16 cell selections in a 4x4 grid). The 14 year olds in their study were more likely than the 17 year olds to base justifications about an event not appearing random on misunderstandings of independence that they thought there were too many long runs or clusters in cells (4 or 5 heads or tails in a row). Furthermore, this tendency only improved slightly with instruction that included experience collecting data themselves.

A misconception about all outcomes being equally likely is commonly termed the *Equiprobability Bias* (Lecoutre, 1992) and involves assigning equal probabilities to events that are not equally likely to occur. For example, if a cooler contains two red and one blue Gatorade, students will expect that if they randomly choose a Gatorade that it has an equal probability of being red or blue because there are 2 colors. Assuming *random sampling* leads to everything being equally likely often leads students to divide one by the number of possible outcomes in the sample space and assign that probability to each outcome. An overwhelming majority of probability situations discussed in school curriculum *are* based on an assumption of equiprobability: students often use

regular six-sided die, two-colored counters, fair coins, etc. (Lee & Lee, 2009). However, many students, particularly ages 11-14, will overgeneralize equiprobable events to situations where the events are not equal in chance (e.g., Amir & Williams, 1999; Chiesi & Primi, 2009; Madsen, 1995).

Over the past year of the DICE project, through an iterative process of item development and interviewing students, we have not only refined items, but we expanded our descriptions of conceptions and misconceptions. 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 normative 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.

It is beyond the scope of this paper to provide details of this conception space. Herein we will focus on misconceptions related to independence and equiprobability of events.

3. Measuring and classifying students' misconceptions

Bradshaw (PI on the DICE project) has been developing the Scaling Individuals and Classifying Misconceptions model (SICM; Bradshaw, 2011; Bradshaw & Templin, 2014) as a blend of an item response theory (IRT) model and a diagnostic classification model, specifically for use with concept inventories. By blending these two state-of-the-art testing paradigms, the SICM model provides an innovative framework for modeling misconceptions as latent variables. In addition to diagnosing misconceptions, the SICM model will also provide an overall measure of a student's ability to reason probabilistically (as in IRT). For teachers, the greatest benefit of the new methodology is that assessment results will classify students according to sets of misconceptions they exhibit on the assessment by providing diagnostically rich but easy-to-interpret feedback based on the results of sophisticated psychometric models.

Concept inventories, which measure misconceptions with more than one item, typically assess or "diagnose" misconceptions based on a subscore calculated by tallying the number of times an option that measures a given misconception was selected (e.g., Garfield & Chance, 2000; Russell, O'Dwyer, & Miranda, 2009). Using subscores introduces potential measurement error by not allowing items to contribute differentially to misconception subscores (Bradshaw & Templin, 2014). Alignment of psychometric and cognitive theory is critical for providing accurate estimates of examinees' cognitive abilities and for advancing cognitive theories (e.g., Bradshaw & Madison, 2016). An important aspect of the DICE project is to further develop the SICM psychometric model to offer a formative assessment system that features psychometrically-based feedback, administers assessments adaptively, and uses automatic scoring to immediately provide diagnoses of student misconceptions related to probability. The SICM model is aligned with theories of learning that drive the development of concept inventories. Developers of concept inventories hold working theories that misconceptions exist as traits to be measured and that the presence of a misconception will influence which incorrect option a student selects in a systematic way, i.e., a student with a given misconception will likely select options that correspond to the

misconception across a set of items created to elicit the misconception response. Thus, if a student chooses a correct response on a given item in our probability concept inventory, we want to have a high level of confidence that they chose the correct response for correct or normative reasoning. Each incorrect response option should also be able to be mapped to reasoning consistent with a targeted misconception. In the end, all items need to be in a closed form where a correct or incorrect response can be scored objectively (e.g., multiple choice, agree/disagree, true/false) and we tag each response option with the most closely aligned conception or misconception that we have been able to show as holding content validity.

Some items are rather straightforward in targeting conceptions such as independence (e.g., given a streak of 6 heads from a fair coin toss what outcome is most likely to be next or are they equal) or equiprobability (e.g., with 14 pink and 18 blue chips in a box which color are you most likely to draw, blue, pink or are they equal?). However, other items are more complex and present students with a situation that allows for assessing whether they tend to reason correctly with independence or if they apply reasoning consistent with several different misconceptions. For example, in the Raffle item below (Figure 1), a scenario is described, rules are given for a raffle, some data is given for past events, and students are asked to choose a statement that is true about the probability of a future event.

Raffle item Version 1

Your school is putting on a school play this weekend. It will be on Friday, Saturday, and Sunday nights. Each night the school will hold a raffle.
Here are the rules of the raffle:

- Each person can buy only one ticket each night
- The school sells exactly 100 raffle tickets each night
- The school draws 1 winning ticket each night
- A ticket can only be used on the night that it was bought.

The seventh grade science teacher, Mrs. Vail, won the raffle on Friday night. Then she won again on Saturday night!
Mrs. Vail has purchased a raffle ticket for Sunday night.

Which statement is true?

A. Mrs. Vail has an even better chance of winning on Sunday because this is her third try. The more tries she has, the greater chance she has to win. **MC3 (higher sample size, more chances)**
B. Mrs. Vail will either win or she will lose. Since there are 2 outcomes her chance of winning on Sunday is 50%. **MC6 (equiprobability)**
C. Because Mrs. Vail won on Friday and on Saturday, her chance of winning on Sunday are much less than they were on Friday and Saturday. **MC4 (independence)**
D. Mrs. Vail has the same chance of winning on Sunday as she did on Friday and Saturday. **CORRECT**

Figure 1. Version 1 of the raffle item mapped to 3 different misconceptions

The Raffle item has had a particularly interesting role in the DICE project thus far. For this paper, we will focus on the ways students have reasoned with the item and the iterative changes made to the item. We will end with our current conundrum on how we should move forward with the item in our concept inventory.

4. Methods for establishing item content validity

The initial phase of the DICE project focuses on the iterative development of the diagnostic items (about 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 use 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 is to determine the extent to which incorrect options indicate the presence of targeted alternative or misconceptions, and whether correct options indicate correct reasoning. In an interview, students are asked to think aloud as they reason through a set of 8-10 items in the online platform using a laptop. We ask students to: (a) reinterpret the question in their own words, (b) verbalize their thinking as they evaluate possible item options, and (c) describe their final answer selection, including why other options were not selected. Lee serves as the lead interviewer while either Famularo or Masters is a notetaker and secondary interviewer asking occasional follow-up questions. Sessions are audio recorded and researchers document student actions (e.g., body language, use of technology-enhanced item features) to supplement the recordings. All interviews are transcribed verbatim.

Thus far, 52 items have been examined through three rounds of interviews conducted with 48 students (12 6th graders, 25 7th graders, 11 8th graders) at middle schools in three different locations in the US. Twenty-two students were female, 25 male, and one who did not answer. The students represent diverse racial backgrounds (19 white, 14 black or African American, 6 Hispanic or Latino, 3 Asian, 6 other or did not specify). Most (23) indicated English is spoken regularly at home, while 21 indicated another language is spoken at home at least some of the time (with 4 non-responses).

5. Examining students' reasoning on the raffle item

This paper focuses on what we have been learning about item writing, modeling ways of reasoning through defined conceptions and misconceptions, and ways students actually reason through several iterations of the content validity process with a single item that began as the version of the Raffle item in Figure 1. The Raffle item has gone through three rounds of cognitive interviews with the intent to ensure the item options

measure the intended targeted conceptions and misconceptions. Analysis of students' reasoning and item revisions were completed between each round. Across three rounds, the Raffle item has been used in cognitive lab interviews with a total of 31 students (see Table 1) and their reasoning analyzed.

Table 1. Age and gender of students interviewed about Raffle item

	6th grade (age 11-12)		7th grade (age 12-13)		8th grade (age 13-14)		
	Male	Female	Male	Female	Male	Female	Totals
Round 1 Interviews March 2018		1			1		2
Round 2 Interviews April 2018		1	1	1	1	2	6
Round 3 Interviews October 2018	4	2	7	6	3	1	23
Totals	4	4	8	7	5	3	31
	8		15		8		(17M, 14F)

Students' responses were coded to determine the extent to which the item option selected by a student was reflective of their understanding as expressed verbally during the interview. 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).

Qualitative coding 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 theoretical conception space. All coding was reviewed by at least two members of the research team until agreement was reached. A few coded responses were reviewed by additional researchers to assist in making final decisions. Different researchers participated in analysis of rounds 1 and 2 interviews. For the third round with 23 students, two researchers coded responses of the Raffle item until consensus was reached, with a consultant occasionally assisting in final decisions.

5.1 Reasoning with raffle item versions 1 and 2

In round 1 interviews, two students were given version 1 of the Raffle item. Both students chose the correct option (option D in Figure 1) and were reasoning correctly that the chance of Mrs. Vail winning on the third night was the same as it had been the previous nights (coded as True Positive). Because one student expressed concern about the item having too much reading and that he had trouble remembering all the details in the item, we edited the item to reduce reading. Several other edits were also made to make the response options parallel so they each began with an expected statement about the chance of Mrs. Vail winning on Sunday followed by a supporting reason.

Raffle Item Version 2

Your school is holding a raffle after the school play on Friday, Saturday, and Sunday nights.
Here are the rules of the raffle:

- Each person can buy only one ticket each night.
- The school sells exactly 100 raffle tickets each night.
- The school draws 1 winning ticket each night.
- A ticket can only be used on the night that it was bought.

Mrs. Vail won the raffle on Friday night. Then she won again on Saturday night!
Mrs. Vail has purchased a raffle ticket for Sunday night.
Which statement is true?

A. Mrs. Vail has a better chance of winning on Sunday because this is her third try. The more tries she has, the greater chance she can win. **MC3**
B. Mrs. Vail has a 50% chance of winning on Sunday night because there are 2 possible outcomes: She will either win or lose. **MC6**
C. Mrs. Vail has the same chance on winning on Sunday as she had on Friday or on Saturday. **CORRECT**
D. Mrs. Vail's chances of winning on Sunday are much less than they were on the other nights because she already won on Friday and Saturday nights. **MC4**

Figure 2. Second version of the Raffle item used in second round of cognitive lab interviews

In the next round of interviews (Round 2), of the six students who worked on the Raffle item (shown in Figure 2), four gave a correct response and expressed correct reasoning (coded as True Positive). Two students chose option B that the chance of winning was 50% and clearly expressed this made sense to them and used equiprobability reasoning. However one also stated that if it was 50% then the response option stating that the probability was the same on Sunday as Saturday and Friday (option C in Figure 2) was also true since it would be 50% each night.

Several students that got the item correct stated the probability was 1% based on the information given about 100 tickets being sold and Mrs. Vail only having one ticket. Thus, they not only expressed they understood the independence of the events, but they demonstrated an understanding of how to quantify the probability as 1% based on information given in the problem.

Thus, after two rounds of interviews, the item had been seen by 8 students and we had strong evidence that many students could correctly reason about the independence of the raffle draws on each night given the rules, with some quantifying the probability as 1%.

Between Round 2 and 3 of interviews, the project team and expert advisors worked on creating a larger conception space of 18 conceptions across three common situations in which probability reasoning can be invoked. These conceptions represent both correct ways of reasoning and common misconceptions. Therefore, in further editing the item, we decided to add 1% to option C, the correct response. This was done for two reasons: 1) the correct response could be distinguished from the response of Mrs. Vail having a 50% chance each night, based on input from one of our student interviewees in round 2 interviews, and 2) reasoning for the correct response could then be mapped to indicating

correct conceptions of "Appreciates Independence" and "Quantification of Probability" in our conception space. In addition, we changed the stem of the item to request to choose the option with which they MOST agree, in case a student thought more than one option seemed reasonable. The revised item is shown in Figure 3.

5.2 Reasoning on raffle item version 3

Unlike past rounds of interviews where only a few students were given the opportunity to work on the Raffle item, the project team decided to use this item as part of a series of questions given to all students to diagnose whether they were exhibiting misconceptions related to independence and equiprobability, which were two of the misconceptions we had not seen regularly in our past interviews. Through Round 2 interviews, we had seen very little evidence of students reasoning incorrectly on some of our items where we anticipated students might use incorrect reasoning related to independence and equiprobability. This was partly due to the chance that a student may not have been given an item during the interview where they had the opportunity to exhibit that misunderstanding. We also suspected, based on literature reviews, that misunderstandings about independence and exhibiting an equiprobability bias might only be shown in a small number of students. Thus, all 23 students interviewed in Round 3 worked with the Raffle item (version 3, figure 3) as their first item.

Raffle Item Version 3

Your school is holding a raffle after the school play on Friday, Saturday, and Sunday nights.
Here are the rules of the raffle:

- Each person can buy only 1 ticket each night.
- The school sells exactly 100 raffle tickets each night.
- The school draws 1 winning ticket each night.
- A ticket can only be used on the night that it was bought.

Mrs. Vail won the raffle on Friday night. Then she won again on Saturday night! Mrs. Vail has purchased a raffle ticket for Sunday night.

With which of these statements do you **MOST** agree?

A. Mrs. Vail has a better chance of winning on Sunday than she had on the other nights because this is her third try, and she won 2 times already.
B. Mrs. Vail has a 50% chance of winning on Sunday because there are 2 possible outcomes: She will either win or lose.
C. Mrs. Vail has a 1% chance of winning on Sunday—the same chance she had on Friday and on Saturday.
D. Mrs. Vail's chance of winning on Sunday is much lower than it was on the other nights because she already won on Friday and on Saturday.

Figure 3. Third version of Raffle item used in round 3 of interviews

As in the other analysis, we coded each students' response as to whether they chose the correct response option (positive) or an incorrect option (negative), and then whether their verbal reasoning when questioned in the interview indicated their response option matched their reasoning (true), or if there was a misalignment between how they reasoned verbally and the reasoning the response option was coded to represent (negative). Table 2 contains aggregate results from coding of 23 students.

Table 2. Students' coded responses and reasoning with Raffle item version 3

	Positive response (option C)	Negative response (options A, B, or D)	Total
True	8	7 (1 A, 6 B)	15
False	3	5 (1 A, 2 B, 2 D)	8
Total	11	12	23

The Raffle item (version 3) performed relatively well, with students' response choice and reasoning matching the expected conceptual reasoning for 15 of the 23 students (65%). For these 15 students, their reasoning during the interview was aligned with the conceptions intended by the response option. For example for 8 students who chose C, the correct option, their reasoning indicated they understood independence and what the 1% represented (8 true positives). Consider the following sample statements made by students:

> *8th grade Male:* "C, because it says each night you can only buy 1 ticket, and there are 100 raffle tickets each night. You only have a 1% chance of winning."
>
> *7th grade Female:* "I think C is kind of accurate because it's like she has always had the same chance of winning, depending on how many people enter the contest of the raffle. And it's one percent because everybody has as around-they have like one percent of winning because there are a hundred raffle tickets each night. So there's a hundred people or so entering the raffle. So she has a one percent chance, just like everyone else would have."

For seven students, they chose an incorrect option and demonstrated the associated misconception in their verbal reasoning (7 true negatives). Six students chose the option representing an equiprobability bias and showed reasoning consistent with this bias. For example, consider the following exchange between the interviewer (I) and an 8th grade female student (St) who chose option B.

> St: I think it would be B that she has a fifty percent chance of winning on Sunday.
>
> I: And why is that?
>
> St: Because even though she won both Friday and Saturday, it doesn't guarantee that she's going to win on Sunday. There's like a win and lose.
>
> I: Ok. All right. And so that gives it a fifty percent?
>
> St: Yeah
>
> I: So why do you think it won't be A that she has a better chance of winning on Sunday?
>
> St: Because even though luck might be on her side, there's not a huge chance, just there's 50 though.

This student seems to use 50% as a way to express a probability of an event when you are not guaranteed to get a certain outcome and you can either win or lose. Her response also hints that she may understand the independence of the events on each night (3rd line) but it is not clear, because she invokes the notion of luck (7th line) as explaining the winning streak. Though she shows the equiprobability bias, she may also misunderstand the independence of events. One student chose option A and clearly demonstrated that winning twice in a row made the probability of winning the third night higher. Thus, that student was coded as a true negative because they misunderstood independence.

However, based on their verbal reasoning, about 35% of the students (n=8) would have provided evidence they held a conception corresponding to the response option they chose when they did not exhibit that conception. Three students chose the correct option (C) but exhibited *incorrect* reasoning. These students were all attracted to the 1% chance at the beginning of the sentence as representing a very low chance. In their justifications, even when pushed by the interviewer, they indicated that it *did* matter that Mrs. Vail had won the past two nights and that made her chances lower the third night (e.g., "I think it's 1% because she already won on Friday and Saturday night. Just a little chance."). They seemed to ignore the end of the response option that stated the chances were the same as they were Friday and Saturday. Thus, these students exhibited a misunderstanding of independence but were not attracted to either A and D responses as being a best option to choose.

Of the five false negatives, two were by students who chose option B but did *not* exhibit an equiprobability bias. In fact, they appeared to misunderstand independence and were attracted to 50% as an indicator for Mrs. Vail being very lucky and having a higher chance of winning on Sunday. For example:

> *6th grade Male:* "It is B, I feel like since she's won two other times. Honestly think that she has a 50% chance of winning this time, because she's gotten lucky, that's one thing to get lucky on a first time, to get a second time that's really cool and lucky so I feel like she has a 50% chance of winning it on Sunday."

For this student, it is not clear if they think 50% makes sense because it is a high value, or if it is a way of expressing being "lucky", which may in fact be a manifestation of the equiprobability bias, where a uniform distribution would suggest a way to express the nature of a random event as being unpredictable and success achievable by "luck". It seems that for 5 students (3 who chose C and 2 who chose B), they believed that winning on previous nights does impact the chance of winning on the third night, and were attracted to the response options that stated a probability value (1% and 50%) rather than options A and D that made general statements about the probability being lower or higher than on previous nights.

Three other students exhibited a false negative by choosing either option A or option D but did not clearly express a lack of understanding of independence. Instead, the students seemed to attribute the rarity of winning of two times in a row to another cause or influence aside from randomness. For example, a sixth grade female chose option A and believed the outcome of winning two nights in a row was so unusual there must be an explanation for this event other than randomness, and since events were not random, Mrs. Vail would have a good chance of winning the third night. She said:

> "Mrs. Vail is probably like upon somebody's favor who is the one like picking the day. So option "A" is the answer. It could be from my background, it's like I see that she was Friday and Saturday and then apparently she bought another ticket. So she probably she had a friend back there, she's like an insider."

When the researcher asked her to assume there was NOT an insider, she said the answer would be C, showing an understanding of independence. Another student (7^{th} grade male) chose option D and reasoned that there would be a lower chance because they would want to have other people win the raffle; thus rationalizing that the people holding the raffle may not be choosing at random and want to have more than one person win. Another sixth grade student also chose D but their reasoning (after a very long exchange with the interviewer) seemed to indicate they were not holding the four bullet points in the problem as given and that they believed the problem was asking

them to find the chances of Mrs. Vail getting more raffle tickets on Sunday. Thus, this student cannot be confidently coded as misunderstanding independence. The scenario given in this item is so unusual, that a student who has a strong intuitive sense of the likelihood of winning two nights in a row (p=.01*.01=0.0001) might believe such evidence suggests rejecting an assumption that tickets are drawn at random or that every person really only gets one ticket.

In summary, related to an understanding of independence, students' reasoning on this item showed that we would correctly provide a source of evidence that eight students understand independence when they chose option C and that one student misunderstands independence by choosing option A. Two students (one choosing A and one choosing D) would provide inaccurate evidence of misunderstanding independence when the problem situation itself seemed to present past events that were so rare that they instead looked for other causes. One student's choice of D would provide inaccurate evidence because the problem context was so complicated they misunderstood the intent and brought in their own interpretation of what they were asked to find (the chance of Mrs. Vail getting more tickets on Sunday). There were also three students who would also provide evidence as correctly understanding independence (choosing C) but really exhibited a misunderstanding of independence and thought 1% was just a way to express a low chance. Likewise, there were two students who would provide a source of evidence as having an equiprobable bias but really misunderstood independence. They chose B as an option because they felt 50% expressed a high probability of winning on the third night. In total, through talking with students, there were six students who truly expressed a misunderstanding of independence, only one of which would provide a strong source of evidence of holding this misconception by selecting options A or D.

Related to the equiprobability bias, the item fared much better. The research team would code six students as holding the equiprobable bias based on verbal reasoning, and those six selected option B. As noted, there were two students who selected B but thought 50% was a way of expressing a high probability because Mrs. Vail had a better chance of winning, not an equal chance of winning or losing. No student chose an option other than B and exhibited equiprobable reasoning.

5.3 Reasoning across items

Psychometrically, using a single item to diagnose the presence of correct reasoning or a misconception is unreliable and may undermine or confound results (e.g., Gigerenzer, 1994). Coordinating students' responses across a collection of items provides a more robust, and psychometrically reliable, diagnosis of understanding. In the next phase of the project, we will be examining students' reasoning across all items they saw in a cognitive interview to evaluate whether they consistently used reasoning related to certain (mis)conceptions in justifying reasoning on items. The current phase of the project focuses on all students' reasoning *within an item* for the purpose of establishing item construct validity. The next phase will include examining students' reasoning *across items*, both qualitatively from interviewed students, and quantitatively through large-scale field test of items that make it onto our concept inventory test. The qualitative process of examining students' reasoning across items has just begun.

As an example, consider the student who had a true negative for choosing option A on the Raffle item (Figure 3) and a misconception about independence. This student also

demonstrated a misunderstanding of independence on an item with a game where a fair coin toss had resulted in 4 heads. When asked what a player should choose next, the student indicated heads "Because, in the last four games, it landed on heads for each game. So I think for the fifth game, it might pick heads again. So, I think he should pick heads." Thus, the student reasoned consistently across both items.

As a second example, consider a sixth grade student who chose option B in the Raffle item that would indicate an equiprobability bias. However, his verbal reasoning indicated he did not understand independence and felt the 50% expressed a greater chance of winning. On several other subsequent items, he clearly exhibited a misconception about independence. For example, on two different items involving a sequence of 5 coin tosses (TTTTT and THTHT), he showed typical misunderstandings about independence by either expecting a result based on positive recency or based on a pattern of results (e.g., Rubel, 2007, Chiesi & Primi, 2009). On a dice item, he also believed that after seeing five 3's in a row from rolling a fair die, 3 has a higher probability of occurring. Thus, on three items he showed a true negative response about misunderstanding independence. However, on another item with an equipartioned spinner with four colors, he indicated an understanding of independence by disagreeing that the probability of landing on green on the 10^{th} spin depends on the previous spins, even when given concrete examples of results from 10 spins that included many greens. Thus, for this student, we have several sources of evidence that he has a misconception about independence. Responses across several items that provide a sequence of 5 outcomes indicate this, while one item response about probability changing after 10 spins indicates an understanding of independence. This type of qualitative analysis of students' reasoning across items will continue in the next phase of the project.

6. Discussion and Significance of Results

The Raffle item seems to have strong content validity for those who understand independence, can quantify the probability as 1% based on the sample space, and for identifying those who may hold an equiprobable bias. However, the item in its version 3 form does not adequately predict whether a student holds a misconception about independence. Some of the following are main points that the project team will need to consider as we move forward with this item:

1. The past outcomes of the same person winning two nights in a row are very unusual and may be overly influencing students' reasoning on this item.
2. The context of the raffle and the details of the set-up of how tickets are bought each night seems to be simple for some students, but provides an opportunity for students to imagine additional aspects influencing a raffle.
3. Values of 1% and 50% are interpreted as low and high probabilities, respectively, and attract those who cannot yet quantify a probability based on a sample space but have a sense of a probability scale from 0-100.
4. Many students seemed to ignore the stated reasoning in the item responses and only seemed to attend to the first part of each response option that contained a statement about the chance of Mrs. Vail winning.

5. Version 2 of the Raffle item had all true tagged conceptions (4 with correct option and 2 with equiprobability bias).
6. Independence seems to be better assessed with simpler items with less possible outcomes (coins, dice) rather than a raffle with 100 outcomes.

Decisions about how the item will need to change have not yet been finalized. The *DICE* project will continue for the next several years to address a specific measurement need in probability 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 concept inventory 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 development of reasoning.

7. Acknowledgements

This research is supported by grant R305A170441, funded by the Institute of Education Sciences. Thank you to Co-PI Roger Azevedo and research assistants Emily Elrod, Madeline Schellman, and Sheri Johnson for their contributions. Thank you to advisors Michael Shaughnessy, Egan Chernoff, and Jan Mokros for input on the Raffle item. A special thanks to Todd Lee for critical conversations about students' reasoning.

References

Amir, G. S., & Williams, J. S. (1999). Cultural influences on children's probabilistic thinking. *The Journal of Mathematical Behavior*, *18*(1), 85-107.

Batanero, C., & Borovcnik, M. (2016). *Statistics and probability in high school*, Rotterdamn. The Netherlands: Sense Publishers.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). *Research on teaching and learning probability.* Cham: Springer.

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education*, *30*(5), 558-567.

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.

Bradshaw, L. (2011). *Combining scaling and classification: A psychometric model for scaling ability and diagnosing misconceptions.* Unpublished dissertation. University of Georgia, Athens, GA.

Bradshaw, L. & Madison, M. (2016). Invariance principles for general diagnostic models. *International Journal of Testing*, *16*(2), 99-118.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403-425.

Chernoff, E. J., & Sriraman, B. (Eds) (2014). *Probabilistic thinking: presenting plural perspectives*. New York: Springer.

Chiesi, F., & Primi, C. (2009). Recency effects in primary-age children and college students. *International Electronic Journal of Mathematics Education, 4*(3), 259-279.

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.* Alexandria, VA: American Statistical Association.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1&2), 99-125.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129-161). Chichester, England: Wiley.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: are we losing an opportunity?* Washington, D.C.: Council of Chief State School Officers.

Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability: Responding to classroom realities. In D. A. Grouws (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 909–955). Charlotte NC: Information Age Publishing.

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

LeCoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557-568.

Lee, H. S., & Lee, J. T. (2009). Reasoning about probabilistic phenomena: Lessons learned and applied in software design. *Technology Innovations in Statistics Education, 3*(2).

Madsen, R. W. (1995). Secondary students' concepts of probability. *Teaching Statistics, 17*(3), 90-92.

Rubel, L. H. (2007). Middle school and high school students' probabilistic reasoning on coin tasks. *Journal for Research in Mathematics Education, 38*(5), 531-556.

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*(2), 414-424.

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.