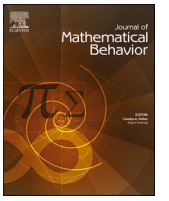
Journal of Mathematical Behavior 71 (2023) 101081

Contents lists available at ScienceDirect

Journal of Mathematical Behavior

journal homepage: www.elsevier.com/locate/jmathb



Validating a concept inventory for measuring students' probabilistic reasoning: The case of reasoning within the context of a raffle

Hollylynne S. Lee [a,*], Hamid Sanei [a], Lisa Famularo [b], Jessica Masters [b], Laine Bradshaw [c], Madeline Schellman [c]

[a] NC State University, USA
[b] Research Matters, LLC, USA
[c] University of Georgia, USA

ARTICLE INFO

Keywords:
Probability
Assessment
Misconceptions
Students' reasoning
Validity arguments

ABSTRACT

Assessing students' conceptions related to independence of events and determining probabilities from a sample space 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 formative concept inventory with strong evidence of validity that uses a psychometric model to confidently predict whether a student exhibits one or more misconception across many items. We illustrate this process in this paper using a particular item with a context of a raffle aimed to measure whether a student reasons with misconceptions related to independence or equiprobability. The results of two aspects of the validity process: cognitive interviews to assess response processes on individual items, and a large-scale administration to examine internal structure of the concept inventory revealed difficulties in assessing students' reasoning about these key probability concepts and trends in the prevalence of misconceptions across grades. Results can provide guidance for others aiming to develop assessments in mathematics education and also support further possibilities for research into understanding students' reasoning about independence and sample space.

1. Introduction

Probability concepts have wide-reaching impacts in students' lives (Batanero et al., 2016) and are a fundamental component of developing statistical literacy for thriving in one's citizenship, workplace, and personal life (e.g., Batanero & Borovcnik, 2016; Franklin et al., 2007; Shaughnessy, 2003). Moreover, the importance of probabilistic reasoning is highlighted in educational standards in the United States, and around the world, which place the ability to reason about probability as a critical skill in middle and high school levels (e.g., Common Core State Standards Initiative, 2010). Even so, teaching probability is often under attended to in schools due to a lack of good resources or teachers' confidence in teaching the topic (e.g., Jones et al., 2007; Stohl, 2005). To help meet this need, the

* Correspondence to: NC State University, 502 C Poe Hall Campus Box 7801, Raleigh, NC 27695, USA.
E-mail address: hollylynne@ncsu.edu (H.S. Lee).

https://doi.org/10.1016/j.jmathb.2023.101081

Available online 30 June 2023

Diagnostic Inventories of Cognition in Education [DICE] project is developing a free web-based assessment system to efficiently provide teachers with actionable feedback about student cognition about several key probability concepts and suggested classroom activities for targeted instruction related to students' needs. The project aims to support learning of probability in middle grades (ages 11–14) by developing and validating diagnostic concept inventories.

A *concept inventory* is a test created specifically to identify students who exhibit *misconceptions* when reasoning. We define a misconception as a conception that is incongruous with expert or normative understanding. Probability misconceptions are widely documented in the decision-making and statistics education literature (e.g., Batanero et al., 2016; Borovcnik & Kapadia, 2014; Kahneman & Tversky, 1972; Konold, 1995), and teachers also have difficulty with their own probabilistic reasoning (Jones et al., 2007; Stohl, 2005). Thus, there is a strong need to create assessments that can efficiently and accurately identify students' cognitive profiles of probabilistic 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 DICE to design and validate items targeting certain misconceptions for use in a probability concept inventory assessment. Second, to discuss nuances in assessing students' misunderstandings of independence and equiprobability in the context of a specific item on our assessment designed to measure if a student exhibits either misconception.

2. Theoretical perspectives to guide research

The term misconception is somewhat contentious because it can be interpreted to have a negative connotation and be associated with a "fix and replace" instructional approach. The discourse on misconceptions involves two contrasting perspectives. Tversky and Kahneman's research (e.g., Tversky & Kahneman, 1974; Kahneman, 2011; Kahneman & Tversky, 1982) posits that misconceptions stem from cognitive biases and heuristics, contributing to systematic judgment errors and negatively affecting comprehension. Conversely, Gigerenzer and colleagues (e.g., Gigerenzer & Gaissmaier, 2011; Gigerenzer & Todd, 1999; Gigerenzer, 2006) highlight the adaptiveness of human reasoning and the potential advantages of heuristics, indicating that misconceptions may originate from misapplication without necessarily reflecting flawed cognition. These divergent perspectives provide contrasting insights into the nature of misconceptions and human reasoning.

The DICE project uses the term misconception to reflect that most misconceptions are a mix of flawed and productive thinking (Schoenfeld et al., 1993; Smith et al., 1994; Swan, 2001). Our broader use of the term reflects that reasoning with a misconception can reflect a degree of sophistication and use of intuitions in reasoning, which are positive student characteristics. In addition, a misconception is often logically formed, and teachers can build from students' initial understandings to assist them in progressing toward an accurate and robust understanding of a given concept (Smith et al., 1994). This paper focuses on two probability misconceptions regarding *ignoring independence* of events and *equiprobability* of outcomes.

Independence is a critical concept in probability. Events are considered independent when the occurrence or outcome from one event is not dependent on another outcome. Consider a bag that contains 5 marbles, 3 blue, and 2 red. If randomly choose a marble from the bag, the probability of choosing a blue marble is 0.6 (3/5). If the marble is returned to the bag after each draw, then the probability of choosing a blue marble on the next draw is also 0.6. These two events (choosing a blue marble on draw 1 and choosing a blue marble on draw 2) are independent of one another, as getting a blue on the second draw is not influenced by the outcome from the first draw since the marble is replaced in the bag. However, if the marble is not returned to the bag (done without replacement), then the probability of choosing a blue on the second draw *is dependent on* whether a blue or red marble is chosen in the first draw, as this would change the sample space with *4* marbles of either 2 blue and 2 red, or 3 blue and 1 red. A misconception about independence, commonly termed the gambler's fallacy (Tversky & Khaneman, 1974), stems from a lack of understanding of independent events, a critical concept in probability. 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., they believe that 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 an empirical distribution will eventually represent a population distribution due to the law of large numbers. Students exhibiting this misconception apply this reasoning to 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 tails, the student might believe that it is more likely for the next flip to be heads, as this will even out the previous flips (e.g., negative recency) (e.g., Chiesi & Primi, 2009). Conversely, a student might believe a tail is more likely because the coin is on a streak of heads (positive recency). Thus, students reasoning incorrectly about independence may use a negative or positive recency approach. Several researchers have shown that students from fifth to ninth grade will exhibit this tendency to ignore independence of events, with more students tending to use negative recency, and the prevalence of ignoring independence being substantial in fifth grade and decreasing for students in 7th or 9th grade (e.g., Fischbein & Schnarch, 1997; Rubel, 2007a).

A misconception about all outcomes being equally likely is commonly termed an 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 reasoning with this bias will expect that if they randomly choose a Gatorade that it has an equal probability of being red or blue because there are two colors. In similar situations, Tarr (2002) reported that students would claim that each possible outcome had a "50–50 chance" because there were two possible outcomes, regardless of the distribution. Assuming random sampling means that everything is 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; Jones et al., 2007; Madsen, 1995).

We know from detailed studies across three decades that students' reasoning about probability can be quite nuanced, and that when given typical multiple choice items, their reasoning (verbal or written) may not always match the intended thinking that an item and response option were meant to evoke (e.g., Groth et al., 2021; Iversen & Nilsson, 2019; Konold et al., 1993; Rubel, 2007a, 2007b). For example, Watson et al. (1997) used or adapted items from prior research (e.g., Green, 1983) and did detailed analysis of students' written justifications for their response selections on multiple choice items. For one of their items related to assessing an equiprobability bias (i.e., picking a name out of a hat with 13 boys and 16 girls and asking whether it is more likely to be a boy, girl or they both have an equal chance), they found that 34% of sixth graders and 10% of 9th graders chose the equally likely option. However, there were different reasons students chose the equally likely option, some of which were well aligned with an expected equiprobable reasoning ("you could get both", "the chance is the same"), but others that indicated awareness of the unequal sample space ("equal but girls have a bit of an advantage") and reasoning that evoked situational or personal experiences ("it depends on how well they are mixed"). Related to ignoring independence, Rubel (2007a) found that when predicting the most likely outcome after a series of 4 heads from a fair coin toss, a large number of fifth graders chose either a negative recency option (tails, 36%) or a positive recency option (heads, 17%), with this decreasing by 9th grade (18%, 10% respectively). However, in her study she found that students' reasoning for their choice did not always indicate a misunderstanding about independence. Some students reasoned that they would expect about equal heads and tails in the long run (correct reasoning) so they would predict a tail in the short run (flawed reasoning). But their responses did not make it clear whether they held any belief about the random nature of the coin toss and independence of each toss. In some cases, students questioned the potential fairness of a coin that would produce four heads in a row, which is also a productive skeptical habit we hope students develop. Such careful attention to students' justification can give us insight into why they make certain choices and should be taken into account when developing any assessment on probability reasoning.

3. Measuring and classifying students' misconceptions

The project is using a modified diagnostic classification model (e.g., DCM; Rupp et al., 2010) to identify if students appear to be exhibiting a misconception. The DCM framework we use is based on literature for modeling misconceptions as latent variables (e.g., Bradshaw & Templin, 2014) and modified for increased measurement precision. For teachers, the greatest benefit of the new methodology is that assessment results will classify students according to misconceptions they exhibit on the assessment by providing diagnostically rich but easy-to-interpret feedback based on the results of sophisticated psychometric models.

Developers of concept inventories hold working theories that misconceptions exist as traits to be measured and that a student reasoning with a given misconception will be likely to select options that correspond to the misconception across a set of items created to elicit the misconception response. 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; Khazanov, 2009; Russell et al., 2009). Using subscores introduces potential measurement errors by not allowing items to contribute differentially to misconception subscores (Bradshaw & Templin, 2014). We know from prior studies (e.g., Groth et al., 2021; Konold, 1989; Konold et al., 1993; Rubel, 2007a; Watson et al., 1997) that students will not always *consistently* apply reasoning across *all* items, giving further support that only using subscores on a concept inventory is likely not the most accurate way to classify students as to whether they exhibit a particular misconception. By using a diagnostic classification approach, we account for that measurement error to provide improved diagnosis of a student misconception. This shift towards diagnostic classification instead of subscores represents an alignment of psychometric modeling and cognitive theory which is important for providing accurate estimates of students' cognitive abilities and advancing cognitive theories (e.g., Bradshaw & Madison, 2016).

4. Designing a diagnostic concept inventory

Hypotheses about how students with misconceptions will respond to assessment items is key to designing a concept inventory. When using multiple choice items, incorrect options are written to be reflections of the types of responses we expect if a student is reasoning with the underlying misconception. If a student chooses a correct response on an item in the probability concept inventory, we want to have high confidence that they chose the correct response for correct or normative reasoning. An incorrect response option should also be aligned to reasoning consistent with a targeted misconception. Items need to be in a form where a correct or incorrect response can be scored by the computer (e.g., multiple choice, agree/disagree, true/false). We then tag specific response options with misconceptions based on the strength of validity evidence.

Many of our items were inspired by or adapted from items used in prior research on students' probabilistic reasoning. Some items in our concept inventory 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 equally likely? With choices: Heads, Tails, Equally Likely). These types of items have been used by many others in research on students' understanding of probability, especially the concept of independence (e.g., Rubel, 2007a; Chiesi & Primi, 2009). On the item with a streak of 6 heads from a fair coin toss, if a student chooses either Heads or Tails, that response is coded as indicative of a misconception of *ignoring independence* (IN). Some items also measure equiprobability in a simple way (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?). If a student chooses the equal option on this item, it is marked as indicating a misconception of *equiprobability* (EQ). This item is similar to the names in a hat item used by Watson et al. (1997) and to items by Green (1983).

Other items we developed 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 one of several misconceptions. For example, in the first version of the item below (Fig. 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. Each response option is tagged as either correct or indicative of a misconception.

The raffle item has had a particularly interesting role in the project as we worked towards collecting validity evidence to support our hypothesized relationships between individual response options and the targeted misconceptions. For this paper, we focus on ways students have reasoned with response options for this item, and the iterative changes made to the item. In doing so, we hope to illustrate our item writing and validation process, while highlighting difficulties in measuring students' understanding or misunderstanding of independence and equiprobability.

5. Methods for collecting validity evidence

Members of the research team iteratively drafted and revised items. We used three phrases and multiple methods to gather evidence of validity based on 1) item content, 2) a students' response processes with an item, and 3) internal structure for a collection of items to predict a misconception classification. We drew upon recommendations in the *Standards for Educational and Psychological Testing* (American Educational Research Association, 2014)concerning using different sources of validity evidence.

In the first phase, we collected evidence of content (or construct) validity to establish whether items were measuring targeted misconceptions. Expert advisors reviewed items and identified: (a) reasoning that could systematically influence students' responses and was different from targeted misconceptions that an item is designed to assess, (b) ambiguities in wording or context that might confuse students or obfuscate an item's intent, (c) item content or context features that might introduce bias or that could be culturally insensitive to a subgroup of students, (d) inappropriate levels of item difficulty for the target population (ages 11–14), and e) the mapping of item response choices to particular misconceptions. Typically experts are asked if the *item* aligns with a target construct. In our study, we asked experts to also attend to whether each *response option* mapped to target misconceptions.

The second phase gathered validity evidence through examining students' response processes to items. This is a second type of validity evidence suggested in the 2014 *Standards*. We used cognitive labs to gather empirical evidence to evaluate whether the items measure what we intend (e.g., Borsboom & Mellenbergh, 2007). Mathematics education researchers have long used this kind of task-based clinical interviewing techniques for understanding students' reasoning (e.g., Goldin, 1997; Zazkis & Hazzan, 1998), and this is especially true for those interested in students' probabilistic reasoning (e.g., Amir & Williams, 1999; Groth et al., 2021; Iversen & Nilsson, 2019). The goal of the labs was to determine the extent to which incorrect options indicate a student is reasoning with a targeted misconception and whether correct options align with correct reasoning. In the cognitive lab interviews, students were asked to think aloud as they reasoned through a set of 8–10 items. In each round, different students saw different sets of items to ensure that each item was seen by several students. 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. One researcher served as the lead interviewer and another project member served as a notetaker and secondary interviewer asking occasional follow-up questions. Sessions were audio recorded and researchers documented student actions to supplement the

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. **(Other reasoning: more trials give 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%. **(EQ)**
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. **(IN)**
D. Mrs. Vail has the same chance of winning on Sunday as she did on Friday and Saturday. **CORRECT**

Fig. 1. Version 1 of the raffle item mapped to different misconceptions.

recordings (e.g., body language). All interviews were transcribed verbatim.

We know that language and cultural everyday experiences can influence students' approaches and reasoning in probability tasks (e.g., Amir & Williams, 1999), and we thus aimed for a diverse pool of students. Seventy-five items were examined through four rounds of cognitive interviews conducted with 66 students (22 age 11–12, 26 age 12–13, 18 age 13–14) at middle schools in three different locations in the US (urban, rural, suburban). Thirty students identified as female, 35 male, and one student chose not to report this information. The students represent diverse racial backgrounds (38% White, 27% Black or African American, 17% Hispanic or Latino, 6% Asian, 12% other or unspecified). Twenty seven (41%) students indicated English was spoken regularly at home, while about half (51.5%) indicated another language was spoken at home at least some of the time (with 5 non-responses).

The cognitive interviews resulted in many items being dropped due to lack of content or response process validity, several being rewritten (some revised more than once), and new items being written based on overall strategies and personal examples we heard from students. We were confidently able to measure students' reasoning on items that targeted four misconceptions: ignoring independence (IN), everything equiprobable (EQ), ignoring relative frequency (IRF), and ignoring sample size (ISS). We split our remaining items into two assessments: *Exploring Probability*, which measured important foundational conceptions (16 items targeting EQ and IN), and *Chance and Data*, which measured more advanced probability tasks using proportional reasoning and understanding the role of sample size in the probability of an event occurring (17 items targeting IRF and ISS). The final version of the raffle item is included on the *Exploring Probability* assessment and measures both EQ and IN.

The third validation phase included a large-scale administration of the two inventories. In accord with a third type of validity evidence discussed in the 2014 *Standards*, we were examining the internal structure of the collection of items and response option choices for predicting the likelihood that a student exhibited one or both misconceptions. In Spring 2020, 28 teachers administered the *Exploring Probability* assessment to 999 students. It will come to no surprise that our data collection was cut short due to the emerging pandemic and closures of schools in March 2020. Forty-one students were considered non-effortful responders as they may have taken too little time (e.g., less than 5 min), skipped more than two items, or appeared to not complete the assessment in a single sitting. Out of the 958 remaining authentic responses, we used students' self-reported grade level as well as teachers' reported grade level of their classes to determine that our sample included: 1.14% (n = 11) fifth graders, 18% (n = 173) sixth graders, 18.6% (n = 178) seventh graders, 35.5% (n = 340) eighth graders, 25.7% (n = 246) ninth graders, and 1% (n = 10) twelfth graders. When asked to report their gender identity, 45.7% (n = 438) students selected female, 43.7% (n = 419) selected male, 2.4% (n = 23) selected non-binary, Third Gender, or Other, and the remaining students (8.1%) either did not respond or selected prefer not to answer. While our sample was majority White (65%, n = 622), it did include students of other racial backgrounds: 81 (8.5%) Black or African American, 15 (1.6%) American Indian or Alaska Native, one Native Hawaiian or other Pacific Islander, 33 (3.4%) Asian, and 76 (7.9%) Multi-racial (note: 13.6% of students either did not respond or preferred not to answer). The majority of students (62.1%) indicated English is spoken regularly at home, while 29.4% (n = 282) indicated another language is spoken at home at least some of the time (note: 81 students either did not respond or preferred not to answer).

6. Validating the raffle item through cognitive interviews

Here we chronicle the exploration of how students reasoned with the raffle item and how, in turn, we tried to model that reasoning in the item response options. We collected evidence of validity based on response processes with the intent to ensure item options measure the intended targeted conceptions and misconceptions. We collected this evidence over four rounds of cognitive interviews using the raffle item, the original version of which is shown in Fig. 1. Analysis of students' reasoning and item revisions were done between each round.

Across all rounds of interviews, we coded each student's 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 meant to represent (negative). Thus, each students' response process was categorized as one of the following:

- true positive: selected the correct option and demonstrating correct understanding,
- true negative: selected an option indicative of a misconception by verbalizing the intended misconception,
- false positive: selected the correct option but demonstrated some other alternative or misconception(s), or
- false negative: selected an option indicative of a misconception without demonstrating reasoning with the intended misconception.

Qualitative coding was further used to evaluate whether there was confusion about an item or item context and to identify if

Table 1
Summary of students' responses and reasoning with all versions of the Raffle item.

	True Positive	True Negative	False Positive	False Negative	Total Students
Round 1 v.1	2	0	0	0	2
Round 2 v.2	4	2 (EQ)	0	0	6
Round 3 v.3	8	7 (1 IN, 6 EQ)	3	5 (3 IN, 2 EQ)	23
Round 4 v.4	2	4 (1 IN, 3 EQ)	1	1 (IN)	8

Note: IN=independence misconception, EQ=equiprobability misconception.

students were using other ways of reasoning aside from our targeted misconceptions. All coding was reviewed by at least two members of the research team until agreement was reached. Table 1 contains aggregate results from coding of 39 students (14 age 11–12, 15 age 12–13, 10 age 13–14) who saw a version of the raffle item in their interview. The next several sections unpack our analysis within each round and how the raffle item was revised accordingly.

6.1. Reasoning with raffle item versions 1 and 2

In round 1 interviews, two students were given version 1 of the raffle item, and both students chose the correct option (see option D in Fig. 1). Both students 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, we edited the item to reduce reading time and remove unnecessary information. Other edits were also made to make the response options parallel in structure so they each began with an expected statement about the chance of Mrs. Vail winning on Sunday followed by a supporting reason. The correct option was also moved to option C.

In the next round of interviews (round 2), six students worked on the revised raffle item (shown in Fig. 2). Four students 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 used equiprobability reasoning when explaining their choice. One of these students also noted that if it was 50%, then the option stating that the probability was the same on Sunday as Saturday and Friday (option C in Fig. 2) was also true since it would be 50% each night.

Two students that answered the item correctly stated the probability of Mrs. Vail winning on Sunday night 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 events, but they demonstrated an understanding of how to quantify the probability as 1% based on information given in the problem.

After two rounds of interviews, the item had been seen by eight students (majority in 7th and 8th grade) 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 rounds 2 and 3 of interviews, the project team and expert advisors suggested edits to the item to add 1% to option C, the correct response. This was done so the correct response that included the phrase “same chance” could be distinguished from the response of Mrs. Vail having a 50% chance each night, based on input from one of our student interviewees. As Figs. 1 and 2 indicate, the first two versions of the raffle item targeted three misconceptions: EQ, IN. and a misconception where a student may reason that if more trials occur, there is a greater chance for a favorable event to happen (e.g., the intuitive rule “more A - more B”, Babai et al.,

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. **(Other misconception: more trials give more chances)**

B. Mrs. Vail has a 50% chance of winning on Sunday night because there are 2 possible outcomes: She will either win or lose. **EQ**

C. Mrs. Vail has the same chance of 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. **IN**

Fig. 2. Second version of raffle item used in the second round of cognitive lab interviews.

2006). We decided to reduce the complexity of what we were measuring on this item. At this point, none of the 8 students had chosen an IN or "more A -more B" response option. Our expert advisors and the research team decided the raffle context could be best used to primarily measure whether a student was using EQ or IN reasoning. Thus option A was also revised to be consistent with ignoring independence so as to possibly attract students who would reason with positive recency rather than negative recency (option D). In addition, we changed the stem of the item to ask students with which option they MOST agree, in case a student thought more than one option seemed reasonable. The revised item is shown in Fig. 3.

6.2. *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** interviewed students to diagnose whether they were exhibiting misconceptions related to independence and equiprobability. Thus, all 23 students in round 3 worked with the raffle item (version 3, Fig. 3) as their first item in the interview.

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 expressed during the interview was aligned with the conceptions intended by the response option. Eight students chose C, the correct option, and their reasoning indicated they understood independence and what the 1% represented. 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."

Seven students chose an incorrect option and demonstrated the associated misconception in their verbal reasoning. One student chose option A and noted that winning two nights in a row made the probability of winning the third night higher. We coded this response as a true negative because they misunderstood independence. Six students chose the option representing an equiprobability bias, and showed reasoning consistent with this bias (all coded as true negative). 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?

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. **(IN)**

B. Mrs. Vail has a 50% chance of winning on Sunday because there are 2 possible outcomes: She will either win or lose. **(EQ)**

C. Mrs. Vail has a 1% chance of winning on Sunday—the same chance she had on Friday and on Saturday. **(CORRECT)**

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**. (IN)**

Fig. 3. Third version of raffle item used in round 3 of interviews.

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. Her response also hints that she *may* understand the independence of events on each night, but it is not clear, because she invokes the notion of luck as explaining the winning streak. Thus, though she shows the equiprobability bias, she may also misunderstand independence.

About 35% of students (8 students) would have been misclassified with a conception corresponding to the response option they chose because that conception did not match their verbal reasoning. Three students chose the correct option (C) but exhibited *incorrect* reasoning (coded as false positive). These students were all attracted to the 1% chance at the beginning of the sentence as representing a very low chance. When prompted to make justifications for their reasoning, the students 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 option that stated the chances were the same as they were Friday and Saturday. The students exhibited a misunderstanding of independence, but were *not* attracted to either A and D options.

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, a 6th grade student explained:

> "It is B, I feel like since she's won two other times. Honestly, I 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 five students (3 false positive and 2 of the false negatives), 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 (measuring IN) 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 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 grader believed the chance of winning two nights in a row was so unusual there must be an explanation for this event other than randomness:

> "Mrs. Vail is probably like up on 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 (7th 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 seemed to indicate they were not holding the four bullet points in the item stem 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 can *not* be confidently classified as misunderstanding independence. A student who has a strong intuitive sense of the very small likelihood of winning two nights in a row ($p = .01 * .01 = 0.0001$) might believe such evidence calls into question an assumption that tickets are drawn at random or that every person only gets one ticket.

Table 2
Five issues and major revisions to raffle item after round 3 cognitive interviews.

Major Issue with Version 3 of Raffle Item	Revision Decision
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.	The problem was changed to only report on winning for one night (Friday), and then asks about results for a second night (Saturday)
2. The context of the raffle and details of how tickets are bought each night seems to be simple for some students, but provides an opportunity for many students to imagine lots of additional aspects of the raffling event.	No changes were made specifically to address this issue. We determined it was necessary to include all information in the item stem, and we had evidence that it was needed for clarity by some students.
3. The values of 1% and 50% are interpreted as low and high probabilities, respectively, and seem to attract those students who can not yet quantify a probability based on a sample space but have a sense of the scale of probabilities being from 0 to 100.	The correct statement with 1% chance was changed to remove the percentage and only state "same chance". The correct response option was revised to be the same as it was in version 2, using the "same chance" language. The statement about winning with a 50% chance was rewritten to simplify and not lead with the 50% chance.
4. Many students seemed to ignore the stated reasoning in the item responses and only attended to the first part of each response option that contained a statement about the chance of Mrs. Vail winning.	We removed "because" statements in three of the option responses. This also reduced the reading load and length of options.
5. When comparing students' reasoning on the Raffle item with 100 outcomes (tickets sold) to other assessment items with less possible outcomes (coins, dice), we noticed better measurement of misunderstanding independence with items with less possible outcomes in the sample space.	The problem scenario was rewritten to only include selling 20 raffle tickets, rather than 100.

In summary, related to an understanding of independence, the item performed well for students that understood independence: eight students were classified by the researchers as verbalizing an understanding of independence and all of those students chose the correct option C.

The item did not perform well for students who misunderstood independence: there were six students who misunderstood independence based on their verbal reasoning and only one of them selected a corresponding misconception option (choice A). Three of the six students selected the correct option (choice C) but exhibited a misunderstanding of independence and thought 1% was just a way to express a low chance. Two students selected the response option corresponding to the equiprobable bias, but the students actually misunderstood independence. They chose B as an option because they felt 50% expressed a high probability of winning on the third night.

Related to the equiprobability bias, the item fared much better. The research team classified six students as exhibiting the equiprobable bias based on verbal reasoning, and all of those six selected option B.

6.3. Revisions to item and reasoning in round 4 interviews

There was strong evidence of validity for the raffle item based on the response processes of students who exhibit understanding of independence and can quantify the probability as 1% based on the sample space (true positives) and students who exhibit an equiprobable bias (true negatives). However, the item in its version 3 form did not adequately capture whether a student held a misconception about independence. Given the multitude of student reasoning exhibited in 23 interviews in round 3, the project team and expert advisors noted the major issues and made significant changes to the item. Table 2 documents the major issues and the resulting revision decision. Fig. 4 shows the revised version used in round 4 interviews.

In round 4, eight students saw the raffle item. Six students' item choice and reasoning matched (75% true positive or true negative): 2 true positives for correct responses, 3 true negatives for equiprobability bias (e.g. "even though she won on one night doesn't mean that she can't win on Saturday night so she still has a 50% chance of winning, but even if she won the night before she could win again"), and one true negative for misunderstanding independence (e.g., winning once made the probability of winning the second night higher).

Two students' reasoning did not match their selected response option. After debating between B (EQ) and C (correct), one student chose C but exhibited equiprobable reasoning (false positive). Similar to a student from round 2, this student argued "I feel like B and C are kind of the same. It is just worded a little different." Thus, for this student the use of 50% and "same chance" was an indication that both options B and C meant the same thing. Removing the "1%" and using "same chance" in the correct option led to a false positive. One student exhibited a false negative by choosing option D (IN) but not clearly expressing a lack of understanding of independence. Instead, the student (a sixth grade male) seemed to attribute the rarity of winning to another cause or influence.

> "I think she has a lower chance [option] D because the winning on Friday was just random and since they sell 20 tickets each night, if they sold a lower amount of tickets each night then she would have 50, and then, so it would go from same, 50, or better chance. Depending on how many tickets they were selling that night."

We saw this type of reasoning in earlier rounds of interviews as well and it is to be expected for students who are beginning learners about probability. This type of false negative occurs when a student attributes probabilities to causes other than randomness and yet chooses option A or D, which indicate misunderstanding independence. This type of false negative is tolerable to us as researchers and teachers. The purpose of our assessment is to give formative feedback to a teacher about possible student misconceptions and misclassified students exhibiting this reasoning would still benefit from targeted instruction related to independence of events.

Raffle Item Version 4

Your school's drama club is holding a raffle for teachers after the school play on Friday and Saturday nights. Here are the rules of the raffle:

- Each teacher can buy only 1 ticket each night.
- The drama club sells 20 raffle tickets each night.
- The principal randomly 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! Mrs. Vail has purchased a raffle ticket on Saturday.

With which statement do you **MOST** agree?

A. Mrs. Vail has a better chance of winning on Saturday than she had on Friday. **IN**
B. Mrs. Vail will either win or lose, so she has a 50% chance of winning. **EQ**
C. Mrs. Vail has the same chance of winning on Saturday as she had on Friday. **CORRECT**
D. Mrs. Vail has a lower chance of winning on Saturday than she had on Friday. **IN**

Fig. 4. Version 4 of raffle item after significant changes, and used in round 4 interviews.

6.4. Final revisions to the raffle item

To help alleviate possible confusion with "same chance" in the correct option, we reworded the correct option to indicate "the chance of winning on Saturday has not changed". We further changed the order of response options to have maximum separation of the EQ option stating 50% and the correct option where the chance has not changed. In addition, we placed the two options measuring a misunderstanding of independence right after one another so they could easily be compared. The final version is in Fig. 5.

7. Results from large-scale administration

For this paper, we report on two aspects of our validation process from the large scale administration. First we examine the response patterns on the raffle item, disaggregated by grade level, for any trends or patterns. We then examine the misconception classifications for the same students from their response patterns on the entire *Exploring Probability* test. This allows us to consider prevalence rates across grades and to compare students' performance on the raffle item with their overall diagnostic classification.

7.1. Response patterns on raffle item

This section considers students' response option selections for the Raffle item across grade levels. Fig. 6 illustrates the proportion of students in each grade level who chose each of the four option responses on the raffle item (Fig. 5).

The proportion of students who chose the correct option (D) improves drastically from 5th to 6th grade, and from 6th to 7th grade, and is fairly stable at about two thirds (65–69%) from grades 7th-12th. This trend aligns with expectation, as we expect older students to be more likely to reason correctly on this item, though we know from other research that both an equiprobability bias and misconceptions about independence can persist until adulthood (Chiesi & Primi, 2009; Fischbein & Schnarch, 1997). The equiprobability bias (EQ) option is chosen much more often than the two options indicative of ignoring independence (IN), especially for 5th-7th grade students, which aligns with our findings from the cognitive interviews (see Table 2). The proportion of students choosing the EQ option

Raffle item (Final) Version 5

Your school's drama club is holding a raffle for teachers after the school play on Friday and Saturday nights. Here are the rules of the raffle:

- Each teacher can buy only 1 ticket each night.
- The drama club sells 20 raffle tickets each night.
- The principal randomly 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! Mrs. Vail bought a ticket for the Saturday raffle. With which statement do you **MOST** agree?

A. Mrs. Vail will either win or lose, so she has a 50% chance of winning on Saturday. **EQ**

B. Mrs. Vail has a better chance of winning on Saturday than she had on Friday. **IN (Positive Recency)**

C. Mrs. Vail has a lower chance of winning on Saturday than she had on Friday. **IN (Negative Recency)**

D. Mrs. Vail's chance of winning on Saturday has not changed. **CORRECT**

Fig. 5. Final version of raffle item used in large scale validation study.

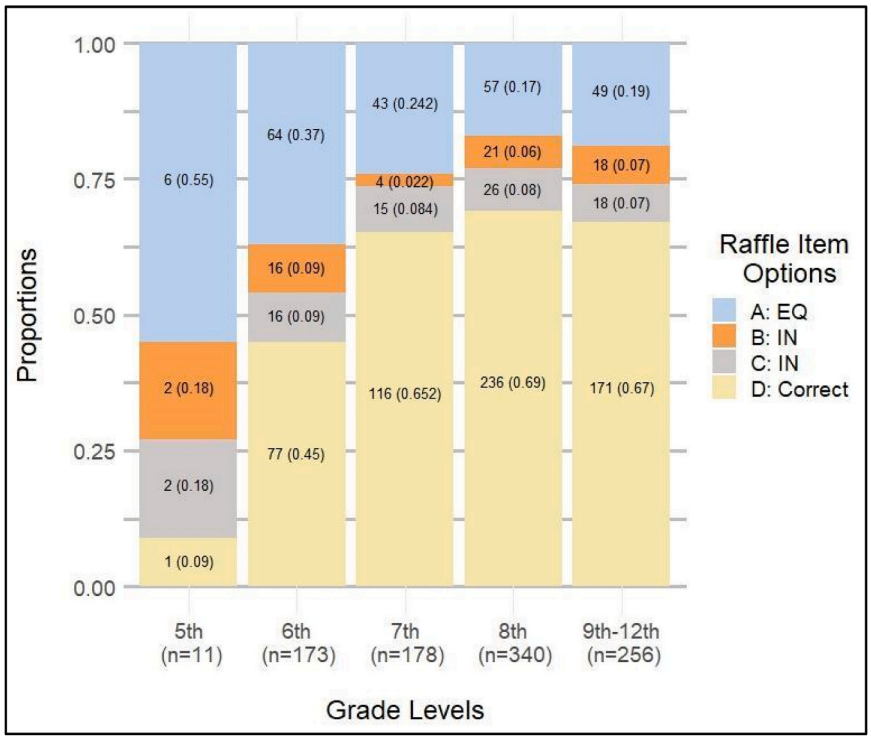


Fig. 6. Frequency and proportion of students selecting response options on raffle item by grade levels.

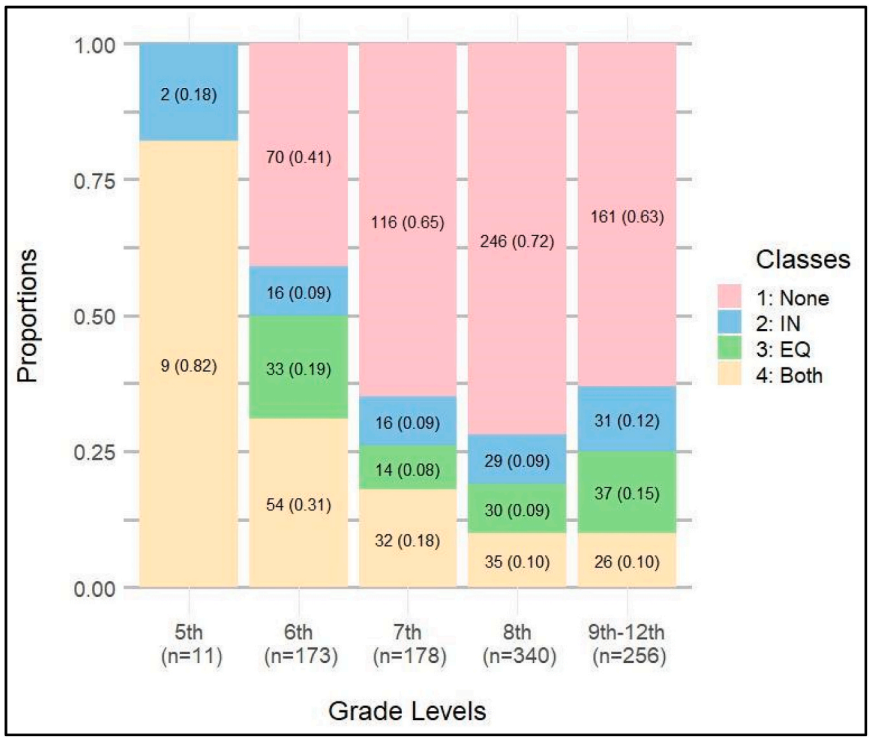


Fig. 7. Most likely classes based on entire assessment by grade levels.

(A) decreases with age, with older students in 8th and 9th-12th grades having the lowest rates of choosing EQ (17–19%). Students were about equally attracted to the negative recency option for IN (C shown in gray) and the positive recency option for IN (B in orange) for all grades except 7th grade, and a slight difference in 8th grade. These findings align with those of other studies that suggest students will reason with negative recency more often than positive recency (e.g., Fischbein & Schnarch, 1997; Rubel, 2007a). It may be that with the structure and context of the raffle item of winning a lottery is more likely to evoke both negative and positive recency reasoning, and is a finding to investigate closely in further studies.

7.2. Misconception classifications

Diagnostic psychometric models were used to obtain statistically stable diagnoses about students' misconception statuses. The diagnostic model simultaneously estimates item parameters and the probability that each student exhibits each measured misconception given their item responses across the entire test. A student likely exhibits a misconception if their probability of misconception appearance is greater than 50%. Therefore, the model provides binary misconception statuses (to answer the question: does the student reason with the misconception or not?) for each student for each misconception.

Because the *Exploring Probability* assessment measures two misconceptions (IN and EQ), each student has two binary misconception statuses: one for IN and one for EQ. The collection of these two statuses is called a *misconception profile*. With two measured misconceptions, we have four possible combinations (referred to in DCM literature as *latent classes* or just *classes*) of misconception statuses. The four misconception profiles predict that:

- Class 1 students do not reason with either IN or EQ.
- Class 2 students do not reason with EQ, but they *do* reason with IN,
- Class 3 students do not reason with IN, but they *do* reason with EQ, and
- Class 4 students reason with both IN and EQ.

Diagnostic models estimate the probability that each student belongs in each latent class. A student's most likely latent class is the class for which they have the greatest probability of membership. The misconception profiles that correspond to the latent classes indicate the students' strengths and weaknesses. Fig. 7 shows the proportion of students in each grade level who were classified by the model into each latent class.

Although we had a small sample of 5th graders ($n = 11$), all of them exhibited at least one misconception on the *Exploring Probability* assessment. In 6th grade, 59% of students exhibited at least one misconception and could benefit from targeted instruction related to these concepts. This aligns with our expectations as the concepts of equiprobability and independence are foundational to

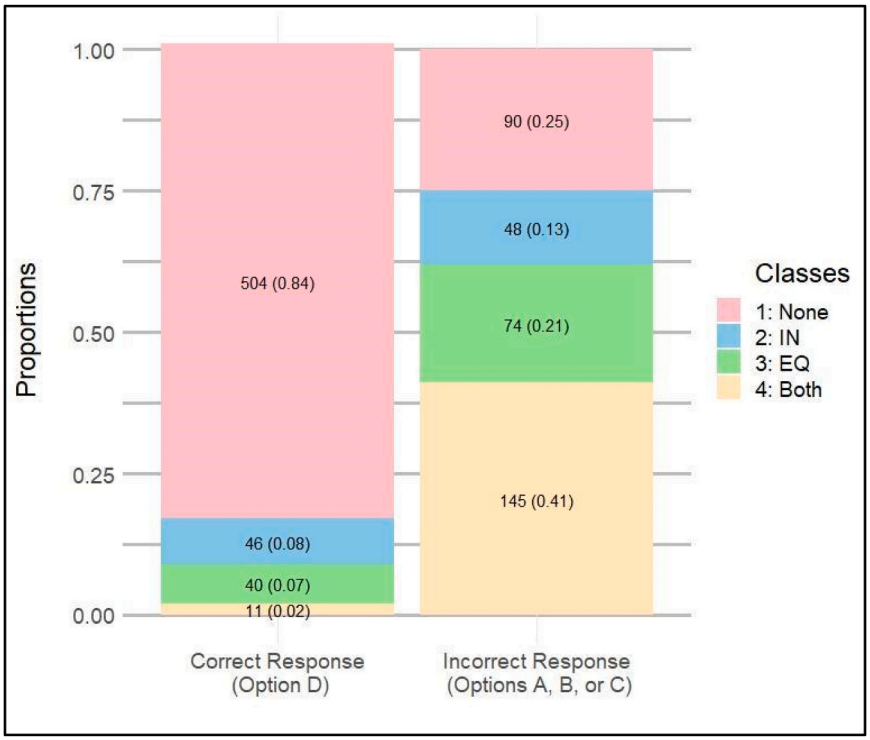


Fig. 8. Relationship between choosing a correct or incorrect response on raffle item and classification based on entire assessment.

probability and have likely not been introduced to many students by mid-year in sixth grade. There is a sharp decrease in this proportion by 7th grade with 35% of students exhibiting at least one misconception. The proportion of students needing support on at least one misconception drops even more for 8th grade (28%) and then increases for 9th-12th (37%) grade students. This indicates that even older students can exhibit misconceptions about fundamental ideas in probability such as equiprobability and independence and further supports our belief that our assessment may be an important tool for teachers to identify those students who may need targeted instruction and educational experiences related to these concepts, perhaps also in college-age students.

We next considered how students' classification on the entire *Exploring Probability* test was related to their responses on the raffle item. Fig. 8 shows the proportion of students in each latent class who either selected the correct option (D) or one of the response options that was hypothesized to align with misconception reasoning (options A, B, and C).

Selecting the correct option (D) in the raffle item is highly related to a student being classified into Class 1 (i.e., exhibiting no misconceptions) on the entire assessment. Fig. 8 shows that 84% of students who did not select any misconception options in the raffle item (i.e., chose D and got item correct) were ultimately classified into Class 1 according to their overall performance. There were 16% of students who correctly answered the raffle item that were classified as holding misconceptions, with 8% later categorized as Class 2 (IN only), 6% as Class 3 (EQ only), and 2% as Class 4 (both). This result shows how a student's correct response to the raffle item might be a strong indicator as to their ultimate classification as reasoning with neither of the targeted misconceptions.

On the other hand, selecting a response option in the raffle item that aligns with misconception reasoning would most likely result in a student being classified as having either one or both misconceptions (Class 2, Class 3, or Class 4). Most students (75%) who did not select the correct option in the raffle item were ultimately classified into Class 2, Class 3, or Class 4 (13%, 21%, and 41% respectively) according to their specific response patterns on the entire *Exploring Probability* test. Only 25% of students who selected a misconception option in the raffle item were later categorized as Class 1 (no misconceptions). This might be an indication that these students either chose one of the misconception options mistakenly in the raffle item, or their reasoning was similar to other students from our cognitive interviews that were coded as "false negative" in their response to this item. A third scenario could be that these students were prone to reasoning with either an equiprobability bias or ignoring independence with items similar in structure and complexity to the raffle. Meaning that they got only raffle item wrong and possibly one or two more items similar to raffle wrong, however they answered most of the other items correctly. The results indicate that responding incorrectly to the raffle item may be a strong indicator that the student reasons with both misconceptions (41% classified as reasoning with both EQ and IN), or at least one misconception (34%). Meaning that they selected a particular misconception option (EQ or IN) in the raffle item and then selected response options aligned with the same misconceptions on several additional items.

8. Discussion and significance of results

Our study has important findings and significant implications across two noteworthy dimensions. The first is the processes we used and lessons learned in writing and validating items for a concept inventory. The second dimension to our findings and implications relates specifically to students' probabilistic reasoning.

8.1. Writing and validating items

The literature in probability education has many examples of a series of items that researchers have used to make sense of students' probability reasoning using open and closed ended items, and asking for written justifications or verbal justifications in interview settings. However, we believe our research is the first, or one of the first, to validate concept inventories to be used by classroom teachers for instructional purposes, as well as researchers. While lengthy, the process of writing items and using a multi-phase validation process, as suggested in the 2014 *Standards* recommendation, that included cognitive interviews and large scale administration with students of the age range for the targeted assessment should lead to a robust validated concept inventory. Such a set of items would have content/construct and response process validity at the item level and collectively be validated as a diagnostic tool for predicting if a student exhibits a few targeted misconceptions related to a key conceptual domain, in our case probability ideas such as independence and sample space. Those working in mathematics education assessment use some of the validation strategies we used, but these strategies are not often used together in a multi-phased process (e.g., Bostic et al., 2019). We hope some of the details of our process provide guidance for others pursuing the validation of such assessments about students' cognition in mathematics.

Rather than only use items that target a single misconception, we felt it was important to have several more complex items on our assessment where we could ascertain if students would be drawn to a correct response, or one of two misconceptions. Though we initially designed the raffle item to include three types of misconceptions as options (to increase the chances of catching all possible reasonings), analysis of cognitive interviews in rounds 1 and 2 showed that this might not be beneficial if the purpose is to measure EQ and IN as primary misconceptions. The evolution of how we revised the raffle item based on some of the nuanced ways students' reasoned is an important lesson in item writing and how seemingly small differences in an item can make a large difference in how students process an item stem and make sense of each of the response options (e.g., 1% in the correct response option could be interpreted as an expression of a low chance rather than an exact quantified probability; students might interpret "50% chance" and "the same chance" as synonymous). The item writing process needs to be iterative in relationship to evaluating students' responses processes with each item.

Through the cognitive interviews, we saw much more consistency with raffle item responses and verbal reasoning using an equiprobability bias. It was much harder to structure the item stem (the context of the raffle) and all the response options to better measure students who reasoned incorrectly about independence of the events. Studies on students' reasoning in probability very rarely

discuss how students reason with different versions of an item and the impact wording changes may have on students' response processes. One notable study was Konold's (1995) research where he noticed a stark difference in how students' reasoned about comparing the likelihood of several series of outcomes from a fair coin toss when asked which sequence was most likely or least likely. The change of one word, "most" to "least", fundamentally changed the structure of Konold's problem for students and evoked different reasoning. We encourage more researchers to examine ways word changes may impact students' approaches and reasoning.

Our assessment can be used for formative purposes as well as in a pre-post scenario to consider if an instructional intervention or educational experience has made a difference in helping students use more normative reasoning in probability and be less likely to reason with certain misconceptions. A small number of high quality items that have been validated can potentially be used by teachers to quickly diagnose whether a student tends to exhibit a particular misconception and then make instructional decisions accordingly to provide support for their learning; and indeed, our finalized assessment system provides suggested instructional activities that are derived from many years of probability education research and practice in the literature.

We are not claiming that any single item can or should be used to diagnose a student's tendency to reason with a misconception. However, our analysis of students' reasoning and response processes with the raffle item, and in relationship to their overall classification from the entire test, suggests that having students complete and justify their reasoning on a more complex item may provide some initial insight into their tendency to reason with or without a misconception. And in fact, the psychometric model showed that some items, like raffle, had higher coefficients and were thus stronger predictors of a students' classification.

8.2. *Trends in students' probability misconceptions*

Many prior studies about students' reasoning related to ignoring independence or the equiprobability bias have reported the prevalence of students' reasoning across grade levels using one or just a few items. The general trend that older students tend to, but not always, reason less with misconceptions about independence or equiprobability has been observed in studies such as Rubel (2007a), Watson et al. (1997), and Fischbein and Schnarch (1999). We saw this same trend in our data, both with the raffle item only (Fig. 6), and in the classifications on the entire test (Fig. 7). However, it is important to note that, different from these prior studies, in the classifications from the *Exploring Probability* assessment, we observed many students in upper middle school (28%, ages 13–14) and high school (37%, ages 14–18) exhibiting misconceptions related to these two key concepts in probability. The prevalence of misconceptions was similar on the raffle item only (31% eighth grade, 33% high school). It is particularly important to note that 17% in eighth grade and 19% of students in high school chose A (in Fig. 5), reflecting that with 20 raffle tickets and one ticket being drawn, there is a 50% chance of winning since you can either win or lose. These students, exhibiting an equiprobability bias, may not understand how to quantify a probability given a sample space. They may need instructional activities that can help alleviate such erroneous reasoning *before* building to more advanced probability activities related to compound events or probability comparisons that are often included in curricula for those grade levels.

Related to reasoning about independence, with the final version of the raffle item (Fig. 5), we had fairly consistent rates of students choosing the response options indicative of ignoring independence from early middle school grades (18%) to high school (14%). Our item had response options available that aligned with both a positive recency and negative recency approach. One interesting finding was that in all grades, there was not a real difference between students' prevalence in reasoning on the raffle item using positive or negative recency reasoning, except for seventh grade which also had the lowest rate of 10% for choosing any response option for ignoring independence (Fig. 6). In seventh grade, many more students used a negative recency approach (8%) than those with positive recency (2%). With some studies reporting on students' selections on a single item, there have been reports that students tend to use negative recency reasoning more often across grade levels. For example, Fischbein and Schnarch (1997) reported that about 35% of students in grades 5 and 7, and 20% in grade 9, used negative recency reasoning in an item with all heads in three fair coin tosses. Rubel (2007a), reported similar trends from a single item about a series of four heads from a fair coin toss, where students tended to use negative recency reasoning more often. In seventh grade, though, Rubel noticed more students indicated a head was more likely (9%, positive recency) rather than a tail (2%, negative recency), which is the opposite of what we saw in the raffle item with more students in seventh grade choosing a negative recency approach. The final form of the raffle item used in the large-scale administration only had two repeated events, winning on Saturday and Sunday, and used a probability of winning of 1/20 compared to a coin toss with probability 1/2 being repeated 3–4 times. Thus, the structure of this item is very different from the fair coin toss items used by Fischbein and Schnarch (1997) and Rubel (2007a). We see potential for further investigating how structure of a problem (i.e., context, probability of success, and number of repeated outcomes) may influence students' reasoning about independence and their proclivity to use a positive or negative recency approach (Sanei and Lee, 2021).

The *DICE* project is addressing a specific measurement and instructional need in probability education by developing an 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 readily available 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 relates to other statistical reasoning skills, and impact of interventions.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hollylynne Lee reports financial support was provided by Institute of Educational Sciences.

Acknowledgements

This paper is supported by grant number R305A170441, funded by the Institute of Education Sciences, United States Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Contributions of the entire project team made this paper possible. Thank you to: Roger Azevedo (co-PI), Emily Elrod (Graduate Research Assistant), and Sherry Johnson (Graduate Research Assistant). Thank you also to expert advisors Chris Franklin, J. Michael Shaughnessy, Jan Mokros, and Egan Chernoff for their input on edits to the raffle item. A special thank you to J. Todd Lee for those critical conversations about the Raffle item and students' reasoning. As of this submission, both concept inventories have been further validated through a second large scale study where we eliminated some items and refined our psychometric model to improve our confidence in students' classifications. The assessment system containing the final concept inventories will soon be publicly available at https://coe.uga.edu/research/labs/dice. To request access to either of the concept inventories for research purposes, please contact the corresponding author.

References

American Educational Research Association , American Psychological Association, & National Council on Measurement in Education, 2014, Standards for educational and psychological testing (Rev. ed.). Washington, DC: American Educational Research Association.

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

Babai, R., Brecher, T., Stavy, R., & Tirosh, D. (2006). Intuitive interference in probabilistic reasoning. *International Journal of Science and Mathematics Education, 4*(4), 627–639.

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,.

Borovcnik, M., & Kapadia, R. (2014). From puzzles and paradoxes to concepts in probabilityE. J. Chernoff, & B. Sriraman (Eds.). *Probabilistic thinking: presenting Plural Perspectives*, 35–73.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education: Theory and Applications* (pp. 85–118). Cambridge, UK: Cambridge University Press.

Bostic, J., Krupa, E., & Shih, J. (2019). *Assessment in Mathematics Education Contexts: Theoretical Frameworks and New Directions*. Routledge,.

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.

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, 2010, The common core state standards for mathematics. Washington, D.C.: Author.

Fischbein, E., & Schnarch, D. (1997). Brief report: The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96–105.

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*.

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

Gigerenzer, G. (2006). Bounded and rational. In R. J. Stainton (Ed.), *Contemporary Debates in Cognitive Science* (pp. 115–133). Blackwell Publishing.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.,.

Goldin, G. A. (1997). Observing mathematical problem solving through task-based interviews. *Journal for Research in Mathematics Education Monograph*, 40–177.

Green, D. R. (1983). A survey of probability concepts in 3000 pupils aged 11-16 years. *Proceedings of the first International Conference on Teaching Statistics* (Vol. 2,, 766–783.

Groth, R. E., Austin, J. W., Naumann, M., & Rickards, M. (2021). Toward a theoretical structure to characterize early probabilistic thinking. *Mathematics Education Research Journal, 33*, 241–261. https://doi.org/10.1007/s13394-019-00287-w

Iversen, K., & Nilsson, P. (2019). Lower secondary school students' reasoning about compound probability in spinner tasks. *The Journal of Mathematical Behavior, 56*, Article 100723.

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. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux,.

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.

Khazanov, L. (2009). A *dia*gnostic assessment *f*or *misconceptions* in prob*ability*. *Paper presented at the Georgia Perimeter College Mathematics Conference in Clarkston, Georgia*.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59–98.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24*(5), 392–414.

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

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

Rubel, L. H. (2007b). The availability heuristic: A redux. *Journal of Statistics Education, 15*(2). https://doi.org/10.1080/10691898.2007.11889467

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press,.

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.

Sanei, H. S., & Lee, H. S. (2021). Attending to students' reasoning about probability concepts for building statistical literacy. *Proceedings of the International Association of Statistics Education Satellite Conference* (available online at) ⟨https://iase-web.org/documents/papers/sat2021/IASE2021%20Satellite%20175_SANEI.pdf?1649974218⟩.

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, pp. *55–175*). NJ: Erlbaum: Hillsdale.

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 developmentG. Jones (Ed.). *Exploring probability in schools: Challenges for Teaching and Learning*, 345–366.

Swan, M. (2001). Dealing with misconceptionsP. Gates (Ed.). *Issues in Mathematics Teaching*, 147–165.

Tarr, J. E. (2002). The confounding effects of "50-50 chance" in making conditionalprobability judgments. *Focus on Learning Problems in Mathematics, 24*, 35–53.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Watson, J. M., Collis, K. F., & Moritz, J. B. (1997). The development of chance measurement. *Mathematics Education Research Journal, 9*(1), 60–82.

Zazkis, R., & Hazzan, O. (1998). Interviewing in mathematics education research: Choosing the questions. *The Journal of Mathematical Behavior, 17*(4), 429–439.