CONFRONTING EXPECTATION IN GRADE 4: TOSSING TWO COINS

Jane Watson[1]
University of Tasmania
Lyn English[2]
Queensland University of Technology

ABSTRACT

This study focuses on the experiences of 91 Grade 4 students who had been introduced to expectation and variation through trials of tossing a single coin many times. They were then given two coins to toss simultaneously and asked to state their expectation of the chances for the possible outcomes, in a similar manner expressed for a single coin. This paper documents the journey of the students in discovering that generally their initial expectation for two coins was incorrect and that despite variation, a large number of tosses could confirm a new expectation.

Keywords: Probability. Expectation. Variation. Coin Tossing. Elementary Students.

[1] Jane.Watson@utas.edu.au
[2] l.english@qut.edu.au

INTRODUCTION

Traditionally the focus of probability learning in the school curriculum could be characterized as focused on Expectation. The theoretical probability of events based on well-defined sample spaces produces an expectation: when tossing a fair coin, P(Head)=1/2 provides the expectation that half "of the time" the outcome will be heads. As school curricula expanded to include trials with concrete materials, students could find variation from the expectation that half of the tosses of a coin would be heads. Moving to the statistics curricula the focus acknowledged Variation because it is variation that necessitates carrying out surveys or experiments. Although acknowledging Variation, however, the purpose is usually to describe the Expectation arising from the data, perhaps as a mean or median. Konold and his colleagues (e. g., KONOLD; POLLATSEK, 2002) characterize this phenomenon as identifying "signal" within "noise." Watson (2005) focused on the concepts of Variation and Expectation as the foundations of the school curriculum in data and chance and claimed that, although the traditional curriculum had emphasized Expectation (in terms of means and medians) before Variation (in terms of standard deviation), in fact young students' intuitive appreciation of Variation developed earlier than their appreciation of Expectation. It was thus suggested that the curriculum should highlight this developmental aspect. This study provided the opportunity to expose students to Variation in both small and large samples to challenge and reinforce developing understanding of Expectation.

Curriculum

The Australian Curriculum: Mathematics (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013) acknowledges variation in the general summary of Foundation to Grade 2, namely, students learn about "presentation and variation of data and a capacity to make predictions about chance events" (ACARA, 2013, p. 8). Further in Grade 3, the "general proficiency" of *reasoning* includes "creating and interpreting variations in the results of data collections and data displays" (p. 34), whereas one of the curriculum descriptions states that students "conduct chance

experiments, identify and describe possible outcomes and recognise variation in results" (p. 38). By Grade 6, students "conduct chance experiments with both small and large numbers of trials using appropriate digital technologies, conducting repeated trials of chance experiments, identifying the variation between trials and realising that the results tend to the prediction with larger numbers of trials" (p. 59).

Other curriculum documents across the world to a lesser or greater extent also acknowledge the importance of a larger number of trials and implicitly, the importance of variation. Campos, Cazorla and Kataoka (2011) reported that the Brazilian curriculum recommends the use of technology "to simulate random experiments that can help students develop an intuitive meaning of probability, observing, for example, the relative frequency of an event over a long run of repetitions" (CAMPOS; CAZORLA. KATAOKA, 2011, p. 8). In comparing the Mexican and Brazilian curricula, Hernandez, Kataoka, Borim da Silva, and Cazorla (2014) noted that although the contents listed in the curricula of the two countries are the traditional, theoretical topics, both countries encourage the use of simulation to assimilate probabilistic concepts. *The Common Core State Standards for Mathematics* (Common Core State Standards Initiative, 2010) emphasizes probability from Grade 7 in the United States and also discusses variation in outcomes with sample size. The *Standards*, however, only explicitly mention simulations for probability in relation to compound events. At the high school level the *Standards* mention computer simulations as useful in several contexts.

PREVIOUS RESEARCH

The history of research into probabilistic thinking did not begin with straightforward questions like the one considered in this study. In writing an early review of research into probability, Shaughnessy (1992) covered topics such as representativeness, availability, the conjunction fallacy, and conditional probability, all of which are much more complex than considering outcomes when two coins are tossed. These early topics reflected research into probabilistic understanding from a psychologist's point of view, where social contexts and people's decision making were considered. Fifteen years later, by the time of the review of probability research by Jones, Langrall, and Mooney (2007), probability was well embedded in the school

curriculum of many countries (e. g., AUSTRALIAN EDUCATION COUNCIL, 1991, 1994; MINISTRY OF EDUCATION (New Zealand), 1992; NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS (United States), 2000). Mathematics education researchers were beginning to ask more detailed fundamental questions about students' intuitions concerning basic events and their probabilities. Jones et al. began by considering the language of chance and the influence it has when one is asked a question about a specific probability. Whereas using language such as, "it's a 50-50 chance" is appropriate when tossing one fair coin, to mean it is equally likely to get a head or a tail, such terminology is not appropriate if there are more than two equally likely outcomes, for example when tossing a die (TARR, 2002). It is also common, however, for similar language to be used for any two or more events that are not equally likely, with Amir and Williams (1999) and Pratt (2000) observing the responses with students dealing with one or two dice. Moritz and Watson (2000) observed a similar claim of 50% when predicting the probability of 4 tails occurring on 4 successive tosses of a coin. Overall evidence from different contexts suggests that phrases such as "50-50 chance" may be used in any situation where outcomes are uncertain or happening "by chance" (LECOUTRE, 1992). This can create confusion when on occasion 50-50 is the correct answer but the explanation is vague or missing.

Although there have been some studies focusing totally or mainly on tossing various numbers of coins (e. g., KONOLD *et al*. 1993; MORITZ; WATSON, 2000; RUBEL, 2006, 2007), the main interest for this study is the research that has included an emphasis on two coins. Carpenter *et al*. (1981) reported on student outcomes from the National Assessment of Educational Progress (NAEP) on the two coin problem and found 60% of 13-year-olds and 69% of 17-year-olds gave the correct answer of 1/2 for obtaining a head and a tail. Somewhat disturbing, however, were the results for the question about the probability of obtaining two heads, where 58% of 13-year-olds and 50% of 17-year-olds again gave the response of 1/2. Carpenter et al. 1981 attributed this inconsistency to "the students' reliance [on] 1/2 as being correct" (CARPENTER et al. 1981, p. 344). One suspects that this observation was a precursor to Lecoutre's (1992) description of the equiprobability bias.

Rubel (2006, 2007) provides strong evidence for this bias after adapting the Carpenter et al. question for a head and tail, surveying 173 students in Grades 5, 7, 9,

and 11, and asking for explanations of the answers. Fifty-four percent of students gave the correct answer of 1/2, 23% stated 1/3, 13% said 1/4, and 11% gave another or no answer. The students answering 1/3 generally explained that there were 3 equally likely outcomes – 2 heads, 2 tails, and one of each. Of those who answered 1/4, some considered the order of the 2 coins and the compound event of one outcome followed by the other, hence misinterpreting the question. Of the 93 students who responded 1/2, 49% (mostly in Grades 9 or 11) explained their response based on some type of sample space argument. Thirty-two percent of the students (mostly in Grades 5 and 7) justified their responses of 1/2 with a 50-50 type answer not addressing the sample space. Rubel did not ask for the probability of two heads as Carpenter et al. did, but Rubel's further analysis of explanations supports the view that one cannot necessarily believe all students understand the appropriate reasoning for the correct answer.

A significant extract from Rubel (2006) illustrates the need for the second part of the current study reported here.

> An example of the 50-50 approach is demonstrated in the following excerpt from an interview with Darnell, a seventh grader. The researcher attempted to create a cognitive conflict for Darnell by adjusting the number of coins in the problem.
>
> Darnell: It's a 50 percent chance each so he has an even chance of getting both. Even if you have two quarters, there's still going to be a 50 percent chance.
>
> Researcher: What if we had three quarters? What's the probability that we get all tails?
>
> Darnell: I still say 50 percent.
>
> Researcher: Why's that?
>
> Darnell: Because unless something affects the way the quarters come down, it's still going to be equal.
>
> Researcher: What if we had 100 quarters? What's the probability that we get all tails?
>
> Darnell: Half-way.
>
> Researcher: So if we threw up 100 quarters, you think we'd have a 50 percent chance that every single one of them lands on tails?
>
> Darnell: Yeah.
>
> Researcher: Okay – what about 100,000 quarters?
>
> Darnell: That's a lot. But it's still 50 percent.
>
> (RUBEL, 2006, p. 52)

Rubel goes on to speculate on what the student meant: perhaps “it is as likely as not to happen”; perhaps “it *could* happen”; perhaps “he doesn’t know what will happen.” These speculations and the possibility that anything can happen again fit Lecoutre’s equiprobability bias. The next research question then should focus on what would be the results of exposing Darnell to many repeated tosses and having him observe the number of times (as a percentage) that all tails occur.

In research before the era of software that could produce many simulated trials for students to observe, researchers did ask survey questions based on the influence of sample size. Fischbein and Schnarch (1997) for example asked about which was more likely, getting heads at least twice when tossing a coin 3 times, or getting heads at least 200 times when tossing a coin 300 times. They also asked an equivalent sample size question based on Kahneman and Tversky’s (1972) Hospital problem: was a larger or smaller collection of births more likely to have 60% or more boys? At the time these were strictly theoretical problems. The development of the software *ProbSim* (KONOLD; MILLER, 1994) allowed Watson (2007) to explore if students who believed in equivalent outcomes regardless of sample size would change their opinions if they observed simulations taking place. In that study of 34 students who initially believed the sample size (10 versus 30) did not matter in an equivalent of the Hospital problem, 27 (79%) changed their minds favouring the better chance of reaching the extreme result with the smaller sample size. The appropriate reasoning for this answer is the underlying variation from expectation, which is larger for statistics that are based on smaller sample sizes and smaller for statistics based on larger sample sizes, a premise covered in many text books and illustrated in Watson (2006).

Although the context for asking questions about increasing sample size was different in the present study, the purpose being to get closer to a theoretical value rather than further away, the technique of repeated sampling is the same. Ireland and Watson (2009) used simulations for tossing a single coin up to 10,000 times to demonstrate to a Grade 6 class that the proportion of heads approaches 50%. This activity was a prelude to considering the trend of outcomes of tossing an ordinary die to approach 1/6 for each of the six numbers with an increasing number of trials. To the authors’ knowledge, using this simulation technique in the school classroom has not been reported before in considering outcomes for tossing two coins.

AIMS

In moving from previous research focusing on answers to specific probability questions and perhaps the follow-up justifications of some students, this study was classroom based, building on students' experiences with a single coin, initially to create a model illustrating the expectation of outcomes. It was of interest to explore students' beginning beliefs and document how first performing trials, then carrying out simulations, led to a change in belief about the underlying model for tossing two coins.

The experiences students had had in the first part of the study with a single coin (ENGLISH; WATSON, 2014) had not only confirmed their initial expectation that the probability of obtaining a head on a single toss was 1/2, but also exposed them to variation across many samples of 10 trials that could challenge their expectation. Simulations with larger sample sizes demonstrated that the percentage of outcomes (heads) approached 50% as the sample size increased (WATSON; ENGLISH, 2013). The aims of this part of the study hence were to:

(i) allow students to express their expectation for all outcomes possible when two coins are tossed;

(ii) document students' experiences and conclusions from tossing two coins 12 times and recording the results, followed by combining these results for the whole class.

(iii) allow students to complete large numbers of trials in the software *TinkerPlots* and document the influence on their beliefs.

METHODOLOGY

Participants

Four grade 4 classes and one grade 4/5 class from a large, middle socio-economic Australian school participated during the first year of a three-year longitudinal study (2012-2014). This paper focuses only on the Grade 4 students whose parents had given ethics approval (n =91; mean age = 9.5 years) in reporting

the findings, 43% of whom were classified as learners with English as their second language (ESL).

Design and Implementation

A design-based research approach was adopted, specifically, a design experiment, which involves engineering an innovative educational environment that supports the development of particular forms of learning and then studying the learning that takes place in the designed environment (COBB; JACKSON; MUNOZ, in press; KELLY; LESH; BAEK, 2008). Given that the teachers' involvement in the study was vital, professional development sessions were conducted in preparation for their implementation of the activity lessons; these were followed by debriefing sessions where researchers and teachers reflected on the students' and their own growth and development.

The long-term purpose for the students by the end of Grade 6 was to have an understanding of informal inference (MAKAR; RUBIN, 2009), to be able to apply this understanding to conducting statistical investigations on their own initiative, and to develop critical thinking wherever statistical literacy contexts arise in their experience. Part of the foundation in the first year for developing this understanding was experiencing variation and expectation as they arise in situations involving data (WATSON, 2005). As well, the tool employed for students to experience and explore data sets constructively was *TinkerPlots: Dynamic Data Exploration* (KONOLD; MILLER, 2011). The activities in the first year were hence focused on introducing students to the software and through use of the software in two very different contexts, experience the relationship of variation and expectation while making predictions and confirming or disconfirming them.

The first activity was a 2-hour benchmarking session where students created a survey and conducted it with their classmates, creating pictorial representations of the results, which they then presented to the class and displayed as posters throughout the school (English & Watson, in press). The purpose was to provide the research team with starting points for the two main activities for the year. The first extended activity introduced the graphing features of *TinkerPlots* as a tool to display data

collected when all students in each class measured and recorded the arm span of one student in the class. This was followed by having all students' arm spans measured and recorded. The difference in the variation in measurements from the two data sets and the related expectation for the accuracy of the estimates of arm spans was a focus of the activity (ENGLISH; WATSON, 2015). The second extended activity of the first year focused on probability and introduced the Sampler in *TinkerPlots* to students in order to carry out large numbers of simulations of trials of tossing one coin or later two coins.

The regular classroom teachers conducted the activity, which occupied an entire school day. This part of the study occupied a 5-day week with one class completing the activity each day. A research team of 4 (including the authors) was in attendance each day to assist the teachers and students, and collect data via video taping of two focus groups (each comprising two students) in each class, camera shots, and notes of observations. A detailed lesson outline was prepared for the teachers, as was a workbook for students. The student workbook was 19 pages long and consisted of instructions for hands-on activities with coins for students, together with questions about the outcomes requiring students' written responses. As well, the workbook included approximately seven pages of explicit instructions with screen dumps for setting up the Sampler in *TinkerPlots*, which the students were expected to create for themselves. A very reduced version of the workbook with space for writing and *TinkerPlots* instructions removed is presented in the Appendix for the second part of the activity involving two coins. The lesson plan for teachers was 12 pages long, including the blank tables to be transferred to white boards for recording student data and suggested solutions for models that the students were asked to create. The introduction to the lesson plan included the relevant descriptors from *The Australian Curriculum: Mathematics* (ACARA, 2013) for Grades 4 and 5 for Probability and links to the state curriculum, which the teachers were being asked to implement that year. Following a request from one of the teachers after the first extended activity, the lesson plan and student workbook were combined for the teachers with the relevant workbook pages interleaved with the lesson plan pages.

Students worked in pairs and each pair had a laptop computer for work with *TinkerPlots* installed. Although the activity of coin tossing and carrying out simulations

was conducted in pairs, students were asked to write their own answers and explanations in their individual workbooks. At the end of the year, 43 students were selected to be interviewed, either in groups or individually, about the activities undertaken during the year. These students had either been members of class focus groups or were observed by teachers or researchers to provide interesting comments during the classroom discussions.

Data

The data analyzed in this report came from the sources listed in Table 1. They are associated with the aims of the study. Some students were absent for part of the day for other school requirements and hence for some items the number of responses is reduced by one or two.

Table 1 - Data collection associated with aims of the study (*n*=91)

Aim		Data
(i)	Expectation for tossing 2 coins	Appendix: Q1 (categories of model)
(ii)	Student data from 12 tosses of 2 coins, experiencing Variation	Appendix: Q3 (agreement with model), Q4, Q7 (range of plots provided)
	Combined class results	Appendix: Q10, Q11 (levels of response)
(iii)	*TinkerPlots* data from large sample sizes, with less Variation	Appendix: QT6, QT8, QT9 (levels of response)
	Final Expectation for tossing 2 coins	Appendix: Q12 (categories of model)

Analysis

The data were clustered by similarities in the understanding displayed in the written responses to four groups of questions: Q1; Q10 and Q11; QT6, QT8 and QT9; and Q12. The criteria for the categories, which generally reflect the correctness of the responses are given when the results are presented. For Q1 and Q12, categories are clustered by whether the basic representation is based on three outcomes or four. For Q10 and Q11 hierarchical categories reflect the growing appreciation of variation and change in pattern as the number of trials is increased. After the introduction of *TinkerPlots*, transcripts from the classroom discussions as well as workbook extracts document the students' impressions of the emerging model. The final understanding

from the day is expressed in the models created in Q12. Q7 responses of students show great variation in the type of plot created and are categorized into nine groupings. Extracts from end-of-year interviews with some students are provided as evidence of the understanding being developed or consolidated in relation to the combination of the class data and creation of a final model after many simulations are completed.

RESULTS

Aim 1: Expectation for tossing two coins

As was to be expected for students with little experience in analyzing the possible outcomes from tossing coins, the mention of three outcomes – two heads, two tails, or one of each – led to the suggestion by many students (74%) of three outcomes, with a mixture of associated probabilities. Although 23% of students could list four outcomes only two could correctly associate the probability of 1/4 with each. Table 2 details the subgroups of responses for models for either three or four outcomes, with 35% of students not assigning specific probabilities to their outcomes and the same percentage assigning inappropriate values, often not summing to 1.

Table 2 - Preliminary models for outcomes of tossing two coins [Q1 in Appendix]

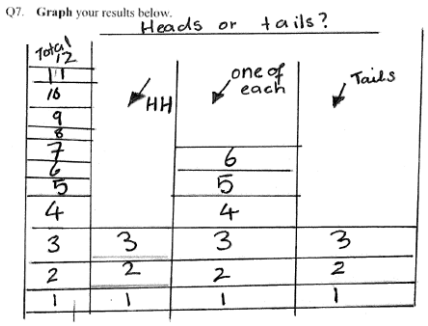
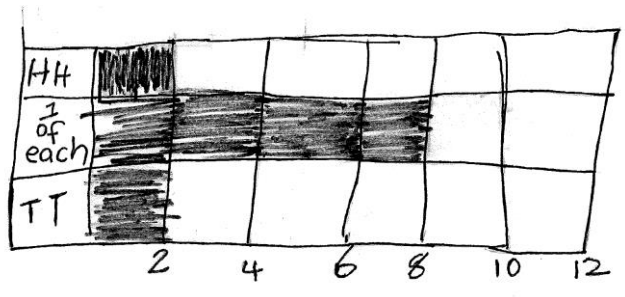
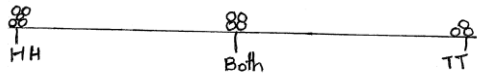
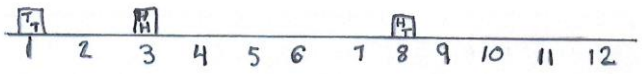
Category	Description	Sub Category	Description	Frequency
A	Four outcomes			21
		A1	Distinguishes 4 outcomes with probabilities	2
		A2	Distinguishes 4 outcomes, no probabilities	13
		A3	Distinguishes 4 outcomes, incorrect probabilities	6
B	Three outcomes			67
		B1	Distinguishes 3 outcomes, probabilities 1/3, 1/3, 1/3	22
		B2	Distinguishes 3 outcomes, no probabilities	19
		B3	Distinguishes 3 outcomes, inappropriate probabilities	26
NA	Idiosyncratic/ No response			3
				91

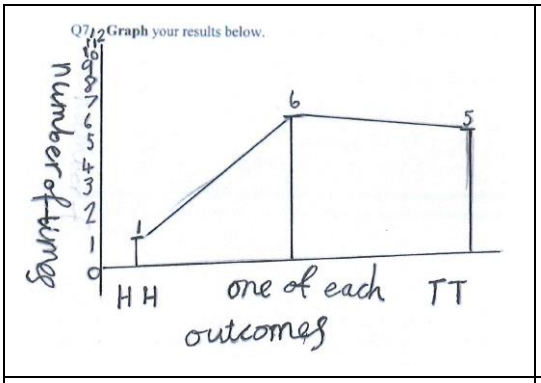
The outcomes show that at the beginning of this part of the activity only two students could state an accurate model for tossing two coins, although 13 could list the four outcomes without assigning a probability. Despite experiencing the earlier part of the activity with the equally likely value of 1/2 for outcomes for a single coin, only 22 students (24%) carried this equal probability idea forward to the three outcome model with a probability of 1/3 for each outcome.

Aim 2: Document students' experiences conducting trials

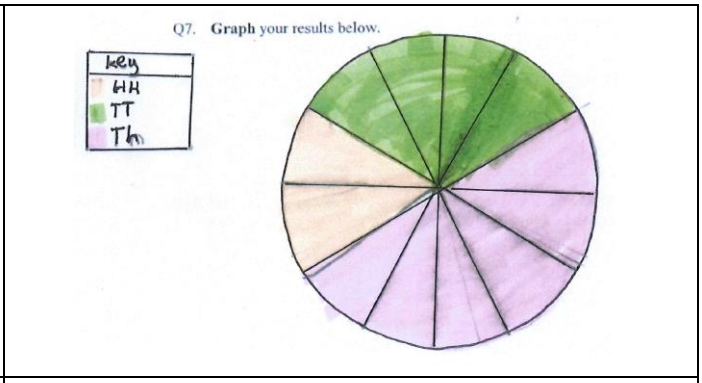
Following through the questions in the workbook, students predicted the outcomes for 12 tosses of two coins, carried out the trials, recorded answers in a table (Q4) and graphed the results in any way they preferred. No instructions were given for the representation. Nine categories of graphs were identified with types, percentages, and examples provided in Figure 1. The main experience students had had with graphing before engagement with this project was using bar graphs but earlier in the project they had been introduced to stacked dot plots through their use of *TinkerPlots*. These were the two most frequently used types of graphs for this task, accounting for 78% of representations. Between two and six students drew value plots, line graphs, pictographs, or pie graphs, with two producing tallies or tables.

Figure 1. Examples of graphs produced of outcomes for 12 tosses of two coins (2% were not complete).

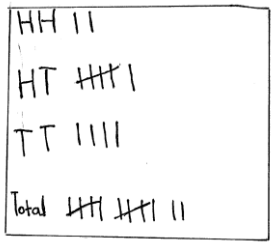
<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Bar graph-vertical (43%)</td><td>Bar graph-horizontal (13%)</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Stacked Dot plot (22%)</td><td>Value type (7%)</td></tr>
</table>



Line graph (3%)



Pie graph (3%)



Tallies (2%)

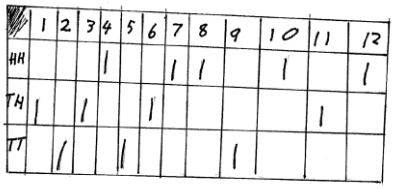
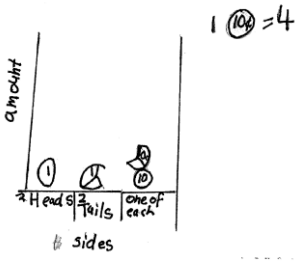


Table (2%)



Pictograph (2%)

The group results, when combined for the entire class, are shown for one class in Figure 2, with students sticking post-it notes on a white board. Similar plots were produced in each class. Question Q10 (Appendix) focused on why the shape of the combined class was about twice as high in the middle as for two heads and two tails. Of interest were the levels of response and particularly those that could go beyond the visual observation to explain the combination of outcomes. Table 3 summarizes the levels of response with examples and frequencies. Sixty-five percent of students had at least an intuitive appreciation of the reason for the observed outcomes.

Figure 2. Combined outcomes from all groups in one of the classes

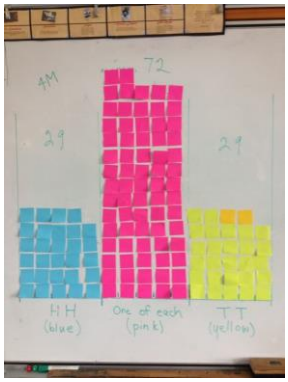


Table 3. Responses to Q10 (Appendix) on the shape of the class graph of outcomes tossing two coins (cf. Figure 2)

Level	Description	Examples	Frequency
2	Explanation of more possibilities/combinations for {H, T}	Because in the middle one there are 2 ways tails and heads/heads tails. Because one of each can count as HT or TH. The centre column has a lot because there is 2 different combinations.	25 (27%)
1	Observation of "more" outcomes, likely, chance	Because when people tossed it lots of people had one of each. It is twice as big because mostly H and T will go together. Because there were two coins, it is most likely one will be heads and one will be tails.	35 (38%)
0	NR, idiosyncratic reason	Well both sides each have a name one is heads and one is tails and so that means it is likely to land on both. I think the centre column is bigger because the person who rolled it might have rolled it longer. I think it is approximately twice the size because it is likely one of the coins will spin less and it is kind of chance and they can be in diff[e]rent order so its like 2 probity [probably] joined together.	31 (34%)

The next question, Q11, was designed to see if students would look at the overall shape (in a proportional sense) of the class plot and their individual plots, rather than details of the numbers involved. Generally students recognized the similarity. The levels of response, example responses, and frequencies are presented in Table 4. Again over 60% of students could express an intuition related to the difference in proportions.

Table 4. Responses to Q11 (Appendix) on the difference/similarity of the class graph and the students' graphs

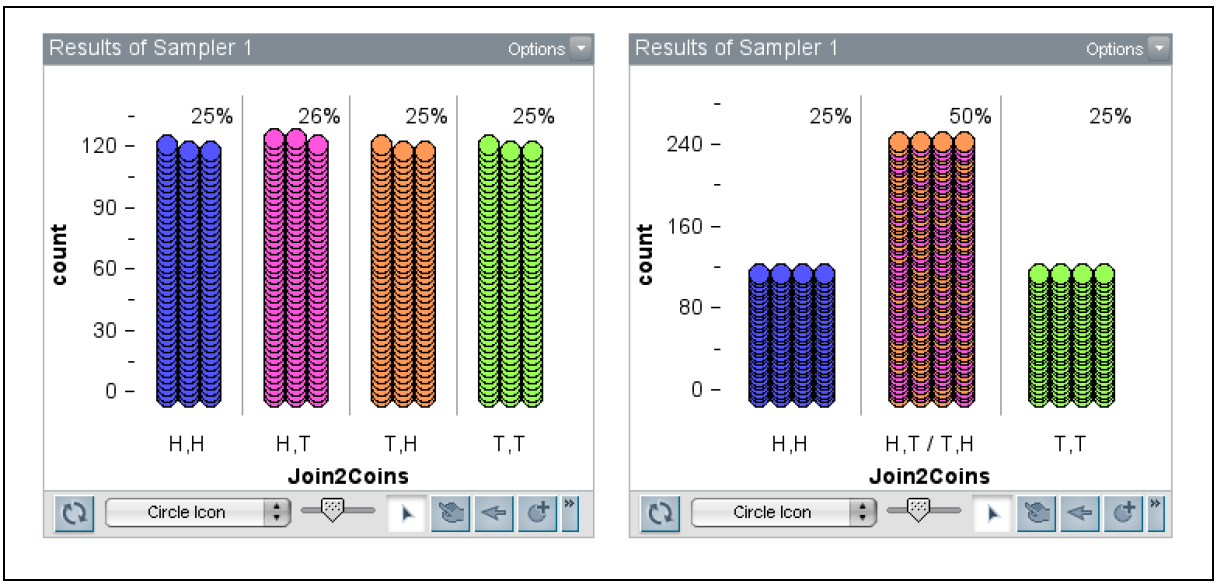
Level	Description	Examples	Frequency
2	Comparison that includes an indication student sees proportions are similar	They both have so much more TH than TT and HH. My one is nearly similar because the HT and TT was the one that I had the most in and HH was the one I had the less.	57 (63%)

1	Basic comparison of numbers	My one is similar to the board's because there is more in the middle. There will be 10 t's of differences because there's lots of people and it's really unlikely for all of us to get the same. Our outcomes were different to the class outcomes because there were more people in the class than me and Nghia. The classes outcomes were spread apart but our graph was only half way.	25 (25%)
0	Unintelligible	Its different because the class is everybodys and with ours the number is different because. That graph shows the whole class's outcome wher[e]as my graph only showed my outcomes and my partners.	11 (12%)

Aim 3: Document the influence of a large number of trials

Students had used the graphing tool of the software *TinkerPlots* (Konold & Miller, 2011) earlier in the year and were familiar with its basic operation. For this activity they were using the Sampler tool, which in this case carried out many simulations of tossing two coins. The students were shown the steps to set up the Sampler and asked to carry out 500 trials and keep track of the number and percentage of each outcome obtained. Not only was the purpose of this part of the activity meant to support the understanding of results approaching 25% for each of two heads and two tails and 50% for one of each, for a much larger number of trials, but also it was intended to reinforce the understanding of the combining of HT and TH from the order tossed by the Sampler. The process of using the Sampler is shown in Figure 3. On the left is the outcome for 500 tosses of two coins showing the four outcomes HH, HT, TH, and TT in four bins. It is then possible to drag and drop one of the icons from the HT bin into the TH bin, resulting in the plot on the right. The plot on the right simulates the same proportions of the outcomes as seen in the class plot (which had only about 120 outcomes). The plots could be labeled with percents to reinforce the relationship to the probabilities of the outcomes.

Figure 3. *TinkerPlots* simulation of 500 tosses of two coins



The questions QT6 to QT11 (Appendix) led students through the above steps asking them to explain what was happening. The question QT6 expected students to respond that the difference between the *TinkerPlots* graph and the class results graph was that there were four columns in *TinkerPlots* (Figure 3, left) and only three in the class graph. Seventy-five percent of students responded appropriately. The other 25% of students responded with other characteristics of the graph, such as “It looks bigger” and “… it’s neater and the icons are circles.”

Question QT8 asked why the graph was more like the class graph after combining two bins for HT and TH (Figure 3, right). Again 75% of students suggested an appropriate reason, with 1/3 of these specifically noting the combining of the two outcomes HT and TH. Still 25% struggled, with no response or responses such as “Not really, this graph is more clustered” and “No because it has two colours instead of one.”

The purpose of QT9 was to give the opportunity for students to mention specifically that the middle column was *twice* the height of the other two. In fact 44% of students did this, whereas 44% made more general comments such as “It looks bigger and fatter.” Only 12% of students did not respond or provided comments that were uninterpretable.

The final question in the workbook (Q12 in the Appendix) moved back from the simulations and asked students to produce a model for tossing two coins and the

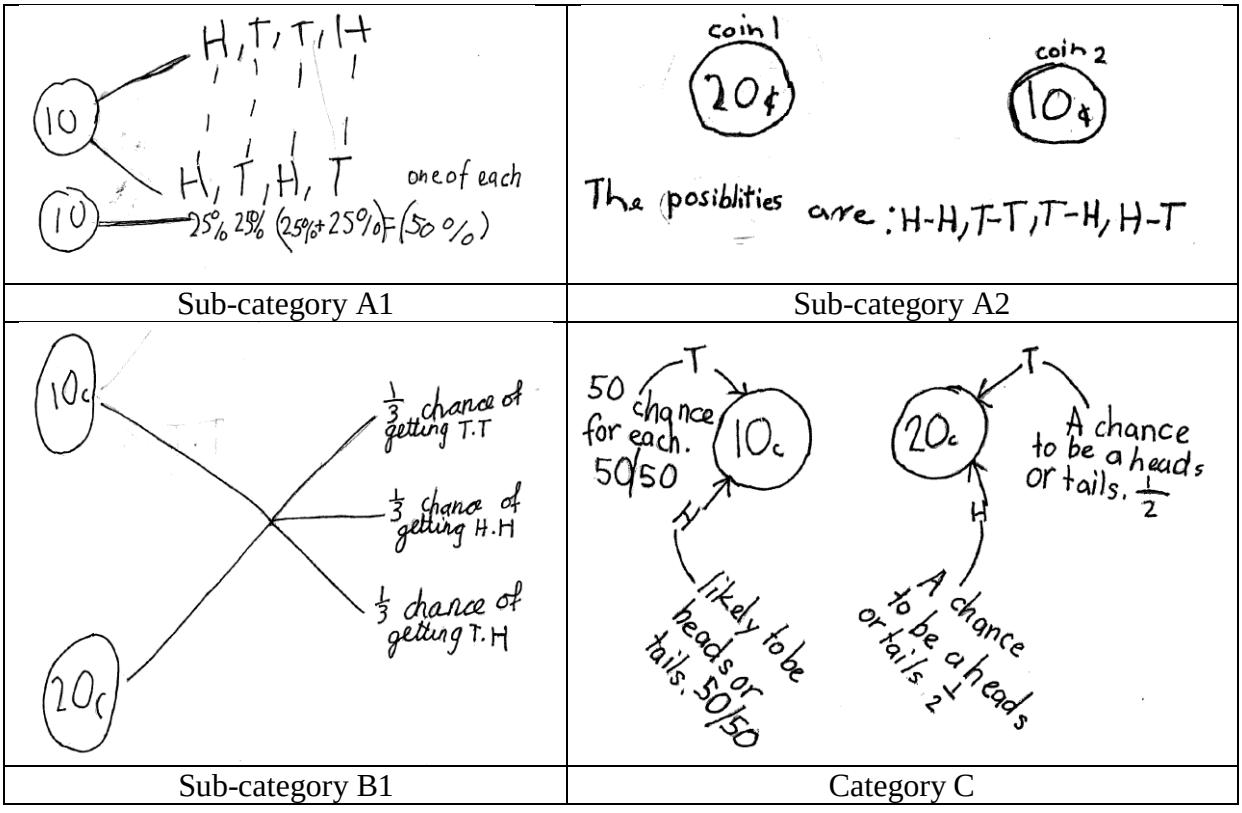
outcomes obtained. As a class with their teachers they had earlier created a model for one coin with arrows and labels of 1/2 for each outcome. For this task, however, there was no assistance and students created their models in their own workbooks. Some pairs of students worked together and produced similar models, whereas other pairs worked as individuals. A wide range of representations was created and for the purpose of this report they have been analyzed as closely as possible to the way that their initial models for question Q1 were analyzed. This allowed for comparison and judgment to be made of improved understanding over the day's activity. The results for the 91 students are shown in Table 5. Categories A1 to A3 and B1 designate the same responses as these categories in Table 2. Category A4, reporting meaningful probabilities (1/4, 1/4, 1/4, 1/4) with no model of coins, was a new group, representing a growth in appreciation of the importance of quarters but not how the probabilities are linked to outcomes. Also categories B2 and B3 from Table 2 were combined into one with either no probability assigned or inappropriate probabilities. There was also a new category, C, that noted responses based on two coins but no indication of outcomes being combined.

Table 5. Final models for outcomes of tossing two coins [Q12 in Appendix]

Category	Description	Sub Category	Description	Frequency	
A	Four outcomes				54
		A1	Distinguishes 4 outcomes with probabilities	26	
		A2	Distinguishes 4 outcomes, no probabilities	15	
		A3	Distinguishes 4 outcomes, incorrect values	4	
		A4	Probabilities of 1/4, not linked to outcomes	9	
B	Three outcomes				6
		B1	Distinguishes 3 outcomes, probabilities 1/3, 1/3, 1/3	3	
		B2/3	Distinguishes 3 outcomes, inappropriate or no probability	3	
C	Two coins with single outcomes but no combinations				22
NA	Idiosyncratic/ No response				9
					91

Figure 4 displays four examples of responses in three of the categories detailed in Table 5. The distinction between sub-categories A1 and A2 was in the lack of labeled probabilities in the A2 group. The sub-category B1 models had not progressed since the beginning of this part of the activity. The category C responses were clear in modeling the first stage for each coin but were unable to combine outcomes meaningfully. Although it was disappointing that 24% of students could start a model with two coins but not combine outcomes, it was encouraging that 59% appeared to appreciate to some extent the existence of four outcomes, with nearly half of these completing an appropriate detailed representation. Only 7% of students continued to create models of three outcomes.

Figure 4. Examples of student models for outcomes of tossing two coins.



Sub-category A1 | Sub-category A2

Sub-category B1 | Category C

During the student group interviews conducted at the end of the year, students' further development of three core understandings was explored, namely, (a) the variation in outcomes when two coins are repeatedly tossed 12 times; (b) the four possible outcomes when tossing two coins, with the probability of obtaining a head and a tail being twice that of two heads or two tails, and (c) predictions becoming closer to actual outcomes with increasing numbers of trials. The interviewer commenced the

interview with the *TinkerPlots* Sampler initially set up to toss two coins just once. It was explained, “I have here a Sampler like we set up to toss 2 coins. You can see it has been Run once. Can you tell me what is happening as we run it 10 times?” (The interviewer changed the Run to 10 and on medium speed to enable the students to explain their observations. The interviewer then expanded the plot to show the results.) The students were asked, “Will it (the plot) look the same if we run the Sampler again?” (The interviewer ran the Sampler a few more times for the students). Their responses to this question (pertaining to understanding (a) above) indicated an appreciation of variation in outcomes, with explanations including, “I don't think so because the first sampler won’t influence the other and pretty much there’s no way to be certain;” “No, because the chance is just really random;” “Because it might not be the same because there’s more chances of others coming, other results coming;” and “Possibly not because like every time you flip it, it’s like you’re flipping it for the first time so the thing before it has nothing to do with the next one.”

Next, the interviewer asked, “Can you remember when we talk about the expected outcome (of tossing two coins once)? We can talk about using a fraction or we can talk about it using a percentage. Can you tell me what those values would be, what those fractions would be or what those percentages would be?” (pertaining to understanding (b) above). For some students, it was necessary for the interviewer to explain the question, such as, “Say, what would be the expected outcome for getting HH? Tell me a fraction or a percentage.” Students’ responses, including the following, indicated an understanding of the four possible outcomes.

> Student A: (for HH) “25%, and ¼”; (for TT) “round about the same as for HH”; (and what about for HT or TH?) “Around 50%”; “yeah, there would be much higher cause you know, joined.”
>
> Student B: “Relevantly the HT and TH, if you combine the both together they will be about half of the thing.”
>
> Student C: (Can you remember what we expect to get if we toss the coins?) “Either Heads or Tails;” (Can you remember our four possible outcomes down the bottom here?) “TT, HH, HT and TH” (… and what fraction is each of those?) “Quarters.”

On further questioning whether each outcome is ¼, Student C agreed and explained that this represents 25%. Student D further commented, “Either 50% … like either,

each of one, but like you combine both of them together for 50%." When asked which two outcomes would be combined, the student indicated that it was the HT and the TH. Further questioning on what percent would be expected for the HH or the TT elicited the response, "HH 25 and joined together (HT/TH) 50 and TT 25."

Students' responses to understanding (c) regarding the number of times the Sampler might need to run for the predictions to become closer to actual outcomes were varied, with students suggesting "100" and some, "near 100,000." A couple of student groups, however, felt they could not provide an estimate because they considered the variation in outcomes made such a prediction difficult. Their responses suggested an understanding of the relationship between empirical estimates and theoretical probabilities was still developing: "We can't tell, it would be hard to tell how many we'd have to do it cause it's the first time every time so we'd just have to keep doing it till we get the 25%;" and "Um, well it could take as many as like, as many as possible because they could be all different every time ... I'm not sure exactly how many times it would take."

DISCUSSION

First, the outcomes of the study are considered in terms of the aims set. Instead of the traditional teaching of probability outcomes, perhaps through introducing tree diagrams, students were asked to make predictions of the probabilities for all outcomes from tossing two coins, expressing Expectation, and then use trials to confirm their predictions, experiencing Variation in the process. It was imagined by the authors that this would create cognitive conflict for many of the students. Further it was intended that this conflict would be resolved by increasing the number of trials and watching the proportion of outcomes stabilize at the true theoretical values. Without "pretest" knowledge of the students' understanding, the authors based their supposition that many students would predict three equally likely outcomes on anecdotal evidence from work with mathematics teachers who had not been previously introduced to the context. By asking students to describe the chances of all possible outcomes, the plan was to give students a structure to create rather than focus only on the chance of

tossing only “a head and a tail,” as for example in Rubel’s (2007) study. This approach may have contributed to taking students away from justifying a single outcome with “anything can happen,” and did result in students accounting for all cases they could imagine. Two-thirds of students imagined three outcomes including “a head and a tail” without distinguishing how this might occur. The stage was hence set for providing conflict for this model by completing repeated trials, first by the students themselves, and then using technology.

The second aim hence was to document the students’ pathways through the trials and their developing understanding of the reducing variation as trials increased and of the proportion of outcomes for “heads and tails” approaching 1/2 of the total and being visually twice that of the proportion of two heads or two tails. The explanations of students became more sophisticated but not all students could reach a high level of language use. This may have been due to the high proportion of ESL students in the classes. In moving from the trials to an abstract model, 90% of students appeared from their representations to be able to visualize the purpose of the activity although 24% could not take this further to combining outcomes from the two single coins. Fifty-nine percent, however, could at least partially complete the process for fair outcomes, often leading to combining HT and TH for a probability of 1/2. In these cases there was no question of 1/2 or 50/50 for “heads and tails” being a random outcome because probabilities for two heads and two tails were also assigned.

The third aim of producing a final model for two coins was intended to link the experimental outcomes to the theoretical model, that is “building” the model as reasonable based on experience rather than just “being told” the theoretical answer. There was no control group in this study but research of others (e. g., RUBEL, 2010; WATSON; CALLINGHAM, 2013) demonstrates some of the confusions that can result when students have not experienced increasing numbers of trials to confirm or dispute initial beliefs. Rubel found that many students who were asked which was more likely, getting 7 heads on 10 tosses of a coin or getting 700 heads on 1000 tosses of a coin, used a mathematical explanation based on the equality of ratios, 7/10 = 700/100, to claim the two events were equally likely. A similar question asked of 247 students by Watson and Callingham based on getting more than 60% heads when either 10 or 30 coins were tossed, resulted in 50% of students claiming the chances were equal. Some

reasoning was the same as that of Rubel's students, whereas other explanations included "60% is independent of sample size," meaning the same thing for both settings.

The context for this study, tossing only two coins and not asking for a comparison of probabilities, is much more basic than those used by Rubel (2010) and Watson and Callingham (2013). The difficulties of older students with these more complex problems suggest that interventions such as those used in this study should begin early, when elementary probabilities are being introduced. Although not all Grade 4 students in this study developed sophisticated ideas, there is confidence that most came to appreciate the variation shown with their smaller sample sizes and the closer approach to the expected proportions with larger sample sizes. It is now possible with software such as *TinkerPlots* to simulate increasingly complex probability problems that students meet across the school years; Konold (1994), for example, gave an excellent example of this in considering families with one son and the overall balance of gender in families. If students become familiar with testing their hypothesized theoretical answers with simulations, they may themselves suggest carrying out trials like those of Watson (2007) in a similar situation. Eventually the importance of sample size may become so embedded that in some situations simulations are no longer needed.

Further evidence to support the authors' claim that opportunities such as those provided in this study should begin early is found in the recent work of Pfannkuch, Arnold, and Wild (2014) who were following Grade 11 students' progress in understanding sampling variability and its relationship to drawing informal inferences. Although taking place in a much more advanced setting with much older students, the students were still dealing with issues of sample size and the variability observed in their samples. In their conclusion Pfannkuch *et al*. (2014) conjectured "that learning to understand and reason about sampling variability will take many years for students to develop the complex network of concepts that underpin statistical thinking.". The current authors suggest that Grade 4 is a reasonable place to start, using basic contexts such as tossing two coins and providing for development across the middle years of schooling with other contexts than probability outcomes.

Although not a specific aim of the study, another outcome for most students was the increasing awareness of the relationship of frequency counts and percentages. As discussed in Watson and English (2013), when tracking outcomes for repeated large sample sizes, plotting frequencies made the variation appear to increase, whereas using percentages, the decreasing variation with increasing sample size was easily seen on plots with the same scale. The part-whole relationship of the number of desired outcomes and the total number of trials linked fractions (often complex and non-intuitive, e.g., 134/500) to percentages that were easily compared across sample sizes (e.g., 27%).

CONCLUSION

Graham Jones in his 2005 reflections on research into children's reasoning about probability at both the elementary and middle school levels included a comment on "limitations with respect to students' thinking about experimental probability and the connection between experimental and theoretical probability" (JONES, 2005, p. 369) that is, "the frequentist approach to probability" (JONES, 2005, p. 368). It is the experimental approach that was the focus of this study, acknowledging the fundamental link to variation, the concept underlying all of statistics. As seen in some of the earlier work on the probability of events such as tossing two coins (e. g., CARPENTER *et al*., 1981), the expectation of a correct answer was based on the understanding of the theoretical sample space underlying the problem. Explanations for incorrect answers were based on misinterpretations of the idea of "50-50." Although students were expected to "know" the answer, it had to be learned from text explaining the theory rather than from physical experience of seeing a pattern develop through many trials. As this study has shown, students require visual evidence of how empirical relative frequency approaches theoretical probability with decreasing variation in large numbers of trials. Such evidence is difficult to produce without the affordances provided by the Sampler in the *TinkerPlots* program; the inclusion of this technology served as a powerful facilitator of this understanding.

The results of this study point to reinforcing curriculum documents, including *The Australian Curriculum: Mathematics* (ACARA, 2013), where activities with

simulations and increasing sample size of the type addressed here are recommended. The idea of asking students to propose a model for tossing two dice, based only on the experience of one die, and then allowing them to test their model with trials was not part of the curriculum 35 years ago. The cognitive conflict, as displayed in this study, is surely one of the features of student experience that should ensure students remember the "answer." This is later tied to the theoretical model, for example through repeated sampling and the law of large numbers (IRELAND; WATSON, 2009). It is hoped that the outcomes of this study will encourage teachers to implement this type of activity to reinforce the understanding expressed formally in curriculum documents.

ACKNOWLEDGEMENTS

This report arose from a research study supported by a Discovery Grant (DP120100158) from the Australian Research Council (ARC). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the ARC. We wish to acknowledge the enthusiastic participation of the classroom teachers and their students, as well as the excellent support provided by our senior research assistant, Jo Macri.

REFERENCES

AMIR, G.S.; WILLIAMS, J.S. Cultural influences on children's probabilistic thinking. **Journal of Mathematical Behavior**, v.18, p.85-107, 1999.

ACARA - Australian Curriculum, Assessment and Reporting Authority. **The Australian curriculum***:* Mathematics, Version *5.0, 20 May 2013.* Sydney: ACARA, 2013.

AUSTRALIAN EDUCATION COUNCIL. **A national statement on mathematics for Australian school***s.* Melbourne: Author, 1991.

AUSTRALIAN EDUCATION COUNCIL. **Mathematics**: a curriculum profile for Australian schools. Melbourne: Curriculum Corporation, 1994.

CARPENTER, T. *et al.* What are the chances of your students knowing probability? **Mathematics Teacher**, v.74, p.342-344, 1981.

CAMPOS, T.M.M.; CAZORLA, I.M.; KATAOKA, V.Y. Statistics school curricula in Brazil. In: BATANERO, C.; BURRILL, G.; READING, C. (Ed.). **Teaching statistics in school mathematics – Challenges for teaching and teacher education**: a joint ICMI/IASE study Dordrecht, The Netherlands: Springer, 2011, p.5-8.

COBB, P.; JACKSON, K.; MUNOZ, C. Design research: acritical analysis. In: ENGLISH, L.D.; KIRSHNER, D. (Ed.). **Handbook of international research in mathematics education**. New York: Routledge, 2013.

COMMON CORE STATE STANDARDS INITIATIVE. **Common Core State Standards for Mathematics.** Washington: National Governors Association for Best Practices and the Council of Chief State School Officers. 2010. Available at: http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

ENGLISH, L.; WATSON, J. Development of fourth-grade students' understanding of experimental and theoretical probability. In: ANDERSON, J.; CAVANAGH, M.; PRESCOTT, A. (Ed.). **Curriculum in focus:** *research guided practice.* In: PROCEEDINGS OF THE 37TH ANNUAL CONFERENCE OF THE MATHEMATICS EDUCATION RESEARCH GROUP OF AUSTRALIA, Sydney, 2014, p.14-222.

ENGLISH, L.; WATSON, J. (2015). Exploring variation in measurement as a foundation for statistical thinking in the elementary school. **International Journal of STEM Educ***ation*, v.*2, n.*3, 2015. doi 10.1186/s40594-015-0016-x

ENGLISH, L.D.; WATSON, J.M. Statistical literacy in the elementary school: Opportunities for problem posing. In: SINGER, F.; ELLERTON, N.; CAI, J. (Ed.). **Problem posing***:* from research to effective practice*.* Dordrecht: Springer, 2015.

FISCHBEIN, E.; SCHNARCH, D. The evolution with age of probabilistic, intuitively based misconceptions. **Journal for Research in Mathematical Education***,* v.28, p.96-105, 1997.

HERNANDEZ, H. *et al*. (2014). Investigation about curricular orientations in teaching statistics in Brazil and Mexico. In: K. MAKAR, K.; DE SOUSA, B.; GOULD, R. (Ed.), *Proceedings of the 9th International Conference on the Teaching of Statistics, Flagstaff, Arizona*. Voorburg, The Netherlands: International Statistical Institute. 2014. Available at: http://icots.info/9/proceedings/pdfs/ICOTS9_ C155_HERNANDEZ.pdf

IRELAND, S.; WATSON, J. Building an understanding of the connection between experimental and theoretical aspects of probability. **International Electronic Journal of Mathematics Education**, v.4*,* p.339-370, 2009.

JONES, G.A. (Ed.). **Exploring probability in school**: challenges for teaching and learning*.* New York: Springer, 2005.

JONES, G.A.; LANGRALL, C.W.; MOONEY, E.S. (2007). Research in probability: responding to classroom realities. In: LESTER JUNIOR, F.K. (Ed.). **Second handbook of research on mathematics teaching and learning**. Charlotte: Information Age, 2007, p.909-956.

KAHNEMAN, D.; TVERSKY, A. Subjective probability: a judgement of representativeness. **Cognitive Psychology**, v.*3*, p.430-454, 1972.

KELLY, A.E.; LESH, R.A.; BAEK, J.Y. (Ed.). **Handbook of design research methods in education***.* New York: Routledge, 2008.

KONOLD, C. Teaching probability through modeling real problems. **Mathematics Teacher**, v.87, n.4, p.232-235, 1994.

KONOLD, C.; MILLER, C. **ProbSim.** [Computer software]. Amherst: University of Massachusetts, 1994

KONOLD, C.; MILLER, C.D. **TinkerPlots:** dynamic data exploration [computer software, Version 2.0]. Emeryville, CA: Key Curriculum, 2011.

KONOLD, C.; POLLATSEK, A. Data analysis as the search for signals in noisy processes. **Journal for Research in Mathematics Education**, v.33*,* p.259-289, 2002.

KONOLD, C. *et al*. Inconsistencies in students' reasoning about probability. **Journal for Research in Mathematics Education***,* v.24*,* p.392-414, 1993.

LECOUTRE, M-P. Cognitive models and problem spaces in "purely random" situations. **Educational Studies in Mathematics**, v.23, p.557-568, 1992.

MAKAR, K.; RUBIN, A. A framework for thinking about informal statistical inference. **Statistics Education Research Journal**, v.8, n.1, p.82-105, 2009.

Ministry of Education. **Mathematics in the New Zealand curriculum**. Wellington: Author, 1992.

MORITZ, J.B.; WATSON, J.M. Reasoning and expressing probability in students' judgements of coin tossing. In: BANA, J.; CHAPMAN, A. (Ed.). **Mathematics education beyond 2000**. Perth, WA: MERGA, 2000, p.448-455.

NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS. **Principles and standards for school mathematics.** Reston: Author, 2000.

PFANNKUCH, M.; ARNOLD, P.; WILD, C. (2014). What I see is not quite the way it is: students' emergent reasoning about sampling variability? **Educational Studies in Mathematics**, 2014. doi: 10.1007/s10649-014-9539-1.

PRATT, D. Making sense of the total of two dice. **Journal for Research in Mathematics Education,** v.31, p.602-625, 2000.

RUBEL, L.H. Students' probabilistic thinking revealed: the case of coin tosses. In: BURRILL, G.F. (Ed.). **Thinking and reasoning with data and chance. Sixty-eighth Yearbook**. Reston: National Council of Teachers of Mathematics, 2006, p.49-59.

RUBEL, L.H. Middle school and high school students' probabilistic reasoning on coin tasks. **Journal for Research in Mathematics Education**, v.38, p.531-556, 2007.

RUBEL, L.H. Connecting research to teaching: is 7/10 always equivalent to 700/1000? **Mathematics Teacher**, v.104, p.144-146, 2010.

SHAUGHNESSY, J.M. (1992). Research in probability and statistics: reflections and directions. In: GROUWS, D.A. (Ed.). **Handbook of research on mathematics teaching and learning**. New York: National Council of Teachers of Mathematics & MacMillan, 1992, p.465-494.

TARR, J.E. Confounding effects of the phrase "50-50 chance" in making conditional probability judgments. **Focus on Learning Problems in Mathematics**, v.24, n.4, p.35-53, 2002.

WATSON, J.M. Variation and expectation as foundations for the chance and data curriculum. In: CLARKSON, P. (Ed.). **Building connections**: theory, research and practice Sydney: MERGA, 2005, p.35-42..

WATSON, J.M. **Statistical literacy at school**: growth and goals. Mahwah: Lawrence Erlbaum, 2006.

WATSON, J.M. The use of technology for creating cognitive conflict in mathematics. In: J. SIGAFOOS, J.; Green, V.A. (Ed.), **Technology and teaching**. New York: Nova Science Publishers, 2007, p.65-74.

WATSON, J.; CALLINGHAM, R. Likelihood and sample size: the understandings of students and their teachers. **Journal of Mathematical Behavior**, v.32, p.660-672, 2013.

WATSON, J.; ENGLISH, L. The power of percent. **Australian Primary Mathematics Classroom**, v.18, n.4, p.14-18, 2013.

Submitted: February, 2014
Accepted: July, 2014

Appendix: Tossing two coins (Workbook Extracts)

Q1. Now think about tossing two coins at the same time. List the **outcomes** that could occur in this case. What is the chance of each outcome? Write a fraction for each outcome.

This time you and your partner are going to toss 2 coins 12 times and record the outcome for each toss.

Q3. **Fill in** the missing numbers below:

I predict we will toss two Heads _____ time/s.

I predict we will throw two Tails _____ time/s.

I predict we will throw one Head and one Tail _____ time/s.

Q4. Trial (tick the box for each toss)

Toss

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
HH												
One of each												
TT												

Q7. **Graph** your results below.

The next two questions relate to the whole class plot that the teacher prepared on the board combining data from the groups.

Q10. **Explain** why the centre column (One of each) is approximately twice the size of the HH or TT columns.

Q11. Write a sentence **describing** the difference between the class outcomes and your own outcomes.

Students are given instructions to collect 500 tosses of two coins and create a plot of the outcomes.

QT6. Is the plot **different** from the one we created for the class results? Why/why not?

QT7. Click on an Icon in the HT column/bin and drag it into the TH column/bin. This will combine these two columns/bins into one, leaving you with three columns/bins.

QT8. Is this more like the graph drawn in class? **Why**?

QT9. What do you **observe** about the size of the combined HT / TH column compared to the HH and TT columns?

QT10. Run the Sampler again to get another set of 500 tosses of the two coins. **Discuss** any variation in your results with your group.

QT11. Run the Sampler for 5 x 500 Repeats, recording your results in the table below. You can add more Repeats of larger numbers in the last 5 rows.

Number of Repeats	HH		One of each		TT	
	N	%	N	%	N	%
500						
500						
500						
500						
500						

On the last page of the Workbook

Q12. Use the space below to create a **model** of all of the possibilities for outcomes when two coins are tossed. To help distinguish the coins call them Coin 1 and Coin 2, or assume one coin is 10¢ and the other is 20¢ (label them accordingly). Ensure you **explain** how your model works.