Andrew Zieffler
Elizabeth Fry

Reasoning about Uncertainty

Learning and Teaching Informal Inferential Reasoning

REASONING ABOUT UNCERTAINTY

REASONING ABOUT UNCERTAINTY

Learning and Teaching Informal Inferential Reasoning

CATALYST PRESS

A CATALYST PRESS PUBLICATION

Published by Catalyst Press, Minneapolis, Minnesota.

Library of Congress Cataloging-in-Publication Data:

Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning / A. Zieffler and E. Fry (Eds.).
"Catalyst Press."
ISBN 978-0692491645 (pbk.)

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTRIBUTORS

DANI BEN-ZVI, The University of Haifa, Israel
HANA MANOR BRAHAM, The University of Haifa, Israel
DANIEL FRISCHEMEIER, Institute of Mathematics, University of Paderborn
JILL FIELDING-WELLS, University of Tasmania, Australia
KATIE MAKAR, The University of Queensland, Australia
LUCIA ZAPATA-CARDONA, Universidad de Antioquia, Colombia
MAXINE PFANNKUCH, The University of Auckland, New Zealand
PIP ARNOLD, Cognition Education Limited and The University of Auckland, New Zealand
ROLF BIEHLER, Institute of Mathematics, University of Paderborn
SIBEL KAZAK, Pamukkale Üniversitesi, Turkey
STEPHANIE BUDGETT, The University of Auckland, New Zealand
SUSANNE PODWORNY, Institute of Mathematics, University of Paderborn

CONTENTS IN BRIEF

CONTENTS

FOREWORD

Every scientific discipline needs a solid research base on which to build theory and areas of inquiry. In education, this research base is also crucial in that it provides a foundation for understanding and improving teaching and learning. Unlike other disciplines, such as mathematics education or science education, the research related to statistics spans many areas of scholarship and teaching, such as mathematics, psychology, and science. Synthesizing research from these different areas has been challenging, due to different terminology, focus, methodology, and target population.

At the *Fourth International Conference on Teaching Statistics* (ICOTS-4), held in 1998 in Singapore, it became clear that there was a need to form a coherent research community to allow scholars studying the teaching and learning of statistics to not only share and discuss their work in detail, but also explore ways to coordinate their research discourse, questions and methods. At that time there were two major challenges to these goals. First, it was difficult to gather researchers from around the world except for ICOTS, which was held only every four years, and even at ICOTS, there was limited time to present and almost no time to discuss research in a deep way. The second challenge had to do with the way researchers used terms to describe important student learning reasoning and outcomes. In particular, there seemed to be no agreement on the definitions of statistical literacy, reasoning and thinking, despite the growing interest in research studying or assessing each of these learning outcomes.

As we recognized this need and tried to come up with possible solutions, the idea for SRTL was born. We wondered: If we held a small, invitational research forum and invited our colleagues to come, independent of a professional organization or conference, but offering the opportunity to share rich segments of video recorded interviews and observations of students, would anyone come? At that time, Dani lived on Kibbutz Be'eri in the south of Israel, and he offered his kibbutz as a place for us to host such a gathering. We obtained a type of endorsement from the informal *International Study Group for the Teaching and Learning of Probability and Statistics*, then chaired by Carmen Batanero, at the University of Granada, Spain. Dani was also able to get support from the Weizmann Institute of Science, where he had been working for several years, and Joan was able to secure a small amount of funding from her department chair (Mary McEvoy) at the University of Minnesota. We sent out an invitation, and held our breath. Would anyone come?

Luckily, the answer was yes. In the summer of 1999, we had 16 scholars from Australia, Belgium, Ireland, Israel, the UK, and the US come to the kibbutz to meet for an intense five days of presentations and discussions. Along with the research we were immersed in, Dani provided outings to enable us to get to know his country, its history and diversity. By the end of SRTL-1, we were an enthusiastic group that saw not only the enormous opportunities and challenges ahead of us, but also the joy of forming a research community with a shared passion for learning how students come to understand and learn statistical concepts and methods. Today, the *International Collaboration for Research in Statistical Reasoning, Thinking, and Literacy* offers scientific gatherings for statistics education researchers every two years. The SRTL research forums, foster collaborative and innovative research studies that examine the nature and development of statistical literacy, reasoning, and thinking, and to explore how educators can develop these desired learning goals for students.

The SRTL research forums have led to many publications that present new research, synthesize and build on previous research, and form connections among related work in other disciplines (see table below).

SRTL Forums and Contributions to Statistics Education

Forum	Theme	Host/Venue	Date	Primary Publication(s)
SRTL-1	Statistical reasoning, thinking, and literacy	Kibbutz Be'eri, Israel	July 18–23, 1999	
SRTL-2	The challenges in describing, teaching, and assessing statistical reasoning, thinking, and literacy	University of New England, Armidale, Australia	August 15–20, 2001	Ben-Zvi, D., & Garfield, J. (Eds.) (2004). *Challenges in developing statistical reasoning, thinking and literacy.* The Netherlands: Kluwer Publishers.

Table continued from previous page

Forum	Theme	Host/Venue	Date	Primary Publication(s)
SRTL-3	Reasoning about variability	The University of Nebraska-Lincoln, USA	July 23–28, 2003	Garfield, J., & Ben-Zvi, D. (Eds.) (2004). Research on reasoning about variability [Special issue]. *Statistics Education Research Journal, 3*(2). Garfield, J., & Ben-Zvi, D. (Eds.) (2005). Research on reasoning about variability [Special issue]. *Statistics Education Research Journal, 4*(1).
SRTL-4	Reasoning about distribution	University of Auckland, New Zealand	July 2–7, 2005	Pfannkuch, M., & Reading, C. (Eds.) (2006). Research on reasoning about distribution [Special issue]. *Statistics Education Research Journal, 5*(2).
SRTL-5	Reasoning about statistical inference: Innovative ways of connecting chance and data	University of Warwick, UK	August 11–17, 2007	Pratt, D., & Ainley, J. (Eds.) (2008). Informal inferential reasoning [Special issue.] *Statistics Education Research Journal, 7*(2).
SRTL-6	The role of context and evidence in informal inferential reasoning	The University of Queensland, Brisbane, Australia	July 10–16, 2009	Makar, K., & Ben-Zvi, D. (Eds.) (2011). The role of context in developing reasoning about informal statistical inference. [Special issue]. *Mathematical Thinking and Learning, 13*(1–2).
SRTL-7	New approaches to developing reasoning about samples and sampling in informal statistical inference	Utrecht University, The Netherlands; Texel Island	July 17–23, 2011	Ben-Zvi, D., Bakker, A., & Makar, K. (Eds.) (2015). Statistical reasoning: Learning to reason from samples [Special issue]. *Educational Studies in Mathematics, 88*(3).
SRTL-8	Reasoning about uncertainty in the context of making informal statistical inferences	University of Minnesota, USA; Two Harbors, MN	August 18–24, 2013	Zieffler, A., & Fry, E. (Eds.). *Reasoning about uncertainty: Learning and teaching informal inferential reasoning.* Minneapolis, MN: Catalyst Press.

Table compiled by Elizabeth Fry

The SRTL research forums have unique features, such as a small size (no more than 25 participants) that allows time for in-depth presentation and discussion of research. There is extensive use of videos to present how students solve problems and reason about statistical information in classrooms or during interviews. Most forums

have included at least one statistician in addition to the educational researchers in order to provide the perspective of the discipline and to give feedback on the research presented. Participants present, discuss and argue about research related to these topics in a format that facilitates becoming acquainted with key researchers and viewing their work in progress. After many SRTL gatherings, since that first one in 1999 in Israel, we hosted the eighth SRTL forum in Minnesota, in the summer of 2013. The theme of that gathering, *Reasoning about Uncertainty in the Context of Making Informal Statistical Inferences*, continued the theme of informal inferential reasoning, but with a special focus on the idea of uncertainty. The papers from that exciting and productive gathering form the chapters of this book.

The eighth SRTL forum built on and expanded the work discussed at previous SRTL gatherings. Recent research on informal inferential reasoning (IIR) suggested it was important to study further ideas and pedagogical approaches related to uncertainty and confidence in the context of reasoning about informal statistical inferences. Assessing confidence about uncertain phenomena and understanding ideas of uncertainty are essential components in making predictions and making judgments about the reasonableness of patterns and trends identified in data. This topic is relevant and important at all levels of schooling, even in the early years. Furthermore, recent developments in statistics educational technology (e.g., TinkerPlots™) can support not only exploratory data analysis approaches to learning IIR but also experimentation with ideas of uncertainty and statistical/probabilistic models as generators of data, modeling and simulations. These developments provided new stimulus for growth in the rethinking and study of the role of uncertainty in helping students develop statistical reasoning.

We are indebted to the co-editors, Andrew Zieffler and Elizabeth Fry, who coordinated the review of the papers, the editing and the formatting of the book. Despite the tendency of SRTL authors to miss or extend deadlines, we are impressed that this volume is going to print at about the same time the ninth SRTL begins in Germany, in 2015. We are also grateful for Dr. Katie Makar, who has taken the reins passed on by Joan, as she moves into her retirement. We know SRTL will be enriched by the thoughtfulness, insights, and high standards Katie brings to this role.

JOAN GARFIELD AND DANI BEN-ZVI
June 2015

PREFACE

The research presented in this volume is the culmination of a two-year process that began in August 2013. That summer, 26 statistics education researchers met in Two Harbors, Minnesota for the *Eighth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (SRTL-8). The theme for the forum was reasoning about uncertainty in the context of making informal statistical inferences.

Over a period of seven days, this group of researchers presented, discussed, and examined research related to the forum's theme of reasoning about uncertainty in the context of making informal statistical inferences. The research at SRTL-8 covered many different facets (e.g., use of technology and students' classroom articulation) and explored several different populations of students (primary, secondary, and tertiary levels), adults, and even teachers of statistics. The chapters in this book constitute a subset of that research, and reflect the variation in both the populations and topics studied.

Chapter 1: Jill Fielding-Wells and Katie Makar investigated 7–8 year-old students' inferential reasoning under uncertainty, using an inquiry-based unit developed around a game of "addition bingo".

Chapter 2: Sibel Kazak studied the use of TinkerPlots™ simulation tools and dialogic talk in small groups to support 10–11 year-old students' articulation of uncertainty in making informal inferences.

Chapter 3: Hana Manor Braham and Dani Ben-Zvi analyzed 13 year old students' articulations of uncertainty during their first steps in exploring sampling distributions in a TinkerPlots™ inquiry-based learning environment.

Chapter 4: Maxine Pfannkuch, Stephanie Budgett, and Pip Arnold explored university students' and workplace volunteers' reasoning processes as they drew *experiment-to-causation* inferences using the randomization test.

Chapter 5: Rolf Biehler, Daniel Frischemeier, and Susanne Podworny investigated the reasoning of preservice teachers about uncertainty in the context of randomization tests facilitated by TinkerPlots™.

Chapter 6: Lucia Zapata-Cardona studied the ideas of uncertainty held by statistics teachers while they worked on professional development activities designed to promote informal inferential reasoning.

The creation of a book, especially an edited volume, takes a village, and we would like to acknowledge the countless hours and volunteers that went into this one. First, the principal credit goes to, of course, the authors and researchers whose work lie at the heart of this book. They also all served as reviewers for other chapters, and their reflections, suggestions, and perspectives were critical during the reviewing process.

We also need to acknowledge the efforts and contributions of both Joan Garfield and Dani Ben-Zvi. As the progenitors of the *International Research Forums on Statistical Reasoning, Thinking, and Literacy*, their experience and encouragement were instrumental throughout this two-year process. Finally, we want to thank both the American Statistical Association and Springer for providing some financial assistance to the participants of SRTL-8 where the research appearing in this book was originally presented.

ANDREW ZIEFFLER AND ELIZABETH FRY
July 2015

CHAPTER 1

INFERRING TO A MODEL: USING INQUIRY-BASED ARGUMENTATION TO CHALLENGE YOUNG CHILDREN'S EXPECTATIONS OF EQUALLY LIKELY OUTCOMES

Jill Fielding-Wells[1] and Katie Makar[2]

[1]University of Tasmania, Australia
[2]The University of Queensland, Australia

Abstract

Children's informal reasoning about uncertainty can be considered a product of their beliefs, language, and experiences, much of which is formed outside of formal schooling. As a result, students can adopt informal intuitions that are incompatible with formal reasoning. Although the creation of cognitive conflict has been considered as one means of challenging students' understandings, prior research in probability suggests that students may simultaneously hold multiple, incompatible understandings without conflict arising. Design-based methodology was adopted to investigate young (7–8 years old) students' inferential reasoning under uncertainty, using an inquiry-based unit developed around addition bingo. This paper selectively reports on students' inferences that initially suggested they were tacitly working from a uniform distribution (equiprobability bias), but shifted as students collected empirical data (from a discrete symmetric triangular distribution). Their inferences were challenged using an argumentation framework, with particular emphasis on the need for defensible evidence. Initial findings suggest potential for argumentation and inferential approaches that make students' conceptions explicit through 'visibilizing' their knowledge.

Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning.
By A. Zieffler and E. Fry (Eds.)

Keywords: Informal statistical inference; Equiprobability; Mathematical inquiry; Argumentation; Statistical modeling; Early years mathematics

1.1 Overview

Do students envision a model when they make inferences from a probabilistic situation? While they probably do not "see" a distributional model as a statistician might, students likely have some implicit model that is used to judge situations. These implicit models are quite limited as students have primarily experienced probabilistic contexts that have either equally likely outcomes (e.g., dice, that would be modeled by a uniform distribution) or fixed proportional outcomes based on categorical data (colored lollies in a jar or spinners, that could be modeled by a bar graph).

In this study, young children (aged 7–8) used inquiry-based argumentation practices to generate and revise inferences as they were tested against experimental data. The aim of the research was to use a focus on evidence to challenge and shift their early inferential models, grounded in a uniform (equiprobability) model, toward a model which aligned with their empirical data (triangular probability distribution). In the context of trying to design and provide evidence of the "best" card that could win at addition bingo, the students encountered conflicts between their expectations (constructed samples based on inferred models) and the outcomes of the game (empirically-generated samples). The data in this paper tell a story of how young children's inferential models and beliefs about randomness became more sophisticated through a focus on inquiry-based argumentation practices as they wrestled with the empirical results of the game.

1.2 Problem

Uncertainty is encountered in everyday contexts, and the likelihood of events can often only be estimated based on previous experiences. For example, prior experience helps us estimate the likelihood that traffic flow will enable us to get to the store before it closes, or if a friend will be late for our meeting. Our response to these uncertainties is to make an inference, or prediction, based on data and expressed with uncertainty (Makar & Rubin, 2009). However, when the experiences we have had early in life do not necessarily provide useful probabilistic conceptions on which to base such experiences, we can develop biases or weakened understanding. Investigating inferential reasoning in relation to probabilistic models may be one way for researchers (and students) to identify and challenge unproductive reasoning.

Although reasoning about uncertainty at the primary level is typically part of the probability strand of the Australian curriculum (Australian Curriculum Assessment and Reporting Authority, 2014), its content focus is often on the language of probability (impossible, possible, likely, certain) and experience calculating simple probabilities using random devices (e.g., coins, spinners, dice) with known (and knowable) probabilities that are characterized by equiprobable outcomes. However, most

experiences in real life are not equiprobable—there are not equal chances of rain or no rain each day—and very often the theoretical outcomes cannot be clearly determined. Equiprobability bias, or a bias towards all outcomes being equally likely, can be very difficult to shift, even in adults with probability training (Lecoutre, Durand, & Cordier, 1990). Very little research has been conducted into addressing equiprobability bias with young students; most research focused on older students and confirmed the difficulty of assisting students to develop understandings more in alignment with accepted theoretical understandings. The research described here strives to begin to fill a gap in such knowledge and explore possibilities for using argumentation practices and a modeling perspective of inference to initially challenge students' expectations of equally-distributed outcomes before assisting students to make informal inferences more attuned to the theoretical distribution representing the outcomes.

1.3 Literature and Background

For young children, probabilistic knowledge of random events is typically informal, grounded in and shaped by their personal experiences, beliefs and language (Amir & Williams, 1999). These elements can influence their learning of probabilistic concepts in school. For example, the concept of "fairness", developed out of familiar situations involving games and friends, is understood to mean that each person has an equal chance of winning. This idea aligns with the equal nature of outcomes when applied to dice and coins, for example, but may be indiscriminately applied to all possible outcomes (e.g., two heads having the same chance as one head and one tail when two coins are flipped). Equiprobability bias is the tendency to assign equal probabilities to all possible outcomes in any event. This is common among both children and adults, even with instruction and across different ages (Lecoutre et al., 1990; Li & Pereira-Mendoza, 2002).

Watson (2006) argues that while children may hold strong beliefs around outcomes being equally likely, they may at the same time hold strong beliefs that outcomes are determined by "luck" or preference, even though these perspectives are contradictory. Neither of these perspectives necessarily align with formal probabilistic reasoning. For example, when determining the sum of two dice, students may believe that all sums are equally likely while at the same time believe that the sum of three is most likely because it is their lucky number. Even if students are explicitly shown the compound structure of an event or engage in data modeling from experiments, little progress has been shown in shaking these perspectives (Lecoutre et al., 1990). If the idea of "chance" is masked in a probability problem (i.e., students are less aware that the outcome is driven by a random experiment), then some success has been shown in moving students towards more conventional ways of representing and solving the problem (Lecoutre et al., 1990).

1.3.1 Informal Inferential Reasoning

In authentic statistical situations, population data are rarely available or "knowable". Despite this, young students are often accustomed to working with (and hence de-

scribing) complete data that do not acknowledge the greater population or mechanism from which they were drawn. As a result, children may not be familiar with thinking beyond the data in front of them (Makar & Rubin, 2009), but rather see them as fixed. This may be one reason why, when students are initially introduced to samples and asked to make claims "beyond" their data, there is a reported tendency for them to express certainty-only (deterministic) or uncertainty-only (relativistic) viewpoints (Ben-Zvi, Aridor, Makar, & Bakker, 2012; Rubin, Bruce, & Tenney, 1991). Building on Rubin, Hammerman, and Konold's (2006) argument of the need for students to develop aggregate thinking in order to make informal statistical inferences, we speculate that there may be benefits to enabling students to not just work from a sample to infer to a population, but also to try to determine the population, and then extrapolate likely samples. By integrating this focus on samples and distributions within randomly-driven contexts, we suspect that children may learn to draw on data to reason probabilistically. However, we feel the use of a strongly evidence-based approach has the potential to challenge students' beliefs more deeply.

We describe informal statistical inference as a generalization (or claim) beyond the data, that uses the data as evidence, and acknowledges uncertainty (Makar & Rubin, 2009). Inferential statistical reasoning—the reasoning that underpins and leads to an informal statistical inference—is nurtured and developed by an inquiry-based learning environment that develops norms and habits around inquiry, statistical concepts and tools, and tasks that challenge students' beliefs (Makar, Bakker, & Ben-Zvi, 2011). Bakker (personal communication, August 2007) further argued that when students make inferences beyond data, they do so with an expectation or model of the data in mind. In this chapter, we explore this idea by explicitly examining the probability distribution models created by young children in seeking to select a sample from an unknown population. Although the children did not consider their work as creating a sample based on a probability model, their actions in developing their samples can be compared to more formal models. In particular, we sought to build on what we suspected would be grounded in equiprobability models and their sense of fairness (uniform distribution), towards the triangular probability distribution that would be expected from the outcome of adding two numbers between 1 and 10.

1.3.2 Models and Modeling in Primary Mathematics

Research has provided several examples of data modeling with young children (e.g., English, 2012; Lehrer & Schauble, 2000). These examples typically focus on ways that children conduct and represent data investigations. Modeling in classrooms often takes one of two approaches. Either models can be used to mathematize, apply and communicate processes using models already known to learners, or models can be used to develop new models (to the learner) through model-eliciting activities. The latter provides children with a sense of agency and opportunity to experience statistics as a tool for learning about the world, and to engage in sense-making and critical analysis (Greer, Verschaffel, & Mukhopadhyay, 2007). Models in this second form have an additional potential in engaging children in learning to envision and antic-

ipate structures in the invariances (patterns and structures) that underpin variability in data, as is done in the discipline (Lehrer & Kim, 2009, p. 116):

> In everyday discourse, variability is often associated with a lack of structure or pattern, as mere difference among data. The disciplined view is very different: Variability is structured as distribution, and the nature of the distribution reflects the operation of a repeated random process (DeGroot, 1975; Thompson, Liu, & Saldanha, 2007). Random is not a synonym for haphazard, but is instead a description of phenomena having uncertain individual outcomes and predictable pattern, given sufficient repetition (Moore, 1990).

The need to reconcile the unpredictability of individual outcomes of random processes with the predictability of patterns associated with aggregated outcomes is critical in statistics education. The perception of variability as "anything goes" is rampant both in understandings about data and in conceptualizing how samples and variability between samples can be harnessed to represent patterns in the population (Lehrer & Kim, 2009; Rubin et al., 1991). Data modeling, even from a young age, can support students in articulating informal inferences from random processes—ones that coordinate the opposing concepts of randomness in everyday discourse with expectations of probability models in the discipline.

Other work with young children has explored their engagement with disciplinary structures and patterns to deepen mathematical learning. For example, seeing and representing patterns and relationships are foundational to early algebra. These experiences "not only [develop] an understanding of common mathematical structures but also a tendency to look for patterns in new situations" (Mulligan & Mitchelmore, 2012, p. 2). Research suggests that children begin their exploration of random phenomenon steeped in at least one of three expectations—that outcomes are (1) completely unpredictable, (2) can be attributed to "fairness" (equiprobability), and/or (3) are deterministically controlled (e.g., favorite number; Pratt, 2005). Therefore immersing children in a context that challenges these notions may support them to anticipate patterns in data, seek underlying structures, and make inferences to models in probabilistic contexts.

1.3.3 Theoretical Framework: Argumentation-Based Inquiry

The data in this study come from a classroom in which part of the students' mathematics learning was conducted using mathematical inquiry. Mathematical inquiry is an approach to teaching and learning where students address ill-structured problems that rely on mathematical (or statistical) evidence (Makar, 2012). An ill-structured problem is one in which the problem statement and/or pathway for solving the problem contain ambiguities that require negotiation (Reitman, 1965). For example, students may address a question like, "Do students at our school eat a healthy lunch?" or "What is the best recipe for play dough?" In these questions, students must negotiate what they mean by ambiguous words like "healthy" or "best" (ambiguities in the problem statement) as well as negotiating a plan for both how they will find out (e.g., data collection, analysis, sample selected) and the criteria by which they will

assess their solution (how will they decide whether lunches overall were healthy or which recipe is the best?).

In problems posed with such inherent ambiguity, there is the potential for students to address the problems in ways that are not mathematical or statistical in nature. For example, the "play dough" question above has the potential to be addressed without using mathematics or statistics at all—with students simply claiming the play dough that has their favorite color is the best. Or using the healthy lunch example above, students may choose to determine a healthy lunch by counting pieces of "junk" food and comparing them to overall numbers of "non-junk" food to make a determination. Although this offers some scope for mathematics (counting), there are more complex and deeper levels of analysis available that are also age-appropriate; for example, fractional representations of quantities of each food group in lunchboxes in comparison to daily dietary recommendations, or average totals of sodium, saturated fat, sugar, and kilojoules in each lunchbox, plotted and graphically represented by year level. The former example would enable students to see where their diets were deficient and could be improved whereas the latter example would enable conjectures to be made about whether lunches became more or less healthy as student's age increased.

One means of creating a discipline-based focus in science education has been the introduction of argumentation practices into the classroom. The use of such practices in science has been shown to have multiple potential benefits; including the potential to increase students' understanding of scientific concepts (Howe & Mercer, 2007), to provide students with opportunities to develop high levels of discipline-specific literacy (Jiménez-Aleixandre & Erduran, 2007), to address what is acceptable evidence and reasoning within the discipline (Simon & Richardson, 2009), and to increase students' contextual knowledge (Zohar & Nemet, 2002).

Toulmin's classical work on argument (Toulmin, 1958; Toulmin, Rieke, & Janik, 1984) describes four elements that can be found in any argument: claim, grounds, warrants/rules, and backing. The *claim* is the initial assertion that identifies the stance and position of the argumentor. *Grounds* provide the support required to enable the claim to be accepted: It is the information that the claim is based upon and that leads to the claim being made. *Warrants and rules* provide for the checking of the grounds to determine whether they offer genuine support for the claim: they are the justification for moving from the grounds to the claim. Warrants are not self-supporting and require backing, which validates the use of the warrant. The *backing* is often implied; however, it is essentially identifiable in a valid argument. To these four essential components, Toulmin et al. add qualifiers and rebuttals in order for the proponent to identify limitations to the arguments or circumstances under which the argument might not hold. Essentially, Toulmin's model is such a complex interwoven structure that it has been criticized for the difficulty inherent in distinguishing between components: particularly between data and warrants (Erduran, 2007) and data, claim, and warrants (Kelly, Druker, & Chen, 1998). To address this difficulty, McNeill and her associates provided a simplified Claim-Evidence-Reasoning model for working with younger children (McNeill & Krajcik, 2011; McNeill & Martin, 2011; Zembal-Saul, McNeill, & Hershberger, 2013). In essence, students make a *claim* (statement of position in

the same vein as Toulmin's claim), provide *evidence* (Toulmin's grounds), and then *reason* how the evidence enabled them to move towards the claim (Toulmin's warrants and backing). This model was adopted as a structural model for argument in this study for working with, and teaching argument to, younger students.

In common usage, the purpose of an argument, and the practice of argumentation, is often to achieve a winning position: to convince an *other* of a particular belief or position, or toward a particular action. Contrariwise, van Eemeren and Grootendorst (2004) proposed a pragma-dialectical model in which the aim is to achieve consensus. That is, the argument is resolved if all parties come to agreement or if opposing views are withdrawn. However, this model may still enable a dominant position, rather than a robust position, to be accepted. Epistemic argumentation (Biro & Siegel, 1992; Lumer, 2010; Siegel & Biro, 1997) seeks to address potential imbalance by providing a goal of collective truth-seeking. While the goal remains to reach consensus, "it is a qualified, justified consensus, where both parties not only share the final opinion but—ideally—their subjective justification for it" (Lumer, 2010, p. 48). Thus, epistemic argumentation comes from a position where the validity of an argument is evaluated through epistemic criteria only (Biro & Siegel, 1992): the argument rests on the quality of the evidence and reasoning advanced, and its acceptability in terms of discipline norms and values.

Another aspect of argument goal and purpose relates to knowledge development. Berland and Reiser (2009) propose three levels of explanation and argumentation—understanding, explanation and persuasion[1]. These levels are aligned with the goals of sense-making, articulation and persuasion, respectively. The goal of understanding (sense-making) is for students to develop a personal sense of that which is being studied. While evidence is at the core of sense-making, this evidence may be based on personal experience, observation, or attempts to incorporate new experiences and knowledge into existing understandings. Essentially this knowledge is internalized and, as such, is largely unavailable to be challenged. One identified benefit to introducing argumentation practices into the classroom is that of "visibilizing" cognitive processes. If student conceptions can be identified through classroom discourse, they are more open to being challenged or enhanced depending on the accuracy of the conception (Jiménez-Aleixandre & Erduran, 2007).

As students engaging in argumentation articulate their understandings to their classroom audience, they necessarily must practice the construction of the argument and prepare to explicate the connections between their claim, evidence and reasoning; for if they do not, it will be requested of them. Thus, as students construct their explanation, and with experience begin to anticipate the delivery, they necessarily engage deeply and critically with the claim-evidence-reasoning dimensions and interactions. Research would indicate that students rarely engage in the persuasion stage (Berland & Reiser, 2009). The stage differs from explanation in that the goal is to convince others of the epistemic acceptability of the evidence and reasoning advanced, and to develop the most robust understandings available.

[1]Note: While Berland and Reiser refer to this as persuasion, they are referring to the goal of persuading others of the veracity of the reasoning and evidence put forward in support of the claim.

In this chapter, we focus on changes in students' probabilistic reasoning as they engaged in an inquiry-based problem centered on a non-equiprobable distribution. Specifically, a class of 7–8 year old students were engaged with the question, "What is the best card for winning addition bingo?" The literature, reported above, suggested that students would likely anticipate complete unpredictability, equally likely outcomes, and/or outcomes based on "luck" or "fairness". We hypothesized that as they moved from sense-making to persuasion, the increased focus on evidence would necessitate evolving ideas about the mismatch between empirical samples of data (from a triangular probability distribution) and students' initial inferential expectations. As the evidence was increasingly challenged, we anticipated a shift in student thinking from a uniform distribution model to one which more closely represented the triangular probability distribution.

1.4 Subjects and Methods

The class engaged in this teaching sequence was comprised of 22 Year 3 students (7–8 years old) from a suburban government school in Australia. At the time the research was undertaken, the class was taught by two part-time teachers with significant experience in the implementation of inquiry-based learning in mathematics. The unit described here was wholly taught by one of those teachers, Ms. Thomson, who collaborated closely with the first author to reflectively design the lesson sequencing and content.

1.4.1 Context

Design research methodology was adopted (Cobb, Confrey, Lehrer, & Schauble, 2003) to develop an inquiry-based teaching unit around the game of addition bingo (lotto): addressing the inquiry question, "*What is the best card for winning addition bingo?*" In addition bingo, all possible combinations of the sum of two numbers (1 to 10) are written on slips of paper and placed in a box. Children have a card (Figure 1.1) consisting of a 5×5 array of self-selected numbers (their predictions of what will be called), allowing for repeated numbers. As each sum is drawn (e.g., $3 + 8$) from the box, children mark off the sum (in this case, 11) if it appears on their card. Players win the game if they are first to mark off all of the numbers on their card. Prior to engaging in this unit of work, the students had undertaken a previous inquiry, "*Can you make a one-liter container?*" Thus the students were developing familiarity with the need to gather evidence in order to answer an inquiry-based question.

To contextualize the learning, the Queensland state curriculum requires that, by the end of Year 3, the students can:

- Make predictions about chance events using simple statements ("It is likely/unlikely that an event will occur"), and
- Organize data in lists, tables, picture graphs and bar graphs.

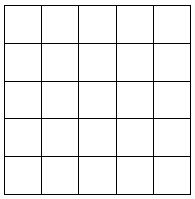


Figure 1.1. Blank Addition Bingo card.

The comparison of experimental estimates of probability and theoretical probability would not be undertaken in the State curriculum until Year 6 or 7; likewise, the recording of numerical probability would be left until such time (Queensland Studies Authority (QSA), 2007).

1.4.2 Data Collection and Analysis

Data collection occurred through various means: Each lesson was videotaped in its entirety and then transcribed for analysis; student work samples were collected; and field notes were also kept to identify salient moments of interest, note the researcher's developing ideas and observations, and to record the informal conversations held with the teacher. In this chapter, the data rely primarily on the video episodes from class lessons, partially informed by the other data collected.

Data analysis was undertaken through several iterations of coding (Corbin & Strauss, 2008). The first iteration served to provide a generalized picture of what was occurring in order to contextualize findings and identify questions that warranted further study. The purpose of this phase was also to identify salient moments that might provide specific insight into both the argumentation process and students' development of probabilistic reasoning, and which would thus indicate potential sections or episodes for further analysis. Examples included moments of insight for either student or teacher, difficulties that were addressed or that might have been problematic in addressing, and students' attempts at using evidence and reasoning. A selection of the episodes and artifacts identified in the initial stages was examined using the claim-evidence-reasoning framework and salient excerpts selected for their ability to illustrate particular aspects of the classroom discourse.

1.4.3 Research Questions

An argumentation-rich unit on probabilistic inference was designed to focus young children on the need to articulate and persuade others through the use of evidence. The research question addressed in this study was:

> *What insights about young students' reasoning under uncertainty and inferential models emerge when a focus on evidence is used to support them in articulating reasoning?*

In particular, when working in a non-equiprobability context,

1. How does a focus on evidence initially challenge students' tenacious equiprobability models?
2. How do students use evidence to articulate and reconcile conflicts between their beliefs, empirical data and inferential models?

1.5 Analysis and Results

As the students progressed through the course of the inquiry, they constructed multiple models to enhance their understanding and to enable them to make inferences about the bingo numbers. These models were purposeful in that the students used them to support and develop their ideas; for this reason, the models themselves, and the use the students make of them, provide insight into students' thinking.

1.5.1 First Model: Uniform Probability Distribution

In students' first round of Addition Bingo, a common initial expectation was that the numbers would be fairly equally distributed. While students likely did not envision a uniform probability model as a distribution when creating their Addition Bingo card, they appeared to draw on an equiprobability assumption. Figure 1.2 shows the distribution of numbers collated from all student cards from the first game. The graph suggests their strong tendency to list all numbers 2 to 20 with a few other familiar numbers to fill in the remaining spaces (note the additional 10s and small even numbers). A few "impossible" numbers were also listed (e.g., 1 and numbers above 20).

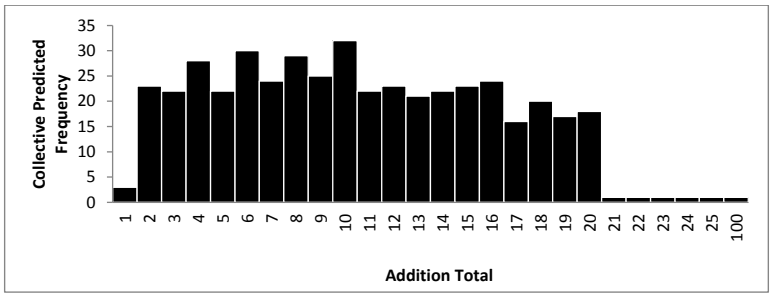


Figure 1.2. Students' aggregated initial frequencies for Addition Bingo.

As the students played the first game of Addition Bingo, they began to recognize problems in the cards they had created (e.g., the number 1 would never be called). As the students in this class were accustomed to negotiation of ideas, these difficulties were not considered by students to be problematic, rather, they saw them as points for discussion. (Note that the numbering system is for reference only and does not imply consecutive remarks.)

01	Gen:	I can't win.
02	Teacher:	You can't win? Why?
03	Lilly:	Clay put a 1 on his card.
04	Teacher:	Put your hand up if you put a 1 on your card? [a few hands go up]. . . I am pleased you realized that part way through Gen. My lowest was $1 + 1$. So what is the lowest number you could use?
05	Students:	Two.

Before the next round, students began to generate strategies that attempted to address the problems of "impossible numbers" that the class encountered in the previous game. As they shared these strategies with the class, students' reasoning suggested they recognized a need to be more attentive to the frequencies of possible outcomes. For example, some students had put numbers such as two or twenty more than once on their card, but in discussion recognized that they could only occur once (i.e., the sum of two can only be represented in a single way). Other numbers, such as 12 and 15, were heard more often than expected. These challenged students' initial ideas of equiprobability model and created a need to seek a more useful model.

Students' initial responses were to fill their cards with these frequently heard numbers. For example, Gideon relied on his memory of hearing 15 come up multiple times. Sirena's reasoning was grounded in the number of possible ways to obtain a sum of 12. At this point the teacher encouraged them to seek evidence for why these numbers were appearing more frequently. A focus on providing evidence encouraged a shift in the class towards identifying the frequencies, as Sirena was suggesting. To facilitate this, the teacher encouraged the students to see if they could find a way of working out which numbers would come up most often.

Several strategies were adopted by students and these emerged over the next few games as they sought methods of finding, testing and providing evidence for the frequencies of each outcome (Figure 1.3). For example, one group counted the number of occurrences of each outcome by pulling each slip of paper from the bucket and tallying them (Figure 1.3; upper-left); other groups sought to record the possible combinations for each number. Some groups relied on the patterns of frequencies (one way to obtain the sum of two, two ways to obtain the sum of three, three ways to obtain the sum of four, etc.) to list the possible outcomes for each number (Figure 1.3; upper-right). One student working alone recorded and tallied the number of ways each number could occur and then displayed the number of tallies on a number line, creating a dot plot (Figure 1.3; lower).

1.5.2 Second Model: Sample Space

As they were reflecting on their progress, students noted that the frequencies they recorded of each sum often differed from their peers. The teacher provided them with an empty addition table as another way to keep track of possible outcomes.

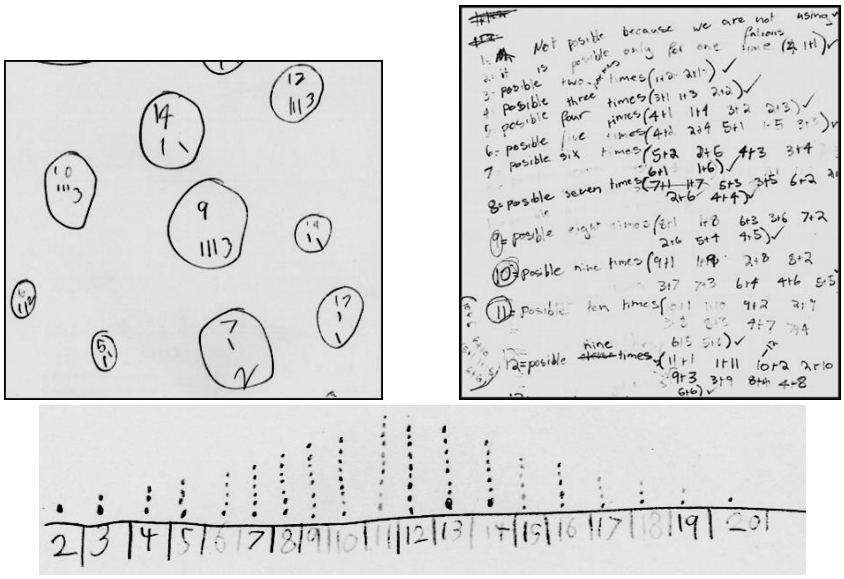


Figure 1.3. Unordered tally marks (upper-left); listing possible outcomes (upper-right); and students' aggregated initial frequencies for Addition Bingo (lower).

Although they did not use the phrase "sample space", they recognized it as a way to ensure that all possible outcomes were recorded.

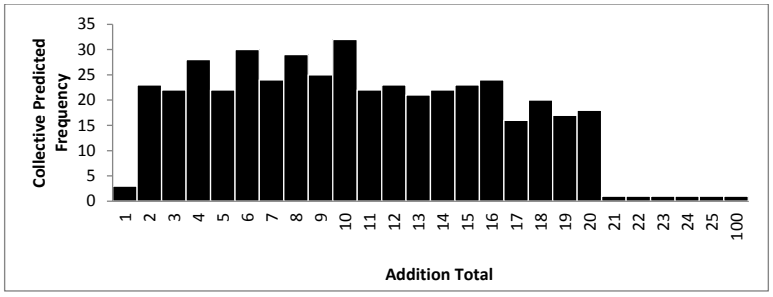


Figure 1.4. Addition table.

The table became a common form of evidence that students used to justify that some numbers were more likely to be selected than others (Figure 1.4). It became the dominant model in the classroom for generating their expectations of which numbers would be called out. There were still tensions, however, in deciding how many of each number to use for their Addition Bingo card as only 25 of the 100 outcomes in the sample space could be used.

06	Gen:	I had 11 is highly likely to be called out; probably, because it is the one that has the most sums and probably they will be pulled out first. But you could still pull out like nine or six or something like that.
07	Teacher:	So you are saying it is not certain, but it is highly likely. Do you have evidence to show me that it is highly likely? If I came over to your book now, do you have the evidence to show me that?
08	Gen:	[indicating an addition table and running her finger along the diagonal] the ones that add to 11 are here, and there's 10 of them. We could still have 10, because there are nine chances. So 10 is probably likely.
09	Teacher:	. . . So would you say that is evidence?
10	Jess:	Yep.

As they worked with collating frequencies, the teacher helped them to use the table to calculate the probability of numbers being selected.

11	Teacher:	So if 11 is the most popular, how many 11s?
12	Byron:	Um, 10.
13	Teacher:	OK so there's ten 11s. How many questions were there in the bucket? [the teacher was referring to the number of sums as 'questions']
14	Byron:	100
15	Teacher:	100. So, what is the probability of drawing out an 11?
16	Gregory:	It is the most common.
17	Teacher:	Yeah. . . . 11 is the most common, there is ten of them. There's 100 different sums in that bucket. What is the probability of pulling out an 11? . . .
18	Sirena:	90 out of 100. I mean 10 out of 100.
19	Teacher:	10 out of 100.

Over the next several lessons, the class stopped periodically to calculate and compare the probabilities of different numbers occurring, and discussing whether "common" numbers were "likely" to come up each time the teacher pulled a slip from the bucket. She often used analogies to help them gain an understanding of these likelihoods (if only two children in the class were going to be given ice blocks [popsicles], is it likely or unlikely that you will get one?). As the students selected their numbers for the games, there was a strong shift towards using primarily common numbers. Students often over-estimated frequencies of the most common sums (e.g., 10, 11, and 12), justifying their choices as those most likely to occur (Figure 1.5).

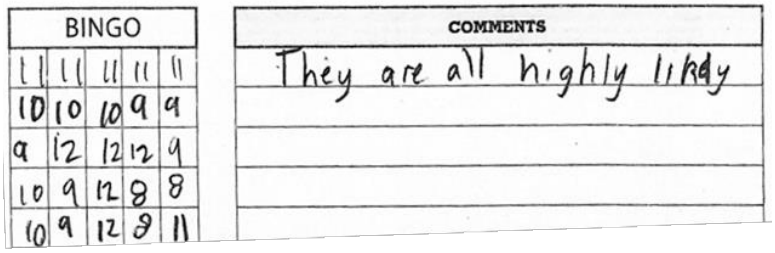


Figure 1.5. A sample Bingo card with reasoning.

Not all of their choices were based on the addition table model, however, as a few students included a mix of common numbers and those they preferred or considered "lucky".

20	Gideon:	I chose the numbers because some are popular and some I put in for no reason, I just felt like it.
21	Teacher:	[repeats what Gideon said] Is that a mathematical way of trying to solve that problem.
22	Students:	No . . .
23	Salena:	I just chose a six and I chose a 16 because it is my lucky number and um I might have a chance of a lucky number
24	Teacher:	Does 16 have a big chance of being drawn out?
25	Troy:	Five out of 100!
26	Teacher:	16 has five out of 100 chance of being pulled out.

Students who filled their card only with the most common numbers from the addition table never won the game. As a model for selecting numbers for their card, there were still problems. Therefore, a discussion ensued about "how many" of the most common numbers should be selected. After one of the games, the teacher had the class examine the numbers selected on the winning card. It was not one that overestimated the common numbers, but was a student who picked a mix of common and less common numbers. Gideon's reasons for picking the numbers were only partially mathematical, so the teacher emphasized that they were looking for a more mathematical approach; that is, an approach that relied on mathematical evidence.

27	Teacher:	Everybody look up this way and let's see what numbers Gideon picked and we will see if we can work out why he won. Let's try and work out, using some maths, why Gideon's card won. OK Gideon, what have you got?

28	Gideon:	[teacher writes the numbers on the board as Gideon reads] 11, 11, 11, 10, 10, 9, 5, 7, 6, 8, 12, 8, 6, 13, 16, 14, 11, 12, 9, 19, 10, 12, 17, 9, 11
29	Teacher:	OK, they are the numbers Gideon chose, and they won. Now have a look at the way I have written that all over the board like that. Have I represented that information very well?
30	Students:	No

1.5.3 Third Model: Paul's Mountain(s)

Students were asked to organize the winning numbers in some sort of logical representation to better visualize the outcomes from the games in relation to the sample space. Paul, who had listed the frequencies of numbers in the addition table onto a dot plot (Figure 1.3; lower), used the same strategy to display the data played in the game (Figure 1.6, represented as sideways tallies).

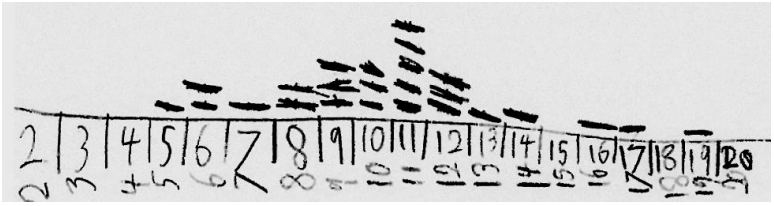


Figure 1.6. Paul's record of Gideon's winning card.

The opportunity to visualize both of these representations (Figure 1.3; lower and Figure 1.6) helped students talk about the difference between the distribution of the sums of the cards in the box (theoretical population) and the distribution of the sums of a winning card (empirical sample).

31	Teacher:	If you have a look at his dot plot that you've got here (Figure 1.6). What can you tell me about the shape compared to his original shape that he had (Figure 1.3; lower). . . . How different is that than his first one? So that is his original dot plot. Remember his original dot plot [stops to discipline]. His original dot plot started at two, went down to 20. It got higher in the middle and went back down again. So what was this representing again? . . . What was this telling me? What was it representing?
32	Clay:	Kind of like a mountain.
33	Teacher:	. . . What was this telling me? What was it representing?

34	Byron:	How many things were in there [the box].
35	Teacher:	. . . If this one shows us all the numbers in the box, have a look at this one, his new one (Gideon's winning card distribution, Figure 1.6). . . . What about it looks different? What about it looks similar? What do you see on here that makes it a bit similar to the other one? Alex, what can you see that makes it a little bit similar? Gen?
36	Gen:	There's a mountain in it.
37	Teacher:	There is a mountain in the middle. This one has got a mountain as well. . . . looking at that, would you say that Gideon picked good numbers?
38	Students:	Yes.

In later rounds, students began to rely more on "Paul's mountain" to select their numbers than the addition table (sample space; see Figure 1.4) used previously. In one round, as the numbers were called out, they were recorded on a dot plot on the board. The teacher then recorded the winning card and the two cards coming second on the same distribution, along with the shape of Paul's mountain superimposed above them (Figure 1.7).

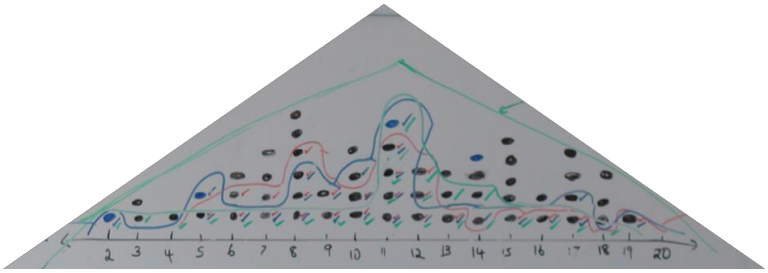


Figure 1.7. The numbers pulled from the box, with the winning three cards. Paul's mountain superimposed above.

In order to make comparisons among the winning cards, the numbers being drawn out and the sample space of all possible numbers, the teacher tried to focus students on interpreting the representations within the context and strategies of the game.

39	Lorena:	It's like Paul's mountain.
40	Student:	Paul's still the biggest
41	Researcher:	Paul's mountain has to be the biggest I suppose.
42	Teacher:	Why is Paul's mountain [the biggest]?
43	Aiden:	It's got all the numbers

44	Teacher:	It's got all the numbers in it. It's got every single number. His mountain has got 100 numbers. . . . What does a bump in the mountain mean?
45	Glenn:	It means that 11 is the highest. The numbers that have the small ones they are not really being called out a lot.
46	Researcher:	. . . OK. So what will you expect to happen next time? It may not happen but if you were allowed to predict what you think might happen, what do you think might happen?
47	Alanna:	Probably something a lot different.
48	Teacher:	A lot different? When I look at your mountain and I look at other people's mountains, what do you notice? Anybody?
49	Sirena:	That umm. That they chose the numbers that came out the most and they didn't choose a lot of the numbers that didn't come out the most.
50	Teacher:	So they chose a lot of numbers that came out the most?
51	Sirena:	But they didn't choose a lot of the numbers that didn't come out the most.
52	Teacher:	So you can still see that people here have chosen the numbers that were in the middle of Paul's mountain. Can you see that from the data there that people have done that? . . . like when you have a look at it, the most of the bumps are here aren't they [8–13]? That is where most of your high stuff is, and I think if you chose mostly from that area, and those people have actually won.

As students continued to play the game and test their strategies, the teacher led discussions that asked them to reason and provide evidence for the numbers they selected. They aimed to connect key statistical ideas to their strategies. Importantly, the discussions were not used to "shame" students who did not select productive strategies, but rather were used to encourage students to (respectfully) challenge one another's thinking with statistical evidence.

53	Teacher:	You said that six and 16 have an equal chance. OK. Can you prove that?
54	Lorena:	Yes, because six has five chances and on the other side (using the symmetry in Paul's mountain), it is the same, five.
55	Teacher:	OK you chose six and 16 because they have equal chances. How many sixes and 16's did you choose?

56	Lorena:	I didn't use six but I chose four 16's
57	Teacher:	. . . But is it [going to] give you the best chance of winning? Why not Jess?
58	Jess:	Because, well, I'm not trying to be mean to Lorena or anything, but like 16 isn't a really popular number so it might not come out as much as four times. It could but it sort of like is only a possible chance of 16 coming up.

1.5.4 Iterations of Games: Accumulating Samples

In order to see the effect of repeated iterations of the game, the students filled out smaller 3×3 cards and played many games quickly. A dot plot was created to keep track of numbers that were called out in each successive game (Figure 1.8). Students were asked to compare the accumulating samples to what they expected to see. (The following comments are selected from the transcript of the discussion, in order, but are not necessarily sequential.)

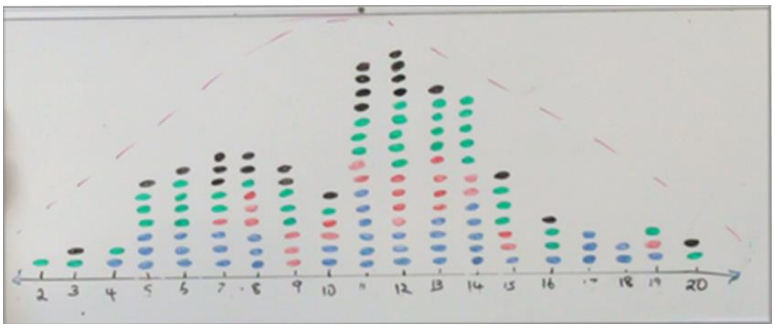


Figure 1.8. Accumulation of sums pulled over multiple games.

59	Teacher:	Is there something that you have predicted to happen but it isn't happening?
60	Isobel:	The 12 is more than the 11.
61	Jess:	With all the colors the 11 is a bit bare, I would expect more, about five or six. But in the second round, it only came out twice and the 12 and the 13 came out more often.
62	Byron:	The 10 is much shorter and I would have predicted that 10 would have come up more.

Most students were comparing the accumulating data to the shape of the distribution they had come to expect from Paul's mountain. Their growing sense of this the-

oretical model (a triangular probability distribution), enabled them coordinate their meanings in terms of the experimental outcomes. Students sometimes combined statistical evidence with common fallacies.

63	Teacher:	Next game. What would you predict might happen?
64	Jess:	I think the 10's will start building up.
65	Teacher:	Why?
66	Jess:	I think because it hasn't been called out. The numbers are just thinking 'Oh we've been called out a lot so we might give 10 a go'
67	Teacher:	So you think the numbers are *thinking* that?
68	Students:	[giggle]
69	Teacher:	Do you think that what you just said is based on any evidence that you know?
70	Jess:	Nooo . . .

Because of the continued focus on evidence, students came to expect these challenges. Throughout the unit, these challenges supported a shift from the equiprobability model, towards a model of the sample space (addition table) and finally the triangular probability model. Rather than being "taught" the triangular probability distribution, clearly not appropriate for this age, the process of negotiation created opportunities to connect each of these models to informal concepts of sample space, theoretical and empirical distributions, compound probability and inference through their expectations, testing and representations. In the end, students were asked to create and then critique their "best" Addition Bingo card. Gen's card below (Figure 1.9) suggests that students had clearly shifted from their original uniform probability model and were able to reason how their card was related to an ideal "best" card to win the game.

1.6 Discussion and Implications

When making an informal statistical inference, do young children infer to a model? We believe they do. Although the children in this paper were not aware of the formal probability models used in the discipline, their inferences suggest that they envisioned and anticipated structure in the variability of the data. That is, they had expectations around how the numbers in the addition bingo game would be drawn. The children accepted the uncertainties of drawing out numbers from a box, but did not approach it as a completely haphazard event. From the beginning, most students accepted the bounds of possible sums (two through twenty) or became aware of them very quickly. Within those constraints, the children initially selected numbers that appeared to have a "fair" chance of being selected (with perhaps a lucky number or two thrown in). This implicit expectation of a uniform probability model was evidenced both individually by the selection of numbers on their card and the collective

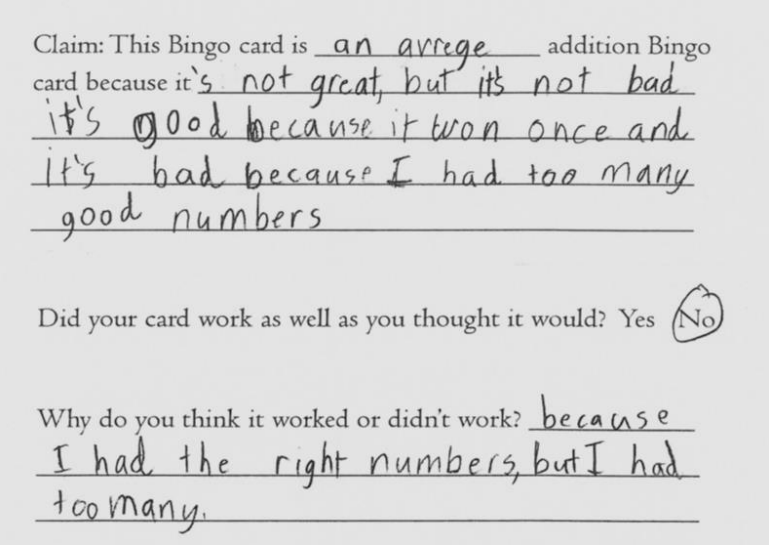


Figure 1.9. Gen's reflection on her final card.

shape of their collated distribution. A uniform probability model is used when the values of the distribution are equally likely; and indeed, three was just as likely to be selected as 11 in the children's initial cards (Figure 1.2). They expected numbers to be equally likely. We argue that they were inferring to a uniform probability model, even though they were not explicitly aware of this.

As numbers were drawn, the uniform probability model was challenged. The children were not surprised by the individual outcomes that were drawn, but by their collective distribution. Through their awareness that some numbers appeared multiple times, the importance of attending to frequency (not relevant in a uniform distribution) became apparent. Not all children let go of their "lucky numbers", even at the end of the unit. However even those who did not, found a way to fit their lucky numbers into a distribution that still recognized basic patterns within a triangular probability distribution: a "mountain" in the middle and low frequencies in the tails. The children who were able to capitalize on their awareness of the structure of the theoretical distribution were more successful in the game.

If we do believe that the children were inferring to a model, a statistical structure, we can think more systematically about the models they engaged with, and the benefits that may have been gained. Mulligan and Mitchelmore (2013) argue that considering children's development of mathematical structure can "provide new insights into how young students can abstract and generalize mathematical ideas much earlier, and in more complex ways" (p. 29). We see connections between the desire to understand and develop children's mathematical-statistical structures and the movement to introduce complex concepts at an informal level in earlier years of schooling. The structures that underpin key concepts in statistics can be introduced through

informal concepts to expose and develop children's sense of structures within the discipline. These structures enable us to work meaningfully with mathematical and statistical ideas in unfamiliar problems. Models, therefore, provide underlying structures that children can learn to depend on, adapt and make meaning from within a problem.

Children's search for frequencies of numbers in the game created a number of experiences and perspectives of frequency, not all of which were model-based. Hearing particular numbers such as 12 and 15 come up multiple times developed an awareness of frequency that may have challenged an equiprobability model, but did not provide a structure from which to anticipate frequencies. Tallying the numbers in the bucket again reinforced the idea that some numbers were more commonly occurring than others but still lacked a sense of pattern that could reveal how the frequencies were structured in the distribution. Several students did recognize patterns once they began listing possible combinations (one way to obtain the sum of two, two ways to obtain the sum of three, etc.) and this initial awareness of a pattern helped them to check their expectations against the frequencies they counted; it was challenged when they realized that there were not 11 ways to make 12, as they anticipated. The sample space supported a way to reliably record all possible outcomes. Students were able to use the sample space to calculate probabilities of outcomes, and make meaning of the small likelihood of even the most frequently occurring numbers. The addition table also revealed patterns in the relationships among the outcomes (e.g., equal sums falling along a diagonal) that enabled the students to be confident in selecting numbers for their cards, although these patterns may have contributed to an over-emphasis on the most common numbers. The sample space in the form of an addition table therefore was a model from which they could work, although it lacked the visual opportunity to envision the structure of variability and "modal clump" in a triangular probability distribution. "Paul's mountain" generated a more visual structure that underpinned the theoretical model.

The theoretical model of a triangular probability distribution represents the variability in the sample space deterministically. The benefit of the theoretical model is in its ability to expose predictable patterns and relationships in the sample space. However, it lacked the sense of randomness that children experienced and came to expect in each game. The familiarity in the mountain that appeared in Paul's tally-dot plot of Gideon's winning card (Figure 1.6) provided this uncertainty. As an empirical model of a triangular probability distribution, Paul's second representation embraced both the structure of the expected outcomes and the sense of uncertainty by showing variability in its divergence from the theoretical distribution. Models are useful when they incorporate the structures and relationships we understand, expect and are productive-in-use; we believe that the class embraced Paul's second model more readily because of this. The two models together set up an opportunity to recognize and express the difference between the theoretical frequencies (what was expected) and the empirical frequencies (what occurred). These two concepts and their relationship will not be met formally for several years, yet the structures visible through these models enabled students to make sense of the game and provided a strong sense of utility in the dot plot representations.

1.6.1 Argumentation-Based Inquiry

We contend that the unit of work reported here illustrates potential for the use of argument-based inquiry in challenging students' alternate probabilistic conceptions, particularly those that have been previously identified as difficult to shift. The nature of epistemic argumentation is such that it requires students to focus on the provision of reasoned evidence. This enables student inferences and claims to be challenged by focusing on the evidence provided rather than taking a 'right or wrong' approach to their learning. In the example here, the students were initially guided into a position where the ideas they had, and the models they were perhaps subconsciously working from, were challenged through the activity itself. Argumentation-based inquiry requires significant scaffolding by a teacher. In this instance, the teacher has used questioning extensively both to support and challenge students' developing understandings of probability (Makar, Bakker, & Ben-Zvi, under review).

The experimental data outcomes of the first game caused students to question the ideas they already held. The teacher was then able to take this moment of conflict and press students to provide evidence that would enable them to identify more likely potential outcomes. The focus on (statistical) evidence supported the development of new models that more accurately reflected the actual triangular probability distribution. Once the students had presented evidence, the multiple representations of evidence were able to be shared with the class. This enabled both the teacher and the students to visibilize student thinking and challenge inaccurate ideas by focusing on the evidence rather than on the student; a process that was clearly a developed norm in this class and that students did not appear to find in the least confronting.

After the students had developed their evidence, they were able to infer from their new distribution models and to make claims regarding likely samples. The distributions they had developed as evidence were not sufficient to provide a decisive answer: the students needed to reason from the distribution (evidence) to the Bingo card (claim). The students were able to provide reasoning, such as seen in the Bingo card in Figure 1.5. This enabled a level of challenge to be provided by the teacher and other students once again as we saw when Jess challenged Lorena (lines 56–58). However, the repeated iterations of the games, and the resultant production of experimental outcomes also served to challenge students' reasoning from evidence to claim.

An additional aspect worthy of note is that of classroom culture. This classroom clearly had established classroom norms of inquiry and negotiation. The students were accustomed to being challenged and to challenging others in ways that were non-confronting but were rather aimed at developing stronger and better responses: essentially students did not progress as if they knew there was a final goal in mind, only a best-case scenario. This may have served the students particularly well in working with the uncertain nature of informal inference. A second and possibly aligned aspect of this sequence of lessons was the engaging nature of the problem. The students enjoyed the game playing aspect and the level of mild competition inherent in playing each round. This level of competition, and a desire to get a good

card so they could win, may well have motivated the students to persevere with the micro-adjustments they were making towards the end.

1.6.2 Implications for Research and Teaching

This study has important implications for research in statistics education.

- *Need for more research.* The research described here is derived from a rich inquiry task undertaken by a single class. As such it is intended to provide specific insights into student thinking rather than develop a transferable approach. Additional research would be beneficial on a much wider scale to determine the potential for argumentation to challenge probabilistic intuitions.
- *Transfer issues.* One particular aspect which would warrant further research is the issue of transfer. (Fischbein, 1987) elaborates on the interwoven nature of probabilistic schema and further research is essential into the extent to which the internalized schema is impacted, both in terms of near and far transfer and longevity of the new understandings.
- *What models are relevant at a young age?* If inferential reasoning assumes that students are inferring to a model, what types of models (beyond uniform distributions) are apparent before students learn about formal probability models and standard distribution shapes?
- *Literature on probability from an inferential perspective.* Much of the classic literature on students' reasoning under uncertainty is situated in literature that does not take an inferential perspective. Given the increased focus on learning probability and statistics through inference, there would be benefit in rethinking and revising classic research from an inferential perspective.

This study has useful implications for teaching probability and statistics to young children.

- *Beyond equiprobability contexts with young children.* Children's early experiences with probabilistic thinking stem from experiences and beliefs established informally and away from the schooling context (Amir & Williams, 1999). The intuitions that develop as a result often remain hard to shift. In terms of equiprobability bias, even in those with formal probabilistic knowledge demonstrate reliance on intuitions (Lecoutre et al., 1990). Early schooling experiences may serve to reinforce these intuitions as students often engage with aspects of probability activities that, while focusing on randomness, remain restricted to events with equiprobable outcomes (e.g., simple coin and dice tossing, spinners). Students may benefit from early exposure to experiments that have unequal outcomes and outcomes that are unknown or difficult to identify theoretically.
- *Focus on evidence may shift/undermine stubborn conceptions.* The use of argument in the classroom has potential to challenge students through several

mechanisms. The first is that structuring an argument, whether formally or informally, requires the student to consider the evidence and present a coherent claim derived from evidence. This necessitates an evidentiary focus as distinct from a "gut" feeling. Second, and dove-tailing with the first, is that the presentation of the argument enables the students to identify the evidence and reasoning attached to the claim. Thus intuitions can be identified and addressed by either classmates or teacher: evidence can be challenged, reasoning can be questioned, and, if necessary, further investigations can be established to create or increase the state of cognitive disequilibrium.

Acknowledgements

This research was funded by the Australian Research Council DP120100690, Education Queensland and an Australian Postgraduate Award.

References

Amir, G. S., & Williams, J. S. (1999). Cultural influences on children's probabilistic thinking. *Journal of Mathematical Behavior*, *18*(1), 85–107. doi: 10.1016/S0732-3123(99)00018-8

Australian Curriculum Assessment and Reporting Authority. (2014). *Australian curriculum: Mathematics v7.1. ACARA*. Retrieved from http://www.australiancurriculum.edu.au/Mathematics/Curriculum/F-10

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—The International Journal on Mathematics Education*, *44*, 913–925. doi: 10.1007/s11858-012-0420-3

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26-55. doi: 10.1002/sce.20286

Biro, J., & Siegel, H. (1992). Normativity, argumentation and an epistemic theory of fallacies. In F. H. van Eemeren, R. Grootendorst, J. A. Blair, & C. A. Willard (Eds.), *Argumentation illuminated* (pp. 85–103). Amsterdam, The Netherlands: International Centre for the Study of Argumentation.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. doi: 10.3102/0013189X032001009

Corbin, J. M., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage Publications.

DeGroot, M. H. (1975). *Probability and statistics*. Sydney: Addison-Wesley.

English, L. (2012). Data modeling with first grade students. *Educational Studies in Mathematics*, *81*(1), 15–30. doi: 10.1007/s10649-011-9377-3

Erduran, S. (2007). Methodological foundations in the study of argumentation in science classrooms. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Ar-*

gumentation in science education (pp. 47–69). Dordrecht, The Netherlands: Springer.

Fischbein, E. (1987). *Intuition in science and mathematics*. Dordrecht, The Netherlands: D. Reidel Publishing Company.

Greer, B., Verschaffel, L., & Mukhopadhyay, S. (2007). Modelling for life: Mathematics and children's experience. In W. Blum, P. Galbraith, H.-W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education* (pp. 89–98). New York: Springer.

Howe, C., & Mercer, N. (2007). *Children's social development, peer interaction and classroom learning: Primary review research survey 2/1b*. Cambridge, England: The Primary Review.

Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education* (pp. 3–27). Dordrecht, The Netherlands: Springer.

Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *International Journal of Science Education*, *20*(7), 849–879. doi: 10.1080/0950069980200707

Lecoutre, M. P., Durand, J. L., & Cordier, J. (1990). A study of two biases in probabilistic judgments: Representativeness and equiprobability. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 563–575). Amsterdam, The Netherlands: Elsevier.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116–133. doi: 10.1007/BF03217548

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 101–159). Mahwah, NJ: Lawrence Erlbaum Associates.

Li, J., & Pereira-Mendoza, L. (2002). Misconceptions in probability. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots6/6g4_jun.pdf

Lumer, C. (2010). Pragma-dialectics and the function of argumentation. *Argumentation*, *24*(1), 41–69. doi: 10.1007/s10503-008-9118-7

Makar, K. (2012). The pedagogy of mathematical inquiry. In R. Gillies (Ed.), *Pedagogy: New developments in the learning sciences* (pp. 371–397). Hauppauge, NY: Nova Science.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173. doi: 10.1080/10986065.2011.538301

Makar, K., Bakker, A., & Ben-Zvi, D. (under review). Scaffolding norms of argumentation-based inquiry in a primary mathematics classroom. *ZDM—The International Journal on Mathematics Education*.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

McNeill, K. L., & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing*. Boston, MA: Pearson.

McNeill, K. L., & Martin, D. M. (2011). Claims, evidence, and reasoning. *Science and Children*, *48*(8), 52–56.

Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, D.C.: National Academy.

Mulligan, J., & Mitchelmore, M. (2012). Developing pedagogical strategies to promote structural thinking in early mathematics. In J. Dindyal, L. P. Cheng, & S. F. Ng (Eds.), *Mathematics education: Expanding horizons (Proceedings of the 35th annual conference of the Mathematics Education Research Group of Australasia).* Singapore: MERGA.

Mulligan, J., & Mitchelmore, M. (2013). Early awareness of mathematical pattern and structure. In L. English & J. Mulligan (Eds.), *Reconceptualizing early mathematics learning* (pp. 29–45). New York: Springer.

Pratt, D. (2005). How do teachers foster students' understanding of probability? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 171–189). New York: Kluwer/Springer Academic Publishers. doi: 10.1007/0-387-24530-8_8

Queensland Studies Authority (QSA). (2007). *Essential learnings (mathematics).* Retrieved from http://www.qsa.qld.edu.au/7296.html

Reitman, W. R. (1965). *Cognition and thought: An information processing approach*. New York: Wiley.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics (icots-3)* (Vol. 1, pp. 314–319). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.ime.usp.br/~abe/ICOTS7/Proceedings/index.html

Siegel, H., & Biro, J. (1997). Epistemic normativity, argumentation, and fallacies. *Argumentation*, *11*(3), 277–292. doi: 10.1023/A:1007799325361

Simon, S., & Richardson, K. (2009). Argumentation in school science: Breaking the tradition of authoritative exposition through a pedagogy that promotes discussion and reasoning. *Argumentation*, *23*(4), 469–493. doi: 10.1007/s10503-009-9164-9

Thompson, P. W., Liu, Y., & Saldanha, L. A. (2007). Intracacies of statistical inference and teachers' understanding of them. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207–231). New York: Lawrence Erlbaum Associates.

Toulmin, S. (1958). *The uses of argument*. New York: Cambridge University Press.

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed.). New York: Macmillan.

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. New York: Cambridge University Press.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zembal-Saul, C., McNeill, K. L., & Hershberger, K. (2013). *What's your evidence? engaging K–5 students in constructing explanations in science*. Boston, MA: Pearson.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62. doi: 10.1002/tea.10008

CHAPTER 2

‘HOW CONFIDENT ARE YOU?’ SUPPORTING YOUNG STUDENTS’ REASONING ABOUT UNCERTAINTY IN CHANCE GAMES THROUGH STUDENTS’ TALK AND COMPUTER SIMULATIONS

SIBEL KAZAK

Pamukkale Üniversitesi, Turkey

Abstract

In this chapter, the results from a design-based research study to support 10–11 years old students’ articulation of uncertainty in making informal inferences are examined. The pedagogical approach taken was Bayesian-like and was mediated by the use of TinkerPlots™ tools and student talk. The notion of uncertainty used in the study focuses on both students’ probability assessments and their personal degree of confidence in judging the fairness of chance games. Six students at a local primary school in Exeter, UK participated in the study during a 9-hour mathematics enrichment program. Sociocultural discourse analysis was used to qualitatively analyze students’ small group discussions. The findings from the study show that the interaction of using TinkerPlots™ simulation tools and dialogic talk in a small group leads to new insights in students’ reasoning about uncertainty as they make informal inferences about the fairness of the games.

Keywords: Informal inference; Bayesian reasoning; Subjective probability; Technology; Dialogic talk; Scaffolding

> The term 'uncertainty' is unquestionably fraught with misinterpretation—especially by non-scientists. I'd prefer the phrase 'how confident am I?', the reciprocal of uncertainty. (Gibbs et al., 2013, p. 6).

2.1 Overview

The concept of inference in statistics refers to "drawing conclusions about populations or processes based on sample data" (Zieffler, Garfield, delMas, & Reading, 2008, p. 40). Formal statistical inference that students encounter in advanced levels includes certain techniques to draw conclusions from data, such as estimation and hypothesis testing. Given the problematic nature of understanding these formal ideas in the context of making inferences by older students (Zieffler et al., 2008), there has been an attempt to begin to develop the foundation for these ideas early on, working with young students to use statistical inference informally (Makar & Rubin, 2009).

Informal statistical inference can be viewed as a way of making informal conclusions (using statistical and probabilistic knowledge) about a population or process from which the data come. The notion of uncertainty plays an important role in making such judgments using data. According to the framework described by Makar and Rubin (2009) informal statistical inference includes making a generalization beyond data, using data as evidence, and using probabilistic language in describing the generalization. So, at the heart of statistical inference is "the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected" (Makar & Rubin, 2009, p. 83), which inherently involves features of uncertainty. Thus, developing the language and understanding of probability in the sense of "statistical tendency, and/or level of confidence or uncertainty in a prediction" (Makar & Rubin, 2009, p. 87) is crucial in reasoning and making decisions based on uncertain data.

This chapter presents how young students articulate uncertainty in the context of fairness in games of chance as they test their hypotheses and update their level of confidence on the basis of the data collected both through physical experiments and computer simulations in TinkerPlots™ (Konold & Miller, 2011). Within this context, the main goal of this chapter is to examine in what ways the combination of using TinkerPlots™ and peer-to-peer dialogic interactions supports students' reasoning about uncertainty in making informal inferences about chance situations through a Bayesian-like approach.

The present study is part of a larger design-based research project which has the overall aim of investigating how to develop young students' conceptual understanding of key ideas in statistics and probability in the context of informal statistical inference through the mediating roles of technological tools and students' talk. The study was conducted during a 9-hour mathematics enrichment program with six primary school students, ages 10–11, in Exeter, UK. The teaching experiment discussed in this paper is one of the initial iterations of the design-based study.

2.2 Problem

In recognition of the role that working with data and making judgments under uncertainty play in daily life and in various fields, statistics and probability have become part of the mainstream school mathematics curricula in primary grades for more than two decades (e.g., Department for Education and Employment, 1999; National Council of Teachers of Mathematics, 2000). The emergence of teaching statistics within school mathematics has led Exploratory Data Analysis (EDA; Tukey, 1977), which involves real data analysis through looking for and describing patterns or trends in data, to become the basis for the data-handling strand in the pre-tertiary mathematics curricula (Biehler, 1986; Shaughnessy, Garfield, & Greer, 1996). In the meantime, as Ainley and Pratt (2001) point out, EDA provided an opportunity for open-ended data exploration by students, using basic concepts of descriptive statistics, while "foregrounding data and making the mathematical model, probability, subsidiary." (p. 7). This has created an artificial separation between data and chance topics, both relevant to uncertainty, in both research and instruction when in fact they are closely related (see Konold & Kazak, 2008). So, one way to build a strong connection between data and topics related to chance is to encourage students to make informal conclusions based on data (Moore, 1990). As it can also be seen in the development of themes of the Statistical Reasoning, Thinking, and Literacy (SRTL) forums in the past decade, the focus has recently shifted towards informal statistical inference, which is an important component of statistical thinking, at all grade levels (K–College). A special issue of the journal *Mathematical Thinking and Learning*, for instance, was devoted to the papers presented at the SRTL-6 forum with a focus on the role of context in developing students' reasoning in informal statistical inference (Makar & Ben-Zvi, 2011).

Since informal statistical inference is considered an end product, the underlying reasoning process leading to that is called informal inferential reasoning (Makar, Bakker, & Ben-Zvi, 2011). According to Fisherian inference, this reasoning process is based on the concept of likelihood which entails: (1) formulating a hypothesis (i.e., null-hypothesis or model), (2) making a judgment that if the hypothesis or model were true, the observed data would have been very unlikely (i.e., intuitively computing a p-value), and (3) rejecting the initial hypothesis or model based on the conditional probability that the observed data would occur by chance (Rossman, 2008). Rossman argues that students do not seem to spontaneously apply this common form of reasoning when making statistical inferences. An alternative approach to statistical inference, which is distinct from that of Fisher, is based on a Bayesian framework. Bayesian inference uses a subjectivist interpretation of probability (Rossman, 2008). In this approach, one would start with a priori probabilities associated with a hypothesis or model based on a belief or previous data, then update these probabilities as new information or data is obtained. It is argued that this form of deductive reasoning seems to be more intuitive than that of Fisher (Albert, 2002; Rossman, 2008). Because of this, there has been a tendency to shift the focus of inference in undergraduate-level statistics courses to a Bayesian framework (Albert, 2002; Díaz, 2010).

The subjective viewpoint of probability is closely connected to Bayesian inference, and is often equated to the degree of belief. Despite its common use in everyday reasoning, subjective estimates of probability have been neglected in the school mathematics curricula (Jones, Langrall, & Mooney, 2007). Hence, there has been a lack of interest in researching this aspect of probability. Given this gap in the literature, and that Bayesian thinking and reasoning tend to be more intuitive than the frequentist perspective in statistical inference, a Bayesian-like approach was adopted to study 10–11 year old students' reasoning about uncertainty in making informal inference in the context of random chance events. More specifically, the *Chips Game Task* described in this chapter is designed to understand students' articulation of uncertainty as they evaluate the fairness of chance games by making an initial hypothesis and expressing their confidence in the likelihood of a particular game actually being fair (or not). Students then revise both their initial hypothesis and their level of confidence as new information is obtained from the data collected through physical experiments and computer simulations. The notion of uncertainty in this task relates to both students' probability assessments and their personal degree of confidence in judging the fairness of the games. Moreover, a link to making informal inference is established by focusing on the probability estimates through experimental data in the task (Konold et al., 2011). Within this context, the following research questions are investigated: (1) How does the combination of using TinkerPlots™ and dialogic interactions in small groups promote students' reasoning about uncertainty in making informal inferences about random events? (2) What are the dialogic mechanisms that help support students' reasoning in the joint activities?

2.3 Literature and Background

Probability is the science of quantification of uncertainty in random processes. The approach used in this study to examine reasoning about uncertainty has roots in the historical development of probability. Hacking (1975) noted that the concept of probability has historically had a dual characteristic: On the one hand is an epistemic notion of probability understood as degree of support by evidence, and on the other hand is a statistical notion of probability concerned with stable frequencies of occurrences of certain outcomes during statistical processes like tossing a coin repeatedly many times. Similarly, Hald (2003) distinguished the two kinds of probability as: subjective probability "used for measuring the degree of belief in a proposition warranted by evidence" (p. 28) and derived from our imperfect knowledge, and objective probability "used for describing properties of random mechanisms or experiments, such as games of chance, and for describing chance events in populations, such as the chance of a male birth" (p. 28).

An implication of this dual nature of probability mentioned in Hacking (1975) and Hald (2003) is twofold. First, the epistemic and subjective notions of probability emphasize personal probabilities relative to our background knowledge and beliefs, and thus enable us to represent learning from experience. Second, the statistical and objective notions of probability are based on the symmetry in the mechanisms of chance

setups, such as equally likely outcomes or the stability of relative frequencies from experiments in the long run. Furthermore, Bernoulli distinguished between theoretical "probabilities which can be calculated a priori (deductively, from considerations of symmetry) and [empirical probabilities] which can be calculated only a posteriori (inductively, from relative frequencies)" (Hald, 2003, p. 247). He then proved the first limit theorem of probability ("Bernoulli Theorem") stating that the probability of a large difference between the empirical probability and the theoretical probability tends to zero as the number of trials increases (Stohl, 2005). The idea that the long-run relative frequency of an event should be very close to the probability of that event is an important corollary of this theorem.

For educational purposes, these different views of probability concepts suggest that when we deal with uncertainty in chance events, we draw upon a variety of evidence, such as personal knowledge or belief, empirical results, and theoretical knowledge. It is also implied that as one learns to appeal to evidence, symmetry of chance setups, and running simulations, one begins to link subjective, empirical, and theoretical estimates of the probability. In particular, young students' personal and experiential knowledge about the world plays an important role in their understanding of probability. Therefore, in this study the students started with formulating a hypothesis about the fairness of a chance game based on their personal knowledge/belief and updated it with the new data from a Bayesian viewpoint (where certainty level is changeable). The assumption was that the simulation of chance experiments would help students interpret probability of events as the relative frequency of outcomes in the long run (where certainty level increases as the number of trials get larger). Then students were expected to provide evidence for the observed results through theoretical analysis of chance events based on the sample space (where certainty level about their hypothesis is the highest).

2.3.1 Theoretical Background

Based on the idea that inference is an end product of inferential reasoning, Makar et al. (2011) recognized the need for understanding and supporting the informal inferential reasoning process that leads to informal statistical inference. Drawing upon their review of relevant literature and analysis of three sixth graders' informal inferential reasoning, Makar et al. claim that informal statistical inference needs to be embedded in informal inferential reasoning, "nurtured by statistical knowledge, knowledge about the problem context, useful norms and habits developed over time, and supported by an inquiry-based environment (tasks, tools, scaffolds)" (p. 171). Within the aim of this study, the design of the learning environment suggested by Makar et al. is seen as particularly relevant to support young students' reasoning about uncertainty. In this research, the design element involves relevant tasks, appropriate computer tools, and talk as scaffolding in peer group interaction. Then these are used to understand how to promote students' emerging reasoning in making informal inferences in the context of chance events in the current research data.

Designing Relevant Tasks. The research indicates that older students have difficulties in understanding the concepts and reasoning related to the common formal statistical inference methods (Zieffler et al., 2008) and this formal inferential reasoning process does not come naturally to students (Rossman, 2008). It has also been argued that a Bayesian approach to making a statistical inference is more intuitive and better reflects the commonsense thinking about uncertainty in daily life (Albert, 2002). Even young students use probabilistic language (e.g., more likely, possible, impossible, always, and rare) to express different levels of uncertainty. Subjective probability, to which Bayesian inference is closely related, is a way of assigning quantities between 0 and 1 to these different levels of uncertainty with beliefs changing based on new evidence (Albert, 2002). The findings of Huber and Huber (1987) suggest that even young children can use personal knowledge or beliefs in the tasks involving ordinal comparison of subjective probabilities about given events in the contexts of sports and gambling. In addition, it is pointed out that children's subjective probability evaluations of events tend to show more stability in the gambling task because of the availability of the objective probabilities, (i.e., the areas in the spinner device used in the task; Huber & Huber, 1987). This suggests that young students' use of subjective probability in making informal inferences may be supported by enabling them to estimate the likelihood of events from other sources as well (e.g., the symmetry in the mechanism of chance setup and relative frequencies). To do so, we need a task which allows students to use subjective, empirical, and theoretical estimates of the probability.

The Bayesian viewpoint seems to be often consistent with people's way of developing intuitions based on learning from their experiences and revising their beliefs as new information is obtained (Falk & Konold, 1992). Furthermore, Hawkins and Kapadia (1984) suggest that subjective probability is utilized to complement the traditional classical and frequentist approaches in teaching probability. Therefore, to support students' informal inferential reasoning, a task was designed using Bayesian-like thinking to develop informal inference where students were expected to state their initial hypothesis (prediction) about the fairness of chance games, provide an explanation, and rate their level of confidence in their hypothesis on a 0%–100% scale. After generating their hypotheses, students were asked to both physically play the game and simulate results from the game using TinkerPlots™ to gather data to support or revise their initial hypothesis and to update their level of confidence. Afterwards, students used the possible outcomes for the combined events to provide a theoretical model for data. They were then expected to be able to explain using this theoretical model how their previously collected empirical results could be used to support their final hypothesis. Through this process, the aim was to highlight the inherent relationship between probability and informal statistical inference in the context of chance games.

Using Appropriate Computer Tools. Several studies have investigated young students' reasoning processes relevant to inference through technology-enhanced tasks. Pratt, Johnston-Wilder, Ainley, and Mason (2008) found that when guessing the hidden numbers in the sides of a die on a computer simulation tool, called *Inference*

Maker, 10–11-year old students tended not to focus on emergent, aggregate characteristics of data. Accordingly, students failed to see the relevance of larger samples in making inferences with a greater confidence, which then would lead to the idea of the Law of Large Numbers. Others reported on effective use of technological tools, in particular TinkerPlots™, to support these understandings with middle school students (Ben-Zvi, 2006; Fitzallen & Watson, 2010). For instance, Fitzallen and Watson (2010) reported that when using TinkerPlots™, students (ages 10–12) generated different kinds of plots that appeared meaningful to them and used these effectively in making their conclusions from data. In their study, the software also facilitated students' thinking processes that involved moving back and forth between making hypotheses and constructing plots in making sense of the data. Additionally, Ben-Zvi (2006) indicated that students used TinkerPlots™ not only as a representation tool, but also as an argumentation tool in expressing their ideas to others.

Other studies (Konold, Harradine, & Kazak, 2007; Konold & Kazak, 2008) conducted using the development version of TinkerPlots™ also revealed how the new probability simulation feature could support middle school students' development of an integrated set of statistical and probabilistic ideas. Findings from Konold and Kazak (2008) suggested that the TinkerPlots™ environment facilitated students' visual reasoning via dynamic graphs where the results accumulated as they were generated by the Sampler (a tool within TinkerPlots™ to model probabilistic processes). The combination of observing the simulation data from multiple trials and sketching only the overall shape and relative heights of stacks seen in the plot enabled students to explore the fit between the expected distribution based on the sample space and the empirical data. These observations led students to perceive 'data as signal and noise', which is a key idea in dealing with situations involving uncertainty (Konold & Pollatsek, 2002). More recent studies presented in the SRTL-8 forum (Ainley, Aridor, Ben-Zvi, Braham, & Pratt, 2013; Braham, Ben-Zvi, & Aridor, 2013; Harradine & Konold, 2013) further documented the benefits of using TinkerPlots™'s simulation and modeling features in promoting young students' statistical understanding and reasoning about uncertainty in the context of informal inference.

Using Talk as Scaffolding in Peer Group Interaction. It is suggested that encouraging talk in a mathematics classroom helps students reflect on their thinking, explore, and form new understandings (Wickham, 2008). Mercer (1996) identified three different types of talk when students engaged in small group work in classrooms: (1) *disputational talk* in which a lot of disagreement between children, individualized decision making and a competitive, rather than cooperative, relationship can be seen; (2) *cumulative talk* in which children tend to simply build on what the other has said in a shared, supportive but an uncritical way; and (3) *exploratory talk* in which children listen to each other actively, ask questions, challenge ideas in a critical but constructive way, and give explicit reasons for challenges. Several studies using the Thinking Together approach, a dialogue-based pedagogy to develop children's collective thinking and learning (Dawes, Mercer, & Wegerif, 2000), found exploratory talk effective in promoting young students' mathematical reasoning and

problem solving when they work together in groups with mathematics software (e.g., Monaghan, 2005; Wegerif & Dawes, 2004).

Another type of talk, called *dialogic talk*, extends the definition of exploratory talk by referring to collaborative and creative engagements with more emphasis on the dialogic quality of the relationships between students (as well as those between students and the shared task) than on the explicit verbal reasoning (Wegerif, 2013). Wegerif argues that dialogic talk entails openness to the other and to otherness in general to the extent that participating individuals are able to listen to each other and to change their minds. From a dialogic perspective inspired by the work of Bakhtin (1981, 1986), dialogic processes refer to the creative leaps required to understand things from the outside position of the witness or the "superaddressee" position in Bakhtin's terms (Wegerif, 2013). According to Wegerif, this creativity is sometimes an emergent effect of the dialogic space that opens up in the gap between different perspectives, including virtual perspective such as that of the superaddressee.

In Kazak, Wegerif, and Fujita (2013), it is argued that the combination of technology and students' dialogic talk can play a critical role in helping students make noticeable shifts forward in their conceptual understanding of probability. The article describes a trajectory of two 11–12-year old students making conjectures about the fairness of a game, involving combined events (an earlier version of the *Chips Game Task* in the current chapter), testing and revising their initial theories based on simulation data using TinkerPlots™. It was found that the dialogic talk helped these two students in several ways. For example, they could articulate their thinking including half-baked or uncertain ideas. The dialogic approach used in the study encouraged students to ask for explanations because they began to feel that making mistakes and showing that they do not understand were acceptable. In this way they could help each other to understand. Moreover, the mechanism behind the initial switch in perspective in one of the students was interpreted as dialogic in the sense that there was an invisible dialogue going on between the student and an absent 'witness'. While taking an outside perspective in reconsidering the problem, he was able to question his initial view and change his mind.

2.4 Subjects and Methods

This exploratory study was carried out with six high achieving Year 6 (age 10–11) students, two boys and four girls (pseudonyms: Ozzy, Jake, Keyna, Flora, Gabby, and Blair), from a local primary school in Exeter, UK. The sample was selected as a convenience sample recruited through their classroom teacher.

2.4.1 Data Collection

The method of research employed was a design study. A design study entails an iterative process to develop theories of students' learning and ways of supporting their learning of domain specific content (Cobb, Confrey, Lehrer, & Schauble, 2003). The initial design is improved through testing and revising conjectures based on continual

analysis of students' reasoning and the learning environment while the experiment is in progress. In this study, the research cycle involved designing instructional materials and a learning environment that supported the desired learning goals in the domain of statistics, conducting teaching sessions, and retrospective analysis. The retrospective analysis of the first cycle was the basis for the new design phase in the second cycle.

During the study, students were seated to work in pairs or groups of three at tables with a laptop and manipulative materials, such as game chips and bags. Each group's work was videotaped to capture the students' interactions around the computer. Additionally, the computer screen of the group including Blair, Gabby, and Flora was recorded using *Camtasia* software (TechSmith, 2011) to capture their work in TinkerPlots™ environment during the *Chips Game Task* (see Section 4.2). Each group also answered questions on the given worksheets for each task. Pre-assessment items were used to evaluate students' reasoning about combined events prior to the probability tasks. The data for analysis included video footage of group work and lessons, computer screen captures, and student artifacts.

2.4.2 Procedures and Tasks

In this study, TinkerPlots™ (Konold & Miller, 2011) was used as the information and communication technology (ICT) tool. TinkerPlots™ is a distinct computer program compared to other graphing or spreadsheet programs as it builds on children's intuitive knowledge about data representations and analysis. It enables students to construct their own graphs when organizing their data by ordering, stacking, and separating. TinkerPlots™ also includes a variety of tools, such as dividers and reference lines, to intuitively analyze data in making inferences. A probability simulation/modeling tool (i.e., the *Sampler*) allows students to build models of random phenomena using variety of devices (i.e., mixer, spinner, bars, stacks, curve, and counter) that can be filled with different elements from which to sample (see Figure 2.1). This tool then enables students to collect measures and outcomes from the sampled elements. Another affordance of the tool is that it allows students to quickly generate a large number of outcomes with each run and to repeat this several times to look at the results from sample to sample.

In addition to the use of TinkerPlots™ to explore data and chance, the participants were introduced to a dialogic way of communicating during group work. Since not all types of student talk that occurred in groups would necessarily result in effective collaboration in joint activities (see Mercer, 1996), certain ground rules were explicitly discussed and practiced with the students in order to set up the conditions for effective talk (Dawes et al., 2000). As an example, here is a set of negotiated expectations for group work in this study: (1) we should make sure that each person has an opportunity to contribute ideas, (2) we should ask each other 'why?' and listen to the explanation and try to understand, (3) we should ask others what they think, (4) we should consider alternative ideas or methods, and (5) we should try to reach an agreement before we do anything on the computer.

The study involved a sequence of tasks designed to develop key ideas and concepts in probability relevant to uncertainty (i.e., randomness, relationship between theoretical and data-centered estimates of probability, the role of sample size, quantifying uncertainty, confidence level, evidence) and to support their reasoning about uncertainty in making informal statistical inferences through students' talk (in groups of 2–3) and their use of computer tools over three sessions each of which lasted about three hours.

Day 1. After a brief introduction to the software, students began to use the *Sampler* tool to build a data factory to make "monkeys" and "teddy bears." This task was intended to familiarize students with some of the data modeling and simulation tools in TinkerPlots™, as well as to practice incorporating the five dialogic talk ground rules (discussed previously) as they worked together around a laptop in groups of three. Later in the class, the entire group had a discussion about the words "random" and "unpredictable." Students were also asked to provide examples for each of the following words relevant to uncertainty: Likely, unlikely, equally likely, most likely, no chance, even chance, certain, uncertain, fair, and unfair.

In the following activity about random events, students initially were asked to write down the sort of results they would expect to get if they were to flip a coin 20 times. After the discussion of their made-up data, we flipped a coin 20 times and compared the results. They also built a model of single coin flipping, graphed and analyzed the results from 50, 100, and 2000 repeated trials, and discussed the variability resulting from those simulations. At the end of the session, the students were asked to discuss in their groups the fairness of the method described in the following context:

> "Carla and her two friends, Justi and Cloe, all want to ride in the front seat of the car on a short trip they are taking. They agree to flip two coins to decide. Carla wins if the two coins come up different. Justi wins if both coins are heads. Cloe wins if both coins are tails" (developed by Konold and Kazak as part of the *Model Chance Project* instructional materials).

Day 2. This day began with a whole-group discussion of a fair method or a fair game, after which the *Chips Game Task* was introduced. Students played a game that involved randomly drawing a game chip from a bag containing one blue and three red chips. Students were only told that the bag contained chips of two colors: red and blue. Students were split into two groups, and each time a red chip, was drawn, one group received a point, and each time a blue chip was drawn, the other group received a point. The game was played 12 times and prior to beginning each new game, students were asked about their level of certainty regarding the fairness of the game. As additional data from each game played was accumulated, students updated their initial conceptions of randomness, chance variability, and uncertainty. Prior to the group task involving reasoning about combined events in a chance game, two assessment items were given to the participants to answer individually. One item involved combined events and the other questioned the fairness of a game involving combined events. The expectations for dialogic talk were revisited before starting the next task.

The *Chips Game Task* was built on the following idea: When events occur randomly, we cannot be certain about what will happen next, but we can analyze and compare the probabilities of particular events provided that we know enough about all the possible outcomes that could happen. When introducing the task, we first asked the groups to discuss whether the following game was fair:

> There are two bags containing game chips of two colors—red and blue. To play the game, you will randomly select a chip from each bag. If the chips are the same color, this group will win. If they are different color, the other group will win.
>
> *Game 1*
> Bag One: Three red chips, One blue chip
> Bag Two: One red chip, Three blue chips

In this group work, students were expected to state their initial hypothesis (prediction) on whether the game was fair or not, along with an explanation of their reasoning. They were also asked to rate their level of confidence in their hypothesis using a scale of 0%–100% (see worksheet in Appendix 1). Students were asked to evaluate their confidence level because, as emphasized in the SRTL-8 theme, assessing confidence about an uncertain event is seen as essential in making predictions and conclusions about the reasonableness of patterns recognized in data. After this, they were asked to physically play the game as many times as they felt that they needed to in order to gather enough data to support or revise their initial hypothesis, and re-evaluate their level of confidence. Students also modeled the game in TinkerPlots™, to collect more data to test their hypothesis through visualizing the probability of two events in the graph, and again re-rate their confidence level (Here too students were allowed to decide how many trials they felt they needed to carry out.) At the end, the students needed to provide an explanation for the empirical results with the expectation that they would work out all the possibilities for the combined event and link the sample space to the empirical distribution (with some scaffolding if needed).

Students also completed the following variations of the task in a similar format:

- Can you make the game (explained above) fair?
- Game 2—Bag One has four red chips and Bag Two has two red chips and two blue chips. Is it a fair game?
- Can you design a fair game using five chips in each bag; using the blue and red chips and the same rules?
- Can you design a game so that the mixture (i.e. ‘red, blue’ or ‘blue, red’) will always win?

Day 3. Students were introduced to the *Random Bunny Hops Task* (Kazak, Fujita, & Wegerif, 2014): “Suppose there are a number of bunnies on land and each bunny can choose randomly to hop only right or left. For each hop, bunnies are just as likely to hop right as left. We want to know where a bunny is likely to be after five hops.” Following a class discussion about how to decide which side the bunnies

might hop, students were asked to make their initial predictions: "Imagine that a bunny is standing on a number line at 0. You flip a coin to decide which way the bunny hops. If the coin lands heads up, it hops one step to the right (i.e., one step along the positive direction). If the coin lands tails up, it hops one step to the left (i.e., one step along the negative direction)." With the objective of students' articulation of uncertainty in the context of making informal statistical inferences, they were expected to predict, produce and analyze data, and compare simulation data with a population model using the simulation features of TinkerPlots™.

2.5 Analysis and Results

Sociocultural discourse analysis was used to analyze the qualitative aspects of the data as detailed by Mercer (2004). The focus of the analysis was the talk of students working jointly on computer-based activities in pairs or groups. Video-recordings of each group's work were viewed to identify key episodes of talk that led to a new insight in students' reasoning during their joint activity. Selected transcribed excerpts of joint activity in the context of the *Chips Game Task* were analyzed in detail to show how students' emerging reasoning about uncertainty was supported by the combination of talk and the use of TinkerPlots™. The method of analysis adapted from Mercer (1996) involved two levels that emerge from a socio-cultural perspective: linguistic and psychological. At the linguistic level, the talk was examined in terms of kinds of speech acts observed in students' exchanges, such as asserting, challenging, and explaining, and students' responses and reactions to each other's talk. At the psychological level, the talk was analyzed as thought and action, such as the visible pursuits of emerging reasoning relevant to uncertainty through the talk in combination with other tools—TinkerPlots™ and physical materials.

The findings of the study are described and discussed in this section by focusing on a close examination of four episodes (from Day 2) where new insights into students' reasoning about uncertainty were identified during the joint activity for one group: Gabby, Blair, and Flora.

2.5.1 Episode 1: Students' Initial Reasoning about Uncertainty in the Fairness of a Game

In Excerpt 1, Gabby, Blair, and Flora were jointly working to decide whether Game 1 was fair or not in order to make their initial prediction on the worksheet (Appendix 1). After a demonstration of the game with the bags by Taro (researcher), the group seemed to agree that the mixture would occur more often and the game was not fair. However, Gabby later inclined to change her idea about the unfairness of the game. Blair disagreed with her and wanted to explain why she thought the game was not fair to Gabby.

01	Taro:	Is it fair?

02	Gabby:	No, it is not fair. Surely it is not fair. Actually, wait, no it is fair.
03	Blair:	No, it is not fair . . . Because look [*Blair tries to get the chips out of each bag but some of them fall off the table. Gabby covers her face and turns over to the worksheet*].
04	Gabby:	Look look look [*looking at each bag's content drawn on the sheet. Blair puts the chips together in two groups as they were in the bags on the table*]. Look so there is. So there is four blues and there is four reds [*pointing to the drawing of the bags on the worksheet*].
05	Blair:	Yeah, but they are in different bags.
06	Gabby:	Yeah I knew that [*Blair moves the groups of chips right above the drawings of the bags on the worksheet*]. But look, look if you got that, and then like, so you've got three chance. Look if you pick the blue for instance [*Gabby still pointing to the chips on the drawing*]. Listen [*tapping Blair's arm, trying to get her attention*] if you pick the blue for instance, then you get three other chances, [*a moment of thought, a new insight*] actually [*hesitating about her answer now*] . . .
07	Blair:	Look. Because look yeah.
08	Gabby:	No it is not fair because there is three [*pointing to the first bag on the worksheet*].
09	Blair:	Yeah okay. [*Pointing to the red chips from Bag 2 on the table*] If I pick the most common one from here [*Bag 2*] is obviously red [*picking a red chip*] and the most common one in here [*Bag 1*] is obviously blue, [*picking a blue chip*], then [*showing both chips in her hands—one blue one red*] they are different.

Students began to make their initial prediction based on subjective probabilities as they used their own personal beliefs about the situation. In particular, Gabby and Blair had different opinions regarding the fairness of the game. Each tried to assert their own ideas with some reasoning (lines 03–06) as seen in disputational talks (Mercer, 2004) and to convince the other. Gabby thought the game was fair because there was an equal number of red chips and blue chips in total with a similar reasoning to the equiprobability bias (Lecoutre, 1992). She seemed to think about the event outcomes additively (i.e. four blue chips and four red chips in total) rather than to use multiplicative reasoning needed to evaluate the probability of the combined events (i.e., six same color chips and ten mixed color chips). Only after Blair challenged her idea, Gabby made a shift in her understanding of the combined nature of outcomes from each bag (line 08). By listening to Blair's explanation she began to

see the outcomes of the game from a different perspective and, as a result, changed her mind.

2.5.2 Episode 2: Students' Reasoning about Uncertainty After Playing the Game

On the worksheet, the group marked their confidence level for the unfairness of the game at 75% after their initial prediction. In the excerpt below, the teacher researcher asked students again how confident they were when they finished playing the game 30 times with the bags. Even though they all felt more confident about the unfairness of the game after the results (same color=13, mixed color=17), their responses varied with no explicit reasoning without teacher prompt.

10	Sibel (Teacher Researcher):	Now what do you think after playing the game?
11	Gabby:	I think I am between 80 and 90 because I still, you still could be.
12	Flora:	Yeah you still produce like pick them up bunch of time.
13	Gabby:	I think I am about 90.
14	Blair:	I think I am about 95.
15	Sibel (Teacher Researcher):	Okay, can you explain why you think 95? Can you agree?
16	Gabby:	Well because of our results [*looking at the worksheet*].
17	Blair:	Because of our results [*they got 13 same color, 17 mixed color*].
18	Gabby:	[*Blair's chair slams into the tripod with noise*] Well we already thought it was 75 and now the results proved that. Now we are very sure but we are not completely certain.
19	Sibel (Teacher Researcher):	Okay, so pick which one is in there [*referring to the scale on their worksheet*].
20	Blair:	Should we get the 90?
21	Gabby:	Which one do you think Flora?
22	Flora:	80 or 90 I would say.
23	Blair:	Eh, make Gabby 90.
24	Flora:	Oh, okay [*she marks it on the scale on the worksheet*].
25	Gabby:	No wait what do you think? [*asking Flora*]
26	Flora:	It is your 90.

27	Gabby:	You sure?
28	Flora:	Yeah.

When Gabby and Blair came to an agreement for their joint decision, students' prediction or hypothesis and explanation about the unfairness of the game seemed reasonable. Their initial confidence level marked at 75% on the scale indicated that they were not yet certain that the game was unfair. After playing the game 30 times, the results came out in favor of the mixture supporting their hypothesis and hence each student individually stated a higher confidence level—between 80% and 95%. Given the data, Gabby's comment, "Now we are very sure but we are not completely certain" (line 18) suggested that their subjective probabilities were updated but the uncertainty in their personal degree of confidence about the unfairness of the game did not completely disappear. Since the talk between students tended to be more cumulative (in the sense of Mercer, 2004) without explicit elaboration, it was difficult to speculate about why they particularly chose 90% confidence level based on the data they had generated.

2.5.3 Episode 3: Students' Reasoning about Uncertainty through TinkerPlots™ Simulations

In the following excerpt, students' talk was around the TinkerPlots™ model of Game 1 that they built to gather more data to test their initial theory about the unfairness of the game. As seen in Figure 2.1, the *Sampler* (left-side) consisted of two mixer devices, one including one red and three blue balls, and the other including one blue and three red balls. These two devices represented the number of red and blue chips in the two bags. The students opted to set the number of trials (*Repeat* value) to 1000. The *results table*—to the right of the *Sampler*—displays the sampled outcomes for each of the repetitions as they are drawn. The plot (right-side) shows the frequency and percentage of the combined outcomes, "the mixed color" and "the same color," for 1000 trials based on the combined outcomes for the two draws: "Blue, red" and "red, blue" then "blue, blue" and "red, red".

In this activity, Gabby was controlling the mouse while others were watching. When they saw the initial simulation results on the screen (mixed color=62%, same color=38%; as in Figure 2.1), Gabby inferred the outcome of "blue and red is probably more likely…I will run it again." In order to be 100% certain that the game was unfair, Gabby later suggested that they run a few more simulations and also increase the number of trials. She also asked other group members for their opinions to involve them in decision-making.

29	Gabby:	I think now we are like. Are you like what percent oh no, are you like hundred percent certain now?
30	Flora:	Yeah, ninety five to hundred.
31	Gabby:	What do you think Blair? Are you like hundred percent sure?

32	Blair:	Yeah. Yeah I think I am about hundred.
33	Gabby:	Because look. We have done it.
34	Blair:	Totally confident
35	Gabby:	Wait we can do it a couple of more times before we say that. So wait should we change this number? Should we change it to three hundred? No three thousand I mean.
36	Flora:	Yeah.
37	Gabby:	Okay. Woo, umm [*looking at the 3000 results that appeared on the plot*].
38	Blair:	Sixty three percent, thirty seven percent.
39	Gabby:	Yeah, they are getting closer [*running the Sampler several times looking at the results on the plot, 62%–38%, 62%–38%, 62%–38%, 61%–39%, 63%–37%, 62%–38%, 63%–37%, 64%–36%, 62%–38%, 63%–37%, 64%–36%*]

Collecting more data very quickly through their TinkerPlots™ model enabled the students to further investigate their initial prediction about Game 1. After the initial simulation results from 1000 trials, students expressed more confidence. When Blair said that she was "totally confident" (line 34), Gabby pointed out that they would need more data before they could become 100% certain (line 35). The group then agreed on running a few more simulations and increasing the number of trials to 3000. This was a major step towards an understanding of the relative frequencies as estimates of probabilities as in Bernoulli's theorem mentioned in Section 5.3 and using them to revise their subjective probabilities. Here the talk stimulated by the TinkerPlots™ simulation results seemed to help these students see the relevance of a large number of trials with several iterations. Gabby's observation of how stable the percentages of the same and different color chips got in the repeated trials showed an insight into an important concept relevant to reasoning about uncertainty.

2.5.4 Episode 4: Students' Reasoning about Uncertainty in Designing a Fair Game

When students were provided with two bags, five red and five blue chips, and asked to design a fair game by placing five chips in each bag, they jointly worked on the task and came up with three red and two blue chips in bag one and two red and three blue chips in bag two. To test their hypothesis, they built a model of their game in TinkerPlots™ and ran it several times with 1000 trials. The results were mostly 52% mixed color and 48% same color, respectively. When the results from one simulation yielded 51%–49%, Flora and Gabby cheered, "Yeay, 51 and 49!" Then Gabby kept clicking the run button repeatedly, watching the results compile on the plot. It seemed like they were looking for the results close to 50%–50%. When asked whether or not

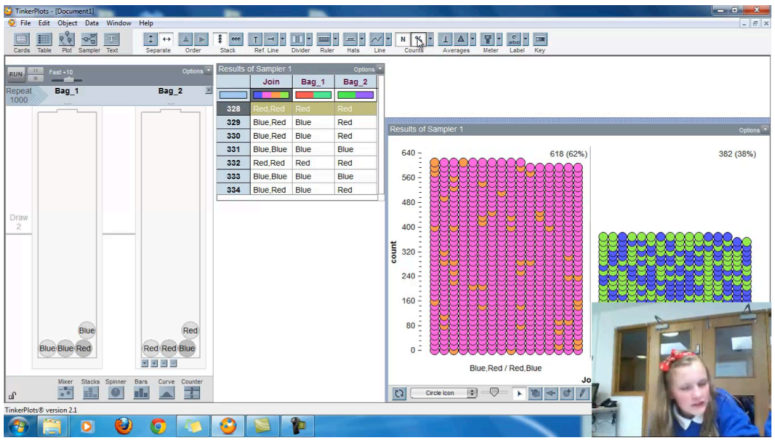


Figure 2.1. Screenshot of students' TinkerPlots™ model and simulation for Game 1.

their game was fair, Gabby said "Yes" with confidence but Blair added, "It's roughly . . . It feels like it's as close as you can because there is an odd like number in each bag." This uncertainty due to the results that were close enough to 50%–50% but not quite even and slightly in favor of the mixture (Blair called this "the fairest game") motivated a further joint exploration that led to a theoretical analysis of the number of possible outcomes for each combined event (mixed color and same color).

In the excerpt below, the teacher researcher joined the group, which was trying to make sense of the simulation results for the game that they initially thought of as fair. This time their joint task developed around the physical material—a group of two blue chips and three red chips and another group of two red chips and three blue chips on the table similar to the contents of the bags in their game. While Gabby still thought it was a fair game, Blair disagreed with her and gave a reason by focusing on the number of red and blue chips in each bag and the amount of mixed color chips that one could get in the game. By manipulating the chips on the table, Blair offered a new idea for why the game could not be fair. Through scaffolding by the teacher researcher, Gabby then took on Blair's idea and began to count the number of possible outcomes for the same color and the mixed color (see Figure 2.2).

40	Sibel (Teacher Researcher):	So when you put two like this, and then like this [*a group of one red chip and three blue chips and a group of one blue chip and three red chips*] huh?
41	Gabby:	Yeah. That's fair.
42	Sibel (Teacher Researcher):	That is fair?

43	Blair:	I don't think it is completely fair, because, there is, like a different amount of reds in each bag, a different amount of blues, but then there is the same amount of the opposite color, like, so, in each bag.
44	Sibel (Teacher Researcher):	Oh what does that mean? Interesting, did you follow her? Gabby?
45	Gabby:	Yeah.
46	Sibel (Teacher Researcher):	Does it matter? She says that they have . . . the same number of, different color?
47	Blair:	The same amount, there is the same number of different colors, like three blues and three reds, like that, but then, there is not the same amount of the same colors, because there is two and three.
48	Sibel (Teacher Researcher):	Huh, in each bag . . .
49	Blair:	So it is almost impossible to get them the same, because if you like move this one for this one and you'd have . . . [*showing two groups: one with one blue chip and four red chips and another with one red chip and four blue chips*]
50	Gabby:	Wait how did you . . . [*rearranging the chips into the original groups—two blue chips and three red chips; two red chips and three blue chips—and starting to count*] there is, one . . . wait, one, two, three, four. No, wait . . . Wait . . . wait wait three, four, five. So . . . so there is five. There is ten times that you can get the same, and then . . .

In this task of designing a new fair game with five chips in each bag, the role of the computer was not directly useful in making a shift in students' reasoning about uncertainty, but raised the need for further exploration. More specifically, interpreting the results generated in TinkerPlots™ required an understanding of sample-to-sample variability versus variability due to the chance setup since the simulation outcomes in percent (with 1000 trials) were close enough to even, like 52%–48% or 51%–49% or even occasionally 50%–50%. Therefore, a theoretical approach was needed to distinguish whether the game was actually fair or not. The previous excerpt illustrated how students' talk, as well as teacher's scaffolding, led to a new insight in students' reasoning about uncertainty. Especially, after Blair's idea of comparing the possible number of same color chips and the possible number of mixed color chips (lines 43 and 47) and the questions posed by the teacher researcher (lines 44 and

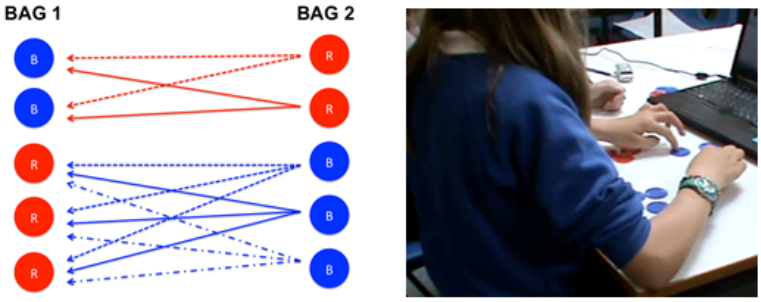


Figure 2.2. An illustration of Gabby's way of counting the number of possible mixed color chips they can get in the game.

46), Gabby seemed to have a new way of seeing the task by reflecting through taking an outside perspective, which was considered as a dialogic switch (Wegerif, 2013). Eventually, she found that there were 12 different ways to get two chips of the same color, but there were 13 different ways to get two chips that were different colors. This led her to conclude that, "there is one more chance that you will get [chips of different colors]." The observed trajectory of students' reasoning that led to this conclusion supports the link between the initial subjective approach used to decide the fairness of the game, and the frequentist and classical approaches to probability.

2.6 Discussion and Implications

The research questions of this study were: (1) How does the combination of using TinkerPlots™ and dialogic interactions in small groups promote students' reasoning about uncertainty in making informal inferences about random events? (2) What are the dialogic mechanisms that help support students' reasoning in the joint activities? The research data were analyzed by focusing on the design element supporting students' informal inferential reasoning—tasks, tools, and scaffolds (student talk) (Makar et al., 2011). This section first discusses the key findings from these analyses as they relate to the subjective approach to probability, scaffolding (through tools and talk), and themes of the dialogic processes. Then, implications for both teaching and research are provided.

2.6.1 Three Emerging Themes in Promoting Students' Reasoning about Uncertainty

Subjective approach to probability. This research has addressed the call for more research on investigating subjective probability and how students understand it (Jones et al., 2007). The findings of the exploratory study attempted to build connections between different views of probability (classical, frequentist, and subjective) and imply that focusing on subjective probabilities can strengthen the link between probability

and informal statistical inference. In a Bayesian-like approach to informal statistical inference adopted in this study, subjective probability is basically considered as, "a person's uncertainty about the occurrence of an event, about the consequences of an action" (Huber & Huber, 1987, p. 304). From the analyses, it became apparent that asking students to begin with a hypothesis (prediction) about the fairness of a game and to rate their level of confidence in it helped them to develop their intuitions leading to conceptualizing subjective probability. More specifically, the empirical data obtained from simulating the game repeatedly were consistently used by the students to revise both their initial hypothesis and confidence level (e.g., Episodes 2 and 3). When there was a higher uncertainty about the occurrences of the combined events ('same color' and 'mixed color') in the empirical results (e.g., Episode 4), the theoretical analysis of all possible outcomes of the combined events (i.e., counting activity as seen in lines 40–50) provided students with additional insight about their initial hypothesis. In this way, they were able to strengthen their level of certainty in relation to the hypothesis about the fairness of the game.

Scaffolding (through tools and talk). TinkerPlots™ and talk played a major role in scaffolding students' reasoning about uncertainty at different stages of the *Chips Game Task*. In particular, the episodes presented in this chapter show how student talk was promoted and evolved to support their reasoning through the use of TinkerPlots™ software. While initially disputational and cumulative types of talk were observed (lines 1–9 and 10–28), in later episodes students switched to exploratory talk, offering explicit reasons for their arguments (e.g., lines 40–50; particularly in Blair's statements). As found in previous studies, it was this exploratory talk that helped students shift their understanding of the problem. Note that Blair's explicit reasoning combined with support by teacher scaffolding led Gabby to count all possible outcomes using the red/blue chips and then to change her mind about the fairness of the game.

Throughout the task, TinkerPlots™ was central to students' reasoning as it scaffolded their talk, including arguments and reasons, and peer interaction by enabling them to collect a large amount of data very quickly and simultaneously see the results in the plot while testing their initial hypothesis. The simulation data in TinkerPlots™ also stimulated students' reasoning as they were updating their level of confidence about their hypothesis (whether the game is fair or not). In designing a fair game with five chips in each bag (Episode 4) particularly, the simulation results from repeated trials with a large amount of data in TinkerPlots™ helped one of the students interpret the game as "roughly" fair. Her comment, "It feels like it's as close as you can because there is an odd like number in each bag," suggested an uncertainty about the fairness of the game to some extent, which then prompted the need for a theoretical analysis of all the possible outcomes to explain the small difference in the occurrences of each combined outcome. Note that during this moment facilitated by talk and software students needed to utilize all three views of probability (subjective, frequentist, and classical) in order to initially recognize the uncertainty, then account for it, and finally resolve it. For example, subjective interpretation of probability came about when students drew upon their personal knowledge or belief to design a fair game and updated their level of confidence based on data. So, at the same time

they relied on frequentist interpretation of probability when they used the relative frequency of each combined outcome in TinkerPlots™ simulations to test their initial hypothesis about the fairness. However, the observed 'almost even' results from repeated trials led a student to question the fairness of their game. Then, an analysis of sample space was needed as seen in the classical approach to probability.

Dialogic processes. In the detailed analysis of the episodes discussed in the previous section, two switches in perspective were noted from a dialogic approach inspired by Bakhtin when a moment of insight helped students shift their reasoning about uncertainty. In lines 1–9, the switch in Gabby's perspective was facilitated by the listening to, and understanding, Blair's justification of why the game could not be fair. Hence the dialogic process behind this switch could be attributed to her ability to see from the perspective of a specific other (Wegerif, 2013). In a later episode (lines 40–50) the dialogic switch in Gabby's reasoning mediated by exploratory talk and teacher scaffolding was interpreted as a result of her reflection by taking an outside perspective (Wegerif, 2013). Behind both switches in perspective is the dialogic quality of the relationship between the students in the group. In these episodes, for instance, we can see trusting, being open to, challenging and critiquing each other's ideas in a constructive way (as opposed to competitiveness), actively listening to each other with understanding, and acknowledging a change of mind by appropriating the perspective of the other.

2.6.2 Implications for Teaching

The emphasis on subjective probability and Bayesian inference in instruction at the pre-university level is very limited or absent. On the other hand, promoting informal statistical inference at early levels of schooling is gaining attention as a result of recent research findings. Subjective probability in the context of a Bayesian-like informal statistical inference seems quite natural, and is worthy of more attention to underpin the link between probability and informal statistical inference, which is important to understand uncertainty. The results also show that students need to experience how different views of probability can be used to quantify uncertainty in mathematics classrooms (see Section 6.1).

The interaction between the use of technology and student talk investigated with a small group of high achieving students in this study is seen as effective for supporting reasoning about uncertainty in making inferences. Promoting this interaction in more typical classroom settings would entail explicitly teaching, discussing, and practicing certain expectations for talking and using computer software for solving problems together as described and suggested by the previous studies conducted in primary mathematics classrooms (Mercer & Sams, 2006; Monaghan, 2005). This study extends these in illustrating how dialogic talk, computer tools and teacher scaffolding can support students' informal inferential reasoning and understanding of probability.

2.6.3 Implications for Research

While several research studies have focused on students' understanding of the relationship between empirical and theoretical probabilities (Ireland & Watson, 2009; Stohl & Tarr, 2002), the link between subjective, empirical, and theoretical probabilities, which historically emerged in relation to each other (Shafer, 1992), has been neglected. Thus, further research is needed on young students' understanding of the relationship between subjective probability and the other approaches to probability.

Finally, the study shows the importance of a dialogic approach inspired by Bakhtin (1986) to explain the switches in perspectives when students have a new insight leading to reason about uncertainty. This dialogic perspective can add to the commonly used theoretical approaches, such as constructivist and socio-cultural, in statistics education research. Hence, future studies should seek to further investigate the advantages of dialogic approach in developing students' conceptual understanding of other statistical topics.

Acknowledgements

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme. The author thanks the STAT-STALK project members, Rupert Wegerif and Taro Fujita; Dani Ben-Zvi, Dave Pratt, Janet Ainley, Hana Manor Braham, and Keren Aridor on their insightful discussion and feedback on an earlier version of the *Chips Game Task*; finally the SRTL-8 participants and anonymous reviewers on their very helpful comments and insights.

References

Ainley, J., Aridor, K., Ben-Zvi, D., Braham, H. M., & Pratt, D. (2013). Children's expressions of uncertainty in statistical modelling. In *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 49–59). Two Harbors, MN: University of Minnesota.

Ainley, J., & Pratt, D. (2001). Constructing meanings from data. *Educational Studies in Mathematics*, *45*, 1–8.

Albert, J. (2002). Teaching introductory statistics from a Bayesian perspective. In B. Philips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics.* Cape Town, South Africa: International Statistical Institute.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays by M. M. Bakhtin* (C. Emerson & M. Holquist, Trans.). Austin, TX: University of Texas Press.

Bakhtin, M. M. (1986). *Speech genres and other late essays* (V. W. McGee, Trans.). Austin, TX: University of Texas Press.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International*

Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (1986). Exploratory data analysis and the secondary stochastics curriculum. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 79–85). Victoria, British Columbia, Canada: University of Victoria Conference Services.

Braham, H. M., Ben-Zvi, D., & Aridor, K. (2013). Students' reasoning about uncertainty while exploring sampling distributions in an "integrated approach"'. In *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 18–33). Two Harbors, MN: University of Minnesota.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. doi: 10.3102/0013189X032001009

Dawes, L., Mercer, N., & Wegerif, R. (2000). *Thinking together: a programme of activities for developing speaking, listening and thinking skills for children aged 8–11*. Birmingham, UK: Imaginative Minds Ltd.

Department for Education and Employment. (1999). *The National curriculum*. London: Author. Retrieved from http://curriculum.qcda.gov.uk/uploads/Mathematics%201999%20programme%20of%20study_tcm8-12059.pdf

Díaz, C. (2010). Psychology students' understanding of elementary Bayesian inference. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics.* Lbjubjana, Slovenia: International Statistical Institute.

Falk, R., & Konold, C. (1992). The psychology of learning probability. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty-first century (MAA Notes #26)* (pp. 151–164). Washington, D.C.: Mathematical Association of America.

Fitzallen, N., & Watson, J. (2010). Developing statistical reasoning facilitated by TinkerPlots. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics.* Lbjubjana, Slovenia: International Statistical Institute.

Gibbs, P., Hanlon, M., Hardaker, P., Hawkins, E., MacDonald, A., Maskell, K., ... Innocent, T. (2013). *Making sense of uncertainty. why uncertainty is part of science*. London: Sense About Science. Retrieved from http://www.senseaboutscience.org/data/files/resources/127/SAS012_MSU_reprint_compressed.pdf

Hacking, I. (1975). *The emergence of probability*. London: Cambridge University Press.

Hald, A. (2003). *A history of probability and statistics and their applications before 1750*. Hoboken, NJ: Wiley-Interscience.

Harradine, A., & Konold, C. (2013). Using data and chance to make conclusions. In *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 5–17). Two Harbors, MN: University of Minnesota.

Hawkins, A. S., & Kapadia, R. (1984). Children's conceptions of probability—a

psychological and pedagogical review. *Educational Studies in Mathematics*, *15*, 349–377.

Huber, B. L., & Huber, O. (1987). Development of the concept of comparative subjective probability. *Journal of Experimental Child Psychology*, *44*(3), 304–316. doi: 10.1016/0022-0965(87)90036-1

Ireland, S., & Watson, J. (2009). Building a connection between experimental and theoretical aspects of probability. *International Electronic Journal of Mathematics Education*, *4*(30), 339–370. Retrieved from http://www.iejme.com/032009/main.htm

Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability: Responding to classroom realities. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 909–956). Charlotte, NC: Information Age Publishing.

Kazak, S., Fujita, T., & Wegerif, R. (2014). Year six students' reasoning about random 'bunny hops' through the use of TinkerPlots and peer-to-peer dialogic interactions. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education: Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).* Flagstaff, AZ: International Statistical Institute.

Kazak, S., Wegerif, R., & Fujita, T. (2013). 'I get it now!' Stimulating insights about probability through talk and technology. *Mathematics Teaching*, *235*, 29–32.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, *12*, 217–230. doi: 10.1007/s10758-007-9123-1

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1). Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1/

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., … Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning*, *13*(1-2), 68–86. doi: 10.1080/10986065.2011.538299

Konold, C., & Miller, C. (2011). *TinkerPlots™ 2.0 beta.* Amherst, MA:University of Massachusetts.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289. doi: 10.2307/749741

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

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173. doi: 10.1080/10986065.2011.538301

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 1–4. doi: 10.1080/10986065.2011.538291

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, *6*(4), 359–377. doi: 10.1016/S0959-4752(96)00021-7

Mercer, N. (2004). Sociocultural discourse analysis: Analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics*, *1*, 137–168.

Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, *20*(6), 507–528. doi: 10.2167/le678.0

Monaghan, F. (2005). 'Don't think in your head, think aloud': ICT and exploratory talk in the primary school mathematics classroom. *Research in Mathematics Education*, *7*(1), 83–100. doi: 10.1080/14794800008520147

Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, D.C.: National Academy.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107–129. Retrieved from http://www.stat.auckland.ac.nz/serj

Rossman, A. (2008). Reasoning about informal statistical inference: A statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19. Retrieved from http://www.stat.auckland.ac.nz/serj

Shafer, G. (1992). What is probability? In D. C. Hoaglin & D. S. Moore (Eds.), *Perspectives on contemporary statistics* (Vol. 21, pp. 93–105). Washington, D.C.: Mathematical Association of America.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 205–237). Dordrecht, The Netherlands: Kluwer Academic Press.

Stohl, H. (2005). Probability in teacher education and development. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 345–366). New York: Kluwer/Springer Academic Publishers.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior*, *21*(3), 319–337. doi: 10.1016/S0732-3123(02)00132-3

TechSmith. (2011). *Camtasia for Mac 2*. Okemos, MI: TechSmith Corporation.

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wegerif, R. (2013). *Dialogic: Education for the internet age*. New York: Routledge.

Wegerif, R., & Dawes, L. (2004). *Thinking and learning with ICT: Raising achievement in primary classrooms*. London, UK: Routledge.

Wickham, L. (2008). Generating mathematical talk in the key stage 2 classroom. In M. Joubert (Ed.), (Vol. 28, pp. 115–120).

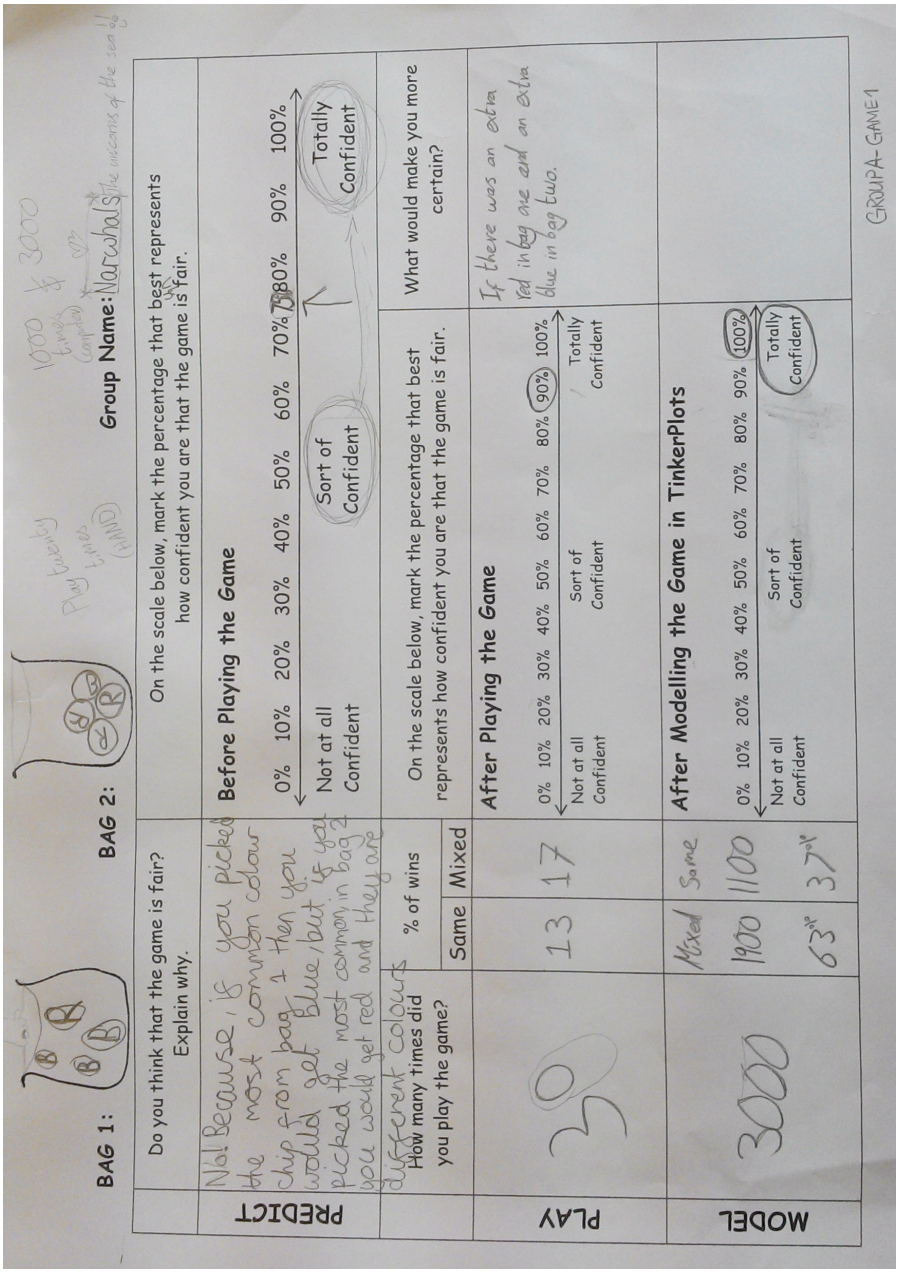
Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. Retrieved from http://www.stat.auckland.ac.nz/serj

Appendix 1: Chips Game Task Worksheet

BAG 1: **BAG 2:** **Group Name:**

PREDICT	Do you think that the game is fair? Explain why.		On the scale below, mark the percentage that best represents how confident you are that the game is fair/unfair.	
			Before Playing the Game 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% Not at all Confident — Sort of Confident — Totally Confident	
PLAY	How many times did you play the game?	% of wins Same / Mixed	On the scale below, mark the percentage that best represents how confident you are that the game is fair/unfair.	What would make you more certain?
			After Playing the Game 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% Not at all Confident — Sort of Confident — Totally Confident	
MODEL			**After Modelling the Game in TP** 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% Not at all Confident — Sort of Confident — Totally Confident	

Appendix 2: Chips Game Task Worksheet Completed for Game 1 by Flora, Gabby, and Blair.



CHAPTER 3

STUDENTS' ARTICULATIONS OF UNCERTAINTY IN INFORMALLY EXPLORING SAMPLING DISTRIBUTIONS

HANA MANOR BRAHAM AND DANI BEN-ZVI

The University of Haifa, Israel

Abstract

The uncertainty in making general conclusions about what is the "true", long run situation is a core idea of formal and informal statistical inference. We analyzed students' articulations of uncertainty during their first steps in exploring sampling distributions in a TinkerPlots™ inquiry-based learning environment. A new "integrated modeling approach" (IMA) was implemented to help students understand the relationship between sample and population. We focused this case study on two students (age 13, grade 7) who had previously participated in the *Connections Project* exploratory data analysis (EDA) activities. Over seven main stages, the students' articulations of uncertainty were shaped by two different views in the way they observed and manipulated the sampling distributions: (1) a move from a global to a probabilistic view and (2) a move from a local–deterministic to a quasi-probabilistic view.

Keywords: Informal statistical inference; Sampling distribution; Modeling; Statistics education; Statistical reasoning; Uncertainty

3.1 Overview

Quantitative information is everywhere, and statistics are increasingly presented as a way to add credibility to advertisements, arguments, or advice. Being able to properly evaluate evidence (data) and claims based on data is an important skill that all students should learn as part of their educational programs. Given the prevalence of surveys in media, statistically literate adults should be able to properly interpret and evaluate messages that contain statistical elements (Gal, 2004). They have to be aware that the meaning of certain statistical terms used in media reports of surveys (e.g., random, representative, reliable, likelihood) may be different than their everyday meaning. Furthermore, they need to have a sense of the power to draw reliable statistical inferences from samples and be able to appreciate the purpose of such activity. In this chapter, we address these issues, and in particular a key phenomenon of articulations of uncertainty, with the help of data coming from the long-term design and research *Connections Project*. In this project, children in grades 4–9 have been performing statistical reasoning in an inquiry-based and technology-enhanced environment. While looking at a pair of seventh graders exploring sampling distributions, we describe the development of their ability to talk and reason about uncertainty.

3.2 Problem

The recognition that judgments based on sample data are basically uncertain is a key idea of formal and informal statistical inference (ISI). Informal inferential reasoning (IIR) includes various elements such as, accounting for, examining, controlling, and quantifying the uncertainty resulting from drawing a random sample in order to infer from it to a population (Pfannkuch, Wild, & Parsonage, 2012). To understand the uncertainty involved in taking a sample, one needs to envision a process of repeated sampling and its relation to the individual sample with the result of a sampling distribution (Saldanha & McAllister, 2014). However, research suggests that students tend to focus on individual samples and statistical summaries of them instead of how collections of sample statistics are distributed (Saldanha & Thompson, 2002), and that students tend to respond in a deterministic way while reasoning about data (Ben-Zvi, Aridor, Makar, & Bakker, 2012). A significant challenge for statistics educators is to enable students to develop a sense of the uncertainty involved in making conclusions from a sample data to a population, and to encourage students to make connections between a process of repeated sampling and the individual sample in order to understand this uncertainty. In this chapter we describe our response to this challenge by experimenting with the "Integrated Modeling Approach" (IMA).

3.3 Literature and Background

In this literature review, we first address the type of uncertainty we refer to in our chapter in relation to formal and informal statistical inference. Second, with respect

to the challenge of supporting students' articulations about uncertainty while making ISIs, we review studies on reasoning about sampling that show students' difficulties in dealing with that uncertainty. In responding to this challenge, we also describe the technological tool that underpins the design of the activities of our research. Lastly, we review two settings that have been used in the research to examine students' IIR and situate our learning environment in relation to those two settings.

3.3.1 Formal Statistical Inference

To learn about real world situations, we collect data and move "beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and that conclusions are uncertain" (Moore, 2007, p. xxviii). In its simplest form, the question of statistical inference deals with the manner of making general conclusions about what the true, long run situation is actually like, based on outcomes of a random sample that can be collected only once. Given only the sample evidence, the statistician is always unsure of any assertion he makes about the true state of the situation. The theory of statistical inference provides ways to assess this uncertainty and to calculate the probability of error in a particular decision. For example, the probability of getting a statistic as extreme as or more extreme than a specific one, given a specified null hypothesis can be calculated. For being able to calculate such probabilities, it is necessary first to make connections between a statistic that is derived from an individual sample and a process of repeated sampling.

3.3.2 Informal Statistical Inference (ISI)

In order to give students a sense of the power of drawing reliable inferences from samples, and given that statistical inference is challenging for most students (Garfield & Ben-Zvi, 2008), ISI and IIR have recently became a major focus of research (e.g., Gil & Ben-Zvi, 2011; Makar & Ben-Zvi, 2011; Makar, Bakker, & Ben-Zvi, 2011; Pratt & Ainley, 2008). ISI is a data-based generalization, which does not involve formal statistical procedures, that includes an articulated component of uncertainty (Makar & Rubin, 2009). IIR is the reasoning process that leads to the formulation of ISIs. IIR includes "the cognitive activities involved in informally drawing conclusions or making predictions about 'some wider universe' from patterns, representations, statistical measures and statistical models of random samples, while attending to the strength and limitations of the sampling and the drawn inferences" (Ben-Zvi, Gil, & Apel, 2007, p. 2).

3.3.3 Sampling Distribution

Understanding the logic behind such ISIs includes "juggling" several ideas, such as: random sampling, sampling variability and relationship between sample and population. However, students can hold contradictory ideas about these relationships: (1) sampling representativeness—the expectation that a sample taken from a population will have characteristics similar to that population; and (2) sampling variability—

the expectation that different samples taken from a population vary from each other and do not match the population (Rubin, Bruce, & Tenney, 1991). Students that hold the first idea have almost an absolute certainty in relation to the sample representativeness of the population. Students that hold the second idea have a big uncertainty in relation to the sample representativeness of the population. Rubin et al. (1991) showed that senior high school students did not integrate these ideas in their reasoning about distributions of sample outcomes, but held instead one idea at a time depending on the given task.

To integrate these contradicting ideas, students need to envision a process of repeated sampling (Shaughnessy, 2007) with the result of a sampling distribution—the idea "that the values of a statistic are distributed somehow with a range of possibilities" (Thompson, Liu, & Saldanha, 2007, p. 209). Engaging students with sampling distributions might support an emergence of a probabilistic view—the ability to make probability statements about sample statistics in order to control or quantify uncertainty.

However, sampling distribution is one of the most difficult concepts in learning statistics (Saldanha & Thompson, 2002). As a result of failing to develop a deep understanding of sampling distribution, students often develop a procedural knowledge of statistical inference. Garfield, delMas, and Chance (2005) listed what students should understand about sampling distributions including, for example, as sample size (n) gets larger, variability of the sample means gets smaller, and students should be able to interpret or apply areas under curve as probability statements about sample statistics. They also listed what students should be able to do with this knowledge about sampling distributions, such as describe the size of the standard error of the mean and the likelihood of different values of the sample mean.

We suggest that in order to develop deep understanding of informal statistical inference, students should be exposed informally first to ideas of sampling variability and sampling distribution over several years starting at early age. Learning these complex ideas in early years is enabled nowadays with technological advancements. Next, we situate the rationale of the learning environment of this study by reviewing a technological tool that guided the design of students' activities, and two types of settings used in previous studies to develop and study students' IIR.

3.3.4 Learning in a Technology-Enhanced Environment

Technological advancements have led to numerous changes in statistical instruction, including new school curricula that introduce advanced statistical concepts as early as the elementary level (Franklin & Garfield, 2006). Technology enables students to organize and represent data dynamically with less emphasis on calculations. Thus, class discussions or activities may focus on "what if" questions by manipulating graphs and instantly seeing the results (Chance, Ben-Zvi, Garfield, & Medina, 2007). Using technology also enables students to experience and participate in the statistical processes in tangible and dynamic ways, which are not available without technology (Biehler, Ben-Zvi, Bakker, & Maker, 2013). For example, simulations can offer ways to understand ideas of long-run patterns and random processes (Garfield, Chance, &

Snell, 2000). Several studies have demonstrated the advantage of dynamic and innovative technological tools, such as Fathom® and TinkerPlots™ (Konold & Miller, 2011) in developing students' statistical reasoning and in supporting their competence in making general arguments via data-based evidence (e.g., Ben-Zvi, 2000; Paparistodemou & Meletiou-Mavrotheris, 2008).

3.3.5 Research on Students' Informal Inferential Reasoning

Two main types of settings have been used in the research literature to examine young students' informal inferential reasoning. The first, Exploratory Data Analysis (EDA) is a learning environment in which students are engaged in real world data investigations where they create surveys to study some question of interest (e.g., Ben-Zvi, 2006; Pfannkuch, 2006; Makar et al., 2011; Makar & Rubin, 2009). In a study by Ben-Zvi (2006), fifth grade students collected and investigated real data about themselves using TinkerPlots™. Following the growing samples instructional heuristic (Bakker, 2004; Ben-Zvi et al., 2012; Konold & Pollatsek, 2002), the students were gradually introduced to increasing sample sizes in order to support their reasoning about informal inference and sampling. The growing samples task design supported students' informal inferential and sampling reasoning by observing aggregate features of distributions, identifying signals out of noise, accounting for the constraints of their inferences, and providing persuasive data-based arguments.

The second setting is probability-based learning environments (e.g., Konold et al., 2011; Pratt, 2000; Pratt, Johnston-Wilder, Ainley, & Mason, 2008), in which students are engaged in manipulating chance devices, such as coins, spinners and dice. Such settings emphasize how probability is used by statisticians in problem solving. For example, 10-year old students who worked with the *Chance-Maker* microworld were able to understand how empirical probability, theoretical probability and sample size are related to drawing valid inferences (Pratt, 2000). In a study by Konold, Harradine, and Kazak (2007) students built models using computer-based simulations, in order to create reasonable approximations of phenomena, ones that take into account signal and noise.

The first setting has a big potential to improve students' use of data as evidence to draw conclusions: When students work on topics close to their world, which makes the task authentic and relevant, they can gain important insights into how statistical tools can be used to argue, investigate, and communicate foundational statistical ideas. However, those settings might lack probabilistic considerations, which are important for understanding the relationship between samples and populations. The second setting might encourage and develop students' probabilistic reasoning: When students manipulate chance devices, they can easily build probability models of the expected distribution and observe simulation data of the model. Then, they can compare simulation data and empirical data to draw conclusions. This strategy of comparing simulated and empirical data introduces students to the logic of statistical inference and emphasizes the key role played by chance variation in statistical inference. Probability settings, however, might lack aspects of authentic data exploration and might exclude the relevance of the situation.

The activities of this study were designed according to the "Integrated Modeling Approach" (described below) to help students in understanding the relationship between sample and population. This approach intends to integrate these two types of settings. In other words, IMA aims to support students' IIR on authentic data while taking into account probabilistic considerations.

The topic of informal inferential reasoning is not yet sufficiently examined in the literature, and specifically lacks studies on the combination of probabilistic reasoning and making ISIs in authentic contexts. This is the focus of the current study.

3.4 Method

This case study focuses on the question: *How can students' articulations of uncertainty emerge while informally exploring sampling distributions in the integrated modeling approach?* In order to address this question, we closely followed the articulations of a pair of seventh grade students (age 13) as they examined a sampling distribution using the sampler in TinkerPlots™. This study is part of the longitudinal design and research *Connections Project* (2005–2015; Gil & Ben-Zvi, 2011) aiming to develop and study children's statistical reasoning in an inquiry-based and technology-enhanced environment for learning statistics in grades 4–9.

3.4.1 Participants

This study involved a pair of students (grade 7, aged 13), Shay and Liron, in a private school in northern Israel. We selected them since they had high communication and thinking skills which can provide a window to their statistical reasoning. They had already participated in two *Connections Project* experiments. In fifth grade (age 11, 2010), they collected and investigated data about their peers using the first version of TinkerPlots™. Following the growing samples heuristic (Ben-Zvi et al., 2012), the students were introduced gradually to samples of increasing sizes, in order to support their reasoning about ISI and sampling. In sixth grade (age 12, 2011), they engaged in both real world data investigations and model-based investigations using TinkerPlots™ chance devices in order to support their reasoning about ISI and sampling. The first co-author observed and guided the students during eight sessions (about 80 minutes each) over a four-week period.

3.4.2 The "Integrated Modeling Approach" (IMA)

The IMA was developed to guide the design and analysis of experimental tasks (as part of Manor's Ph.D. study) to help students learn about the relationship between sample and population. It is comprised of data and model worlds. In the data world, students collect a real sample, frequently through a random sampling process, in order to study a particular phenomenon in the population. Students choose a research theme, pose questions, select attributes, collect and analyze data, make informal inferences about a population and express their level of confidence in the data. Students

also begin to model real world phenomena using statistics by moving from real world questions to statistical ones. However, they do not necessarily account for probabilistic considerations (e.g., the chance variability that stems from the random sampling process). In the model world, students build a model (a probability distribution) of an explored (hypothetical) population and produce data of random samples from this model. Hence, they pay attention to a model and to the random process that produces the outcomes of samples from this model. Due to randomness, the details vary from sample to sample, but the variability is controlled. That is, given a certain distribution of the population, the likelihood of certain results can be estimated.

In the IMA learning trajectory, students iteratively create connections between the two worlds by working on the same problem context in both worlds. They begin their exploration in the data world (the first dotted trajectory in Figure 3.1) by choosing a meaningful research theme, formulating a question, making an initial conjecture based on their contextual knowledge, building a questionnaire, planning how to draw a sample and collecting real small sample data (represented by a small dotted circle in Figure 3.1). While exploring the sample data they start making sense of it and search for typical characteristics or trends in the data to make ISIs. In the end of this part, they make a second version of their conjecture (the second triangle in Figure 3.1) about the population based both on their contextual knowledge and the sample data results.

As a motivation to move to the model world, the students are asked to express their level of confidence in the sample data that they had collected in relation to the second version of their conjecture and to consider what is the minimal sample size needed to draw reliable inferences about the population with a reasonable confidence level. At this point, the students are first introduced to the model world. They are told: "Imagine you were almighty and could know what characterizes the population. What do you think a random sample from this population would look like? Could you find in this imaginary world the minimal sample size that could represent the population well?"

To do that, the students begin their exploration in the model world (the first lined trajectory in Figure 3.1). They build in TinkerPlots™ a model of the hypothetical population according to their second version of their conjecture and then they simulate sample data from this model. They explore the variability between simulated samples, compare them to the model and gradually enlarge sample size to reduce the variability between samples. Agreeing on the minimal sample size by which they can draw conclusions with confidence, they move again to the data world (the second dotted trajectory in Figure 3.1) to collect real sample data of that size. In the data world, they collect more sample data and formulate the third version of their conjecture (the third triangle in Figure 3.1) about the population based on both their contextual knowledge and the second sample data. They also explain their level of confidence in the data based on what they have learned in the model world.

The continuous trajectory in Figure 3.1 describes integrative transitions between the worlds which might occur to improve the model in relation to different issues, like the dependency between attributes in the model, the shape of distributions of attributes in the model. Our hypothesis is that the IMA can support students' de-

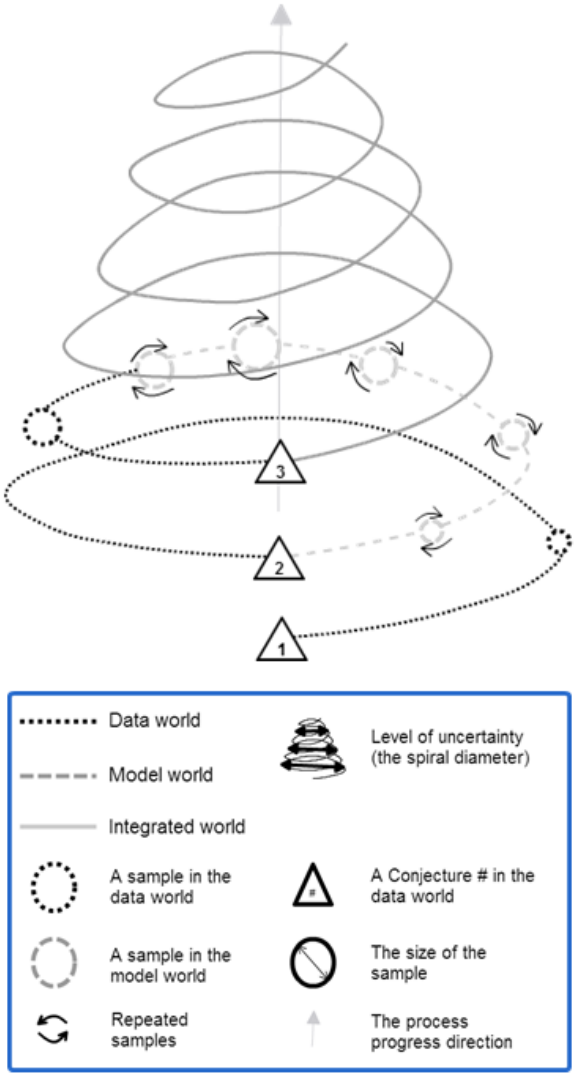


Figure 3.1. The Integrated Modeling Approach (IMA) model.

velopment of reasoning about uncertainty when making ISIs by experimenting with transitions and building connections between the two worlds.

A central feature of the IMA is the use and the study of TinkerPlots™ (Konold & Miller, 2011), a dynamic interactive statistics software developed to support young students' statistical reasoning through investigation of data and statistical concepts. TinkerPlots™ is designed in a "bottom up" manner; that is, it builds on young learners' previous knowledge (Konold, 2002). Using TinkerPlots™, young learners can

start exploring data and creating their own graphs without having previous knowledge of conventional graph types. Unlike other software, in which students choose from a menu of ready-made plots, in this software, students can organize their data using three simple operations: ordering data according to the variable values, separating data into categories, and stacking data.

Using the software, the inquiry process usually does not end with the creation of one representation. The dynamic nature of this software encourages learners to explore data in different repeated representations, testing various hypotheses. In this way, the tool encourages reasoning about data during comparison of distributions, examination of correlations between variables and identification of trends.

TinkerPlots™ also includes a "sampler", which allows learners to design and run probability simulations. The sampler allows learners to build a data distribution of a population and draw random samples from this population in an animated visual way: The items are removed one by one from the population distribution to the sample distribution. Learners can then plot the samples' results, giving a visual representation of the outcomes over many samples.

By using a probabilistic simulator, such as the one in TinkerPlots™, students have an opportunity to explore relationships between data and chance (Konold & Kazak, 2008) by means of one technological tool. They can learn about the sample data using data exploration tools and build a hypothetical probabilistic model of the population from which the sample was taken. They can then examine the sample results in relation to this hypothetical model.

3.4.3 Data and Analysis

We performed a retrospective analysis after each session (to re-direct the next session) and also after the entire teaching experiment was completed. Data collection included students' responses[1] and gestures (captured using *Camtasia*), researchers' observations, and students' artifacts (e.g., data representations that students drew). All students' verbalizations were carefully transcribed. Interpretive micro-analysis (e.g., Meira, 1998)—a microgenetic method (Chinn & Sherin, 2014)—was used to analyze the data. It is a systematic, qualitative and detailed analysis of the transcripts, which takes into account verbal, gestural, and symbolic actions within the situations in which they occurred. The validation of the data analysis was performed within a small group of statistics education researchers (including the co-authors). The researchers discussed, presented, advanced, or rejected hypotheses and interpretations, and inferences about the students' reasoning and articulations. The goal of such an analysis was to infer students' articulations of uncertainty as they explored the sampling distribution. Initial interpretations grounded on data were reviewed by the researchers and triangulated by a group of expert and novice peers. During these triangulation meetings, hypotheses that are posed by the researchers were advanced

[1] All students' conversations were originally said or written in Hebrew, and have been translated for this chapter. As part of the micro-analytic study, we closely examined the meaning of every word to make sure the translation was as close as possible to the original intention of the contributor.

and/or rejected, until a consensus was reached. Triangulation was achieved only after multiple sources of data validated a specific result (Schoenfeld, 2007) to achieve "trustworthiness" (Lincoln & Guba, 1985).

3.4.4 The Setting

In the current study, the actual learning trajectory (a total of about 11 hours, Table 3.1) was comprised of four activities that were designed according to the IMA. In the data world, the students planned a statistical investigation. They chose a research theme, posed a question, formulated a hypothesis, and decided on the sampling method and sample size (Activity 1). In the model world, they built a hypothetical TinkerPlots™ model for the distribution of the population based on their research hypothesis (Activity 2). In order to encourage the students to examine the connections between the two worlds, they were asked "what if" questions about optional real data results while they were exploring the sampling distribution informally (Activities 2 and 3). Finally, the students explored data and models in both worlds by examining the real sample results in relation to their hypothetical models of the population (Activity 4).

Table 3.1
The Actual IMA Learning Trajectory

Act.	Activity Title	Session Title	Session Themes	Statistical Ideas and Concepts	Time (Min.)
0	Preparatory	Preliminary Interview	What do you remember from the *Connections Project* last year?	A sample size required to get significant conclusions about a population, and levels of confidence in results from that sample	40
1	A research plan about teenagers: From a sample to population	Session 1.1: Research plan—Part I	Choose a research theme, pose research question and survey questions, choose population, sampling method and sample size	Considerations in formulating a research question and survey questions, choosing sampling method, random sampling, and sample size (absolute size vs. population percentage)	90
		Session 1.2: Research plan—Part II	Formulate a hypothesis and express confidence level in the results	Research hypothesis (verbal and visual expressions) and informal confidence level	100
2	Music among teenagers: From a population to samples	Session 2.1: Model in TinkerPlots™—Part I	Build and run a TinkerPlots™ model (no relationships between the attributes)	A "simulated database of 30 cases model" for examining whether a sample size of half the population size is similar to the population, informal level of certainty in the sample size	90

Table 3.1 – continued from previous page

Act.	Activity Title	Session Title	Session Themes	Statistical Ideas and Concepts	Time (Min.)
		Session 2.2: Model in TinkerPlots™—Part II	Build and run a second TinkerPlots™ model (with relationships between the attributes)	Level of certainty in sample size, sampling with and without replacement, sample size in relation to population size, sampling variability, representativeness.	90
		Session 2.3: Sampling distribution (includes the episode of this chapter)	Build a third TinkerPlots™ model and explore the sampling distribution	Sampling distribution, probability of sample results, range, mode, degree of inaccuracy, level of uncertainty, accept/reject interval of sample results, and informal hypothesis testing	90
3	Cellular phones among teenagers	Exploring a hidden sampler with samples	Draw ISIs by exploring samples of a hidden sampler	The range (shows the possible inaccuracy of samples) and variability of the sampling distribution, sample size, and frequency of sample results	80
4	Music among Teenagers	Exploring real sample data	Explore real sample data in relation to a sampling distribution and draw ISIs	As the population is bigger, a certain percentage of the population will be more similar to the population, and informal hypothesis testing	60
Total time (minutes):					640

3.5 Results

We focus on the students' discussions during one episode of exploring sampling distributions during Session 2.3 (described in Table 3.1). In order to put this episode in context, we provide first a general description of students' responses during the entire learning trajectory.

3.5.1 General Account

Preliminary interview. In order to evaluate the students' starting point, we interviewed them about their participation in the *Connections Project* in the previous year. During this interview the students discussed informally the sample size required to draw reliable conclusions about a population and levels of confidence in results from that sample. While Shay argued that a large sample (bigger than 100) is required to infer reliably about a population of 600 students, Liron thought that a sample of 100 students (or $1/6$ of the population) was sufficient. Shay explained that one cannot rely on small samples since repeated sampling would yield very different results, while the differences between larger samples would be significantly smaller. Accordingly, they informally estimated their confidence level in a sample size of $1/6$ of the population: Shay—30% and Liron—75%. They concluded the interview by disagreeing on the sample size needed for reliable conclusions: Shay—at least $1/2$ of the population, and Liron—$1/3$ to $1/2$ of the population.

3.5.2 Activity 1. Learning about Teenagers: from a Sample to Population.

In the first activity, we asked the students to plan a research project related to teenagers about a subject that interested them.

Session 1.1: Research Plan (Part I). *Choose a research theme and pose research question and survey questions.* Shay and Liron decided to study music practices and preferences among teenagers. They suggested seven research questions and discussed how to formulate them while raising practical considerations. For example, they discussed two optional formulations: "How much time do you listen to music each day?" versus "How often do you listen to music?"

Choose population, sampling method and sample size. While Liron was interested in studying the topic among seventh grade students in their school, Shay preferred a larger population made of seventh graders that were "culturally similar." Liron, a guitar player, explained that knowing the music preferences of his peers would direct him to play their favorite music type and become more popular among his friends. Ultimately, Shay and Liron decided to study the topic among seventh grade students within their school (about 120 students) and take a random sample of an equal number of children from each seventh grade class, but did not agree about the sample size.

Session 1.2: Research Plan (Part II). *Formulate a hypothesis about the expected results of the research.* Liron and Shay were asked to describe their hypothesis verbally and visually. They created bar graphs for one or two attributes according to their hypothesis. For example, they separately drew bar graphs of the favorite music types among seventh grade students in their school (Figures 3.2 and 3.3).

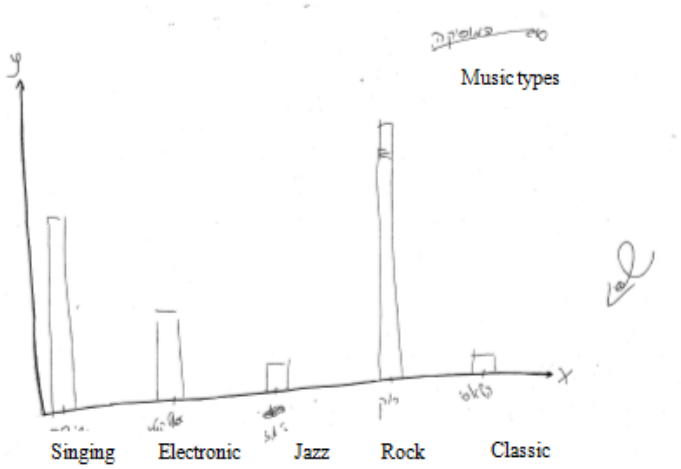


Figure 3.2. Shay's hypothesis regarding favorite music types.

3.5.3 Activity 2. Music among Teenagers: from a Population to Samples

Session 2.1: Model in TinkerPlots™ (Part I). After a short introduction to modeling and simulation in daily life, we asked the students: "Imagine that you had the power to find out what characterizes the music listening habits of teenagers and had findings about all the seventh graders in your school: a) What will the distribution of these findings look like? b) What will a random sample from this distribution look like?" Using the TinkerPlots™ sampler, Shay and Liron built a model that included seven attributes according to their hypotheses. They used the "stacks" device[2] and added the "show count" option (Figure 3.4).

At this point, Shay had a clear plan regarding the method of using the TinkerPlots™ sampler, while Liron was still trying to understand the meaning of the model they built (Figure 3.4). Before they drew a random sample from this model, Shay explained that they had created a simulated database of 30 cases (note that the stacks

[2]The TinkerPlots™ stacks device is used for entering a large number of duplicate elements. In a stacks device, one can simply edit the labels along the bottom axis. Then, using a cursor, one can drag the top of each stack to adjust the number in that stack and can also choose the "Show Count" option and edit the number that appears above each element type.

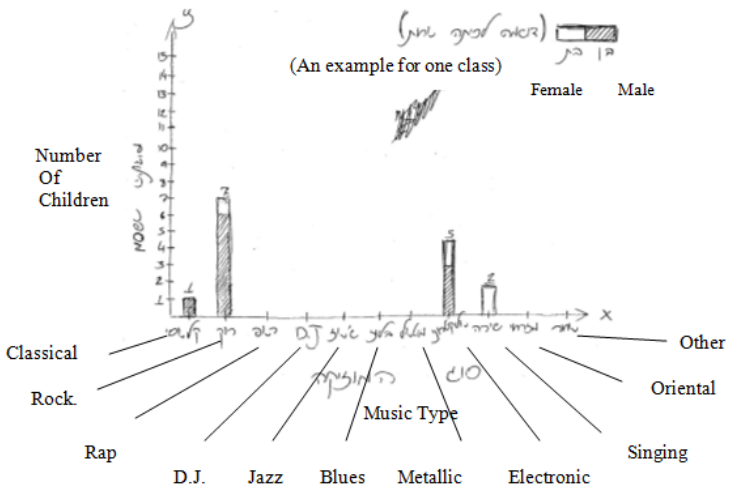


Figure 3.3. Liron's hypothesis regarding favorite music types by gender.

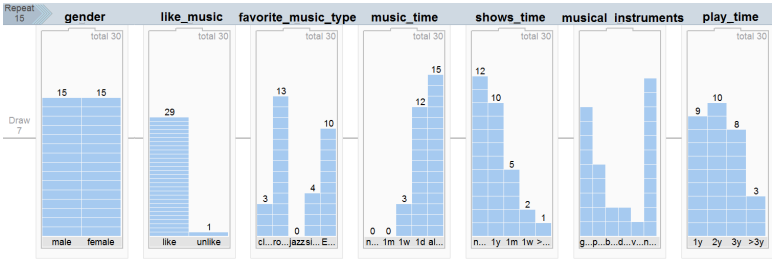


Figure 3.4. Shay and Liron's first model in TinkerPlots™.

in each device sum up to 30) to examine whether a sample size of half the population would be similar to the population. They drew random samples of size 15 and compared the attributes of the sample to the database one at a time. When Shay found a discrepancy between the sample result and the database, he claimed that the sample was not reliable.

Liron wondered how the sampler would represent the relationship between two attributes (musical instrument and frequency of listening to music). Shay explained to him that there was no reason to expect a relationship between the attributes because they had not yet constructed a relationship model (Figure 3.4) and the sampler had drawn the data randomly. Furthermore, Liron seemed to be confused between real data and simulated data. For example, having explored several attributes in the simulated sample, he argued that his hypotheses were correct. Shay disagreed: "It's not true because we haven't yet done the research. It is obvious that if we take a sam-

ple from our hypothesis [model], it will be similar to our hypothesis." Liron agreed, with a bit of confusion.

Session 2.2: Model in TinkerPlots™ (Part II). Shay and Liron started building a second model with seven attributes and 30 cases, but this time they added relationships between attributes according to their hypotheses. After they had entered two attributes (gender and favorite music type) to their model (Figure 3.5), Shay drew a random sample of size 15 (Figure 3.6).

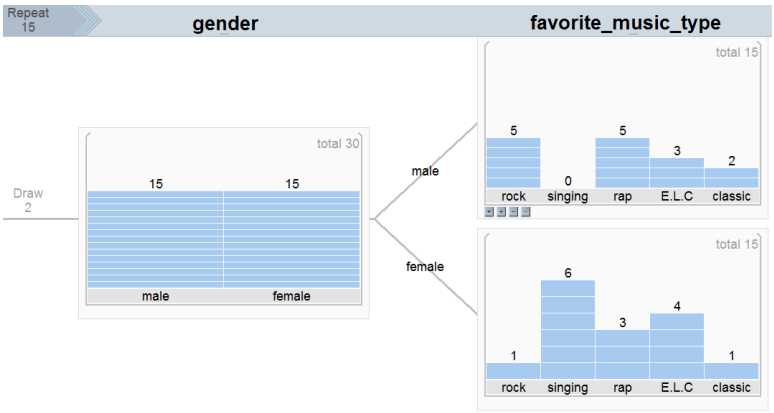


Figure 3.5. Shay and Liron's second model with two interrelated attributes.

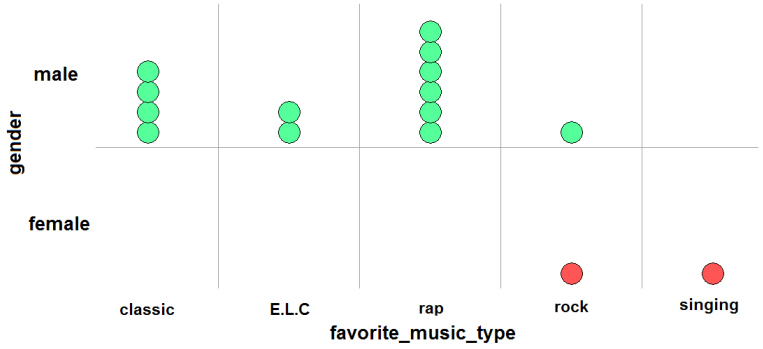


Figure 3.6. A random sample of size 15 taken from the second model (Figure 3.5).

They were surprised to observe that the sample (Figure 3.6) had 13 boys and two girls although they had built the model with an equal number of boys and girls (Figure 3.5). After drawing another sample, and again obtaining a similar result, they ran the sampler slowly and discovered that the sampler was drawing samples with replacement. Shay insisted on changing the sampling method to sampling without replacement. They then continued to draw more samples, each half the size of the

population, while increasing the number of cases in the model. Shay assumed that, "the bigger the population is—the smaller relatively is the sample size you have to take to learn about the population."

Session 2.3: Sampling Distribution (The Focused Episode). Exploring Samples from the Model. Shay and Liron built a third model with seven interconnected attributes according to their hypotheses. They entered 120 cases into the model (identical to the real population size) and examined whether a sample size of half the population would be similar to the model. They explained that in reality they intended to collect 40 to 60 cases (Liron) or 60 cases (Shay). They first drew a sample of size 120 without replacement (all 120 cases of the model, see Figure 3.7) to check if their model was compatible with their hypothesis. They then drew several random samples of size 70 from the model and explored the sample plot of favorite music type. To assess their confidence level in samples of size 70, they compared these sample plots to the population plot (Figure 3.7) to decide if they could learn from them about the hypothetical population.

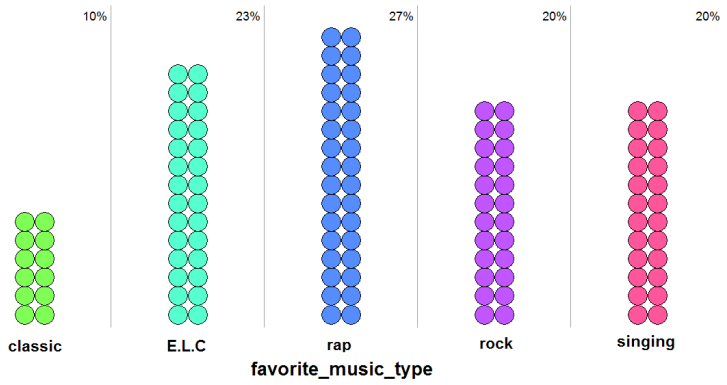


Figure 3.7. A random sample without replacement of size 120 taken from the third model.

Exploring the sampling distribution. At this point, the interviewer reminded Shay and Liron that there is a TinkerPlots™ option of simultaneous collection of data from many samples. They decided to collect 100 random samples and explore the sampling distribution of the statistic: Percentage of students whose favorite music is rock (see a detailed analysis of this exploration in Section 6.5.5 below).

3.5.4 Activities 3 and 4

To refine students' understanding of the connections between the two worlds, they were given a third activity. The students were asked to study a hidden TinkerPlots™ sampler with unknown data distributions (built by two other students) by exploring random samples drawn from this sampler. The idea behind this activity was to make a clearer distinction between data from the model and real data.

In the last activity Shay and Liron compared the real sample data of size 60 (collected by them) with their hypothesis. In response to the interviewer's suggestion to use the sampling distribution, Liron determined an interval of proportions he agreed to accept as "correct results." Based on both the real sample results and this interval, he decided to accept his hypothesis. As for the results which fell outside the interval, Liron said that, "if it's a little far out of the range, then that's fine, but it's not really a conclusion." Shay said that he had learned from the simulations that a sample size does not depend on a percentage of the population, "the bigger the population is—the more similar to the population a certain percentage of the population will be."

3.5.5 Main Results (Session 2.3)

We identified seven key stages (Table 3.2) in Shay and Liron's articulations of uncertainty. These stages present the ways they accounted for, quantified, controlled and decreased uncertainty while exploring the sampling distribution.

Table 3.2
Seven Stages in the Students' Articulations of Uncertainty

Stage	Stage Title
1	Accounting for uncertainty in sampling representativeness
2	Accounting for uncertainty due to sampling variability
3	Shay's discovery: Quantifying uncertainty
4	The students' views of uncertainty collide
5	Control of uncertainty: Better chance, but is it accurate enough?
6	Decrease of uncertainty by increasing sample size
7	Liron's and Shay's conclusions: Quasi-probabilistic vs. probabilistic view

Stage 1: Accounting for Uncertainty in Sampling Representativeness. Shay and Liron started this episode with a deep interest in examining whether a sample size of half the population would be similar to the whole population. They were curious to know the required sample size in order to make good conclusions about the population, a sample which they were actually going to collect. They did not succeed to resolve this issue by comparing a few repeated samples drawn from the model with the model, and therefore decided to explore a sampling distribution (Figure 3.8) of 100 random samples, each of size 70, of the statistic: the percent of students whose favorite music is rock (`%ROCK` in abbreviation). But, they first drew a plot of a sample of size 120 without replacement (i.e., all the 120 cases of the model; Figure 3.7), and saw that the `%ROCK` was 20% in their model.

They examined the representativeness level of samples size 70 by using two measures in the sampling distribution (Fig 3.8): a) The difference between the mean of

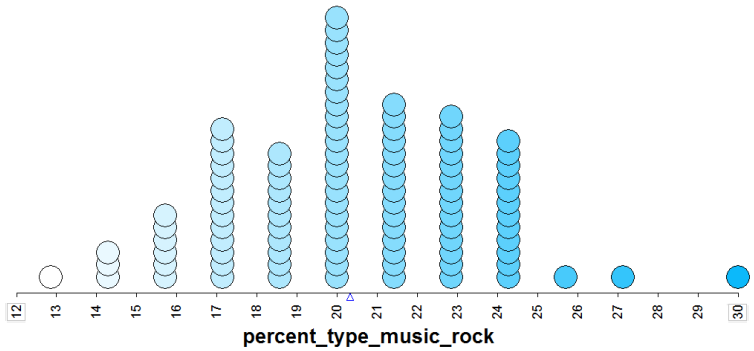


Figure 3.8. A `%ROCK` sampling distribution of 100 samples size 70.

the sampling distribution (the blue triangle) and the `%ROCK` in the model (20%); and b) the modal value (20%), and commented:

01	Shay:	Indeed the mean is close [to the original `%ROCK`], but this is obvious.
02	Liron:	You are like, yes, yes [cynically referring to what he thinks Shay is going to say]: "I'm going to prove that he [Liron] is wrong."
03	Interviewer[3]:	What do you see here [Figure 3.8] now? Can you describe it?
04	Liron:	That the mean is correct.
05	Shay:	The mean will obviously be correct.
06	Liron:	[Referring to interviewer's question in Line 03] That most of the tests [samples] showed that . . . the percentage of kids who like rock is 20, which is 20 percent of kids who like rock.
07	Shay:	But that is obvious, Liron.

Unlike Shay, who claimed that the results were self-evident, Liron expressed high confidence in the representativeness of samples size 70. He was satisfied that the sampling distribution's mean was close to the original `%ROCK` and that the mode was equal to it.

Stage 2: Accounting for Uncertainty due to Sampling Variability. The interviewer nudged students to further reason about the sampling distribution (Figure 3.8):

[3]The interviewer in all the quotes is the first co-author.

18	Interviewer:	. . . What can you learn from this [Figure 3.8]?
19	Shay:	I say that it is not accurate enough.
20	Liron:	. . . So each piece of data here, every time it . . . that we made a new mixing [took a new sample], once it was 27%, another time it was 30%, each time it was some other percent, but it was 20% the most.
26	Interviewer:	And if it was 19%, would that be right or not? Or 18 [percent]?
27	Liron:	18, no [would not be right].
28	Shay:	19 [percent] is still reasonable, but let's say 18, and 23, that would be going over the line.
29	Liron:	20 is, however, correct.
30	Shay:	And there are quite a few [results that are 23%].

Liron first clarified to himself the simulation process that led to the sampling distribution at hand, and concentrated on the signal of the sampling distribution. Unlike him, Shay was more attentive to the resultant "noise" [Line 19]. Trying to sway Liron's attention from the mode, the interviewer asked him about a range of results, but he remained focused on the equality between the mode and the original `%ROCK` and provided only deterministic utterances [Line 29] (in the sense of Ben-Zvi et al., 2012). Being aware of sampling variability, Shay accounted for the uncertainty involved in this process [Lines 28 and 30].

Stage 3: Shay's Discovery: Quantifying Uncertainty. In response to Shay's referral to reasonable results [Lines 28 and 30], the interviewer suggested to group the data into intervals instead of discrete categories. The students used dragging to create a sampling distribution presented as continuous style vertical bins (Figure 3.9), and added relative frequencies in percentages.

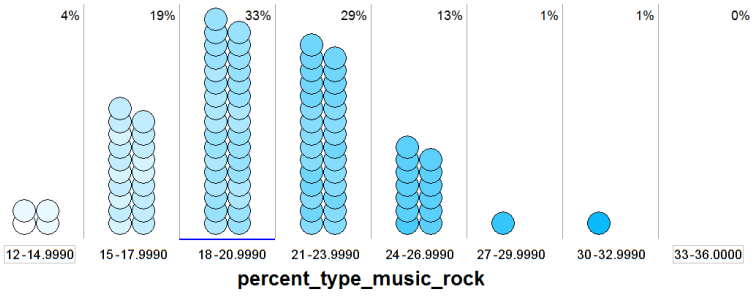


Figure 3.9. A continuous style vertical bins sampling distribution with percentages, sample size 70.

33	Interviewer:	Tell me what you see here [now in Figure 3.9]?
34	Liron:	But in fact I have proved what I have done!
35	Interviewer:	How have you proven it?
36	Liron:	33% is incorrect, then.
37	Interviewer:	No. What do you see here?
38	Shay:	Oh [with excitement]! Here the percentages are the probability that it will come out like this [that one sample will fall in this interval]!
39	Liron:	[uncertain] What?

The new representation (Figure 3.9) led Shay to find a way to quantify the uncertainty involved in the modeling and sampling process. He correctly interpreted the percentages of a certain bin as the probability that a sample result will fall in the interval covered by that bin [Line 38]. Unlike him, Liron was not pleased with this new representation because the modal bin included several values rather than just 20% (the value of the original `%ROCK`). He preferred the former representation (Figure 3.8) since the mode there was equal to `%ROCK`, and therefore changed the graph (Figure 3.9) back to the previous one (Figure 3.8).

Stage 4: The Students' Views of Uncertainty Collide. Liron changed manually (rather than by dragging) the interval width to one (Figure 3.10). They then used Shay's discovery from the previous stage to quantify the probability that a sample statistic will be equal to the original `%ROCK`:

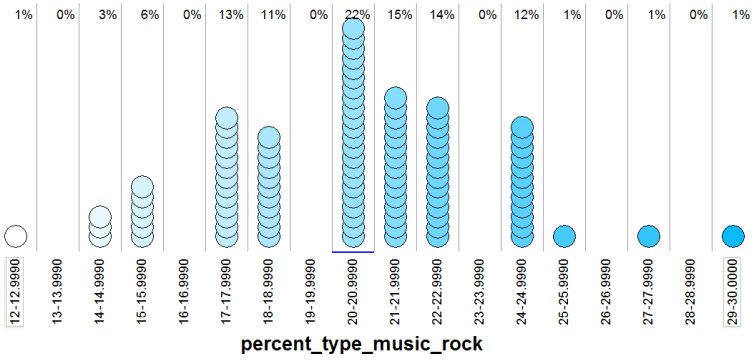


Figure 3.10. A sampling distribution with bin width of one, sample size 70.

51	Shay:	Pay attention [Figure 3.10]: The probability... that it [the sample statistic] will be correct is...
52	Liron:	22%!
53	Shay:	22%, and that's not a lot.

55	Liron:	Why not a lot? Everyone [most of the data] however is here [points at the center]... But this means that... that it happened the most times.
62	Shay:	If we take one hundred samples it [the mode] will obviously come out close.
63	Liron:	It happened most often, right? So, think about it for a second: If it happened most often, every time that we mixed [took a random sample], it most often produced data that is in these percentages.
68	Interviewer:	Right, but is this "most often" significant in relation to other [results]? 21 [percent] also appears a lot.
69	Shay:	No. No. [Not significant]. A probability of 22%!
70	Liron:	But this is the idea. That it [the 20–20.999 interval in Figure 3.10] is the largest. It occurred more times!
71	Shay:	Listen, Liron. A probability of 22% is ridiculous.

The students' interpretations of the probability of 22% exposed their different points of view on likelihood and confidence: Liron increased his confidence in samples of size 70 since the mode was equal to the original `%ROCK` and he could state the likelihood that samples of size 70 will be equal to the original `%ROCK`. His interpretation of the sampling distribution was deterministic and local by focusing only on the "correct" result without considering that a probability of 22% is relatively small. For Shay, who viewed this sampling distribution globally (in the sense of Ben-Zvi & Arcavi, 2001) and probabilistically, the 22% probability further decreased his confidence in random samples of size 70.

Stage 5: Control of Uncertainty: Better Chance, but is it Accurate Enough? Trying to move the students' focus from a single "correct" result to a range of results, the interviewer asked whether getting 22% in a random sample will be too far from the original `%ROCK`. This question led to the following discussions.

77	Shay:	But Liron, let's say that up to three should be enough...
78	Interviewer:	What is "up to three"?
79	Shay:	Of, an inaccuracy of up to three is good enough. Write down three [asking Liron to change the bin width to three, Figure 3.11].
81	Shay:	And the probability is one-third... If we are satisfied enough [compromise] with [a range of] 18 to 21 [rather than precisely 20].

The interviewer tried to understand why Shay said that the interval 18 to 21 had "an inaccuracy of up to three" [Line 79]. In response, Shay explained that it was

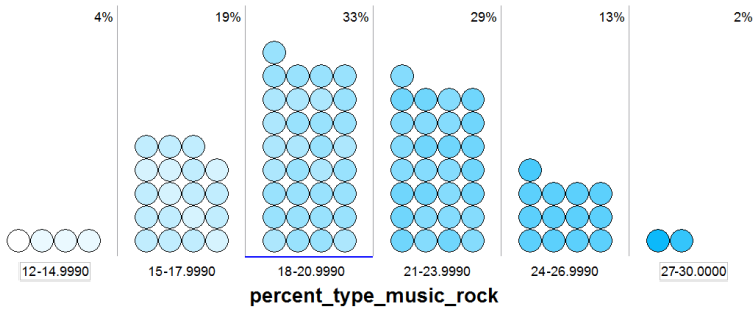


Figure 3.11. A sampling distribution with bin width of three, sample size 70.

close enough to the original %ROCK. The interviewer clarified that a deviation of three from 20% means plus or minus three, (i.e., 17–23%). Shay then created several graphs (Figures 3.12–3.14) trying to get a symmetrical range around the 20% by using the dragging option in TinkerPlots™ but did not succeed.

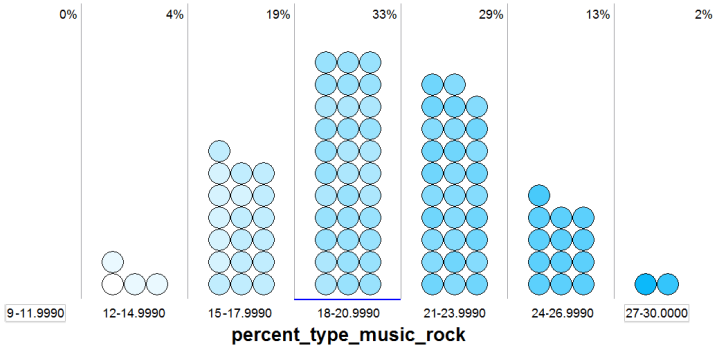


Figure 3.12. A sampling distribution with minimum value nine, sample size 70.

102 Shay: Come on. It is more or less. Everything is more or less a probability of one-third. It is less than fifty and I'm not satisfied with it [Figure 3.12].

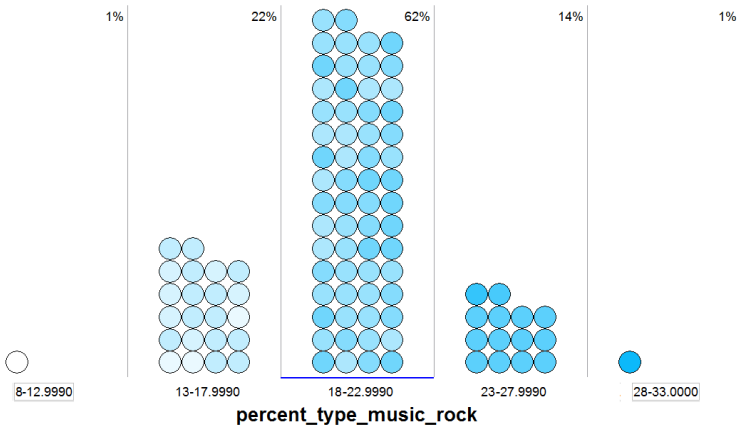


Figure 3.13. A sampling distribution with bin width of five, sample size 70.

109 Shay: 18 to 23. What do you think, Liron? It's a better chance [62%], but is it accurate enough? [Figure 3.13].

110 Liron: Well, let it go, you see? The highest result will always be between 18 and 23 [Figure 3.13].

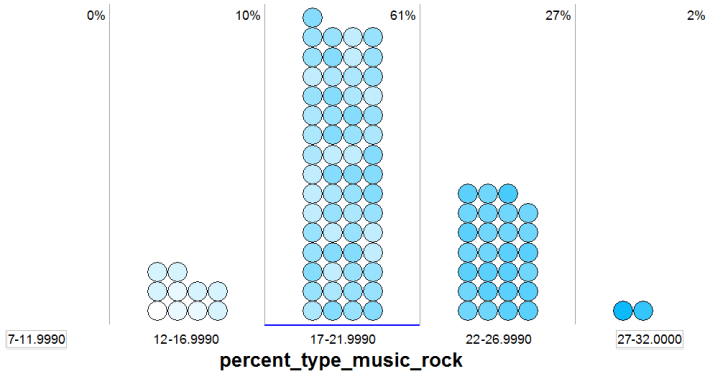


Figure 3.14. A sampling distribution with minimum value seven and maximum value 32, sample size 70.

115 Shay: It is obvious that the probability [of the center interval around 20%] will be the highest. The probability that it would turn out correct is higher than the one that it would turn out incorrect. The question is how much higher it is.

Liron strengthened his confidence level by observing that the center bin that included the value of the original %ROCK had large probability in all these graphs. Shay examined each graph differently (Figures 3.12–3.14): He observed the probabilities that a sample fell in different ranges around %ROCK in order to control the uncertainty and decrease it. The students' different perspectives made Shay articulate the key issue in examining this sampling distribution: What is the probability of the center bin that I accept as certain enough to use a sample size 70? [Line 115] When Shay eventually created another graph (Figure 3.15), the interviewer suggested to sum the percentages of the three center bins (20 ± 3) to get 75%.

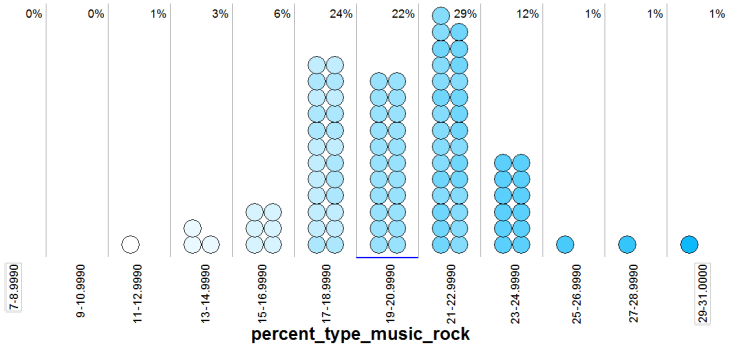


Figure 3.15. A sampling distribution with bin width two, sample size 70: The probability of the 17–23 range is 75%.

129	Shay:	Even if the chance is 75 [%]... [Figure 3.15]
130	Liron:	Wait, it [referring to the 19–20.999 bin] is the lowest here!
131	Shay:	There is a 25% chance that it will not turn out right, and I cannot take a risk of 25%.
132	Liron:	Stop! Then it is not correct, right? Because it is suddenly lower [than the frequencies of other intervals]. [Cursing], and it is 20.
133	Interviewer:	[Correcting Liron] These are all the numbers that are 19 to 21.
134	Liron:	That is not correct.
136	Shay:	Not to mention the fact that it is far... And if there is a deviation of three to four percentages, than it could be meaningful because it could bypass other data [type of music].

Examining the new sampling distribution (Figure 3.15) with a probabilistic view, Shay focused on the probability of getting "wrong" results and defined it as a risk he

refused to take. He was aware of the negative aspect of the risk [Line 131] and the compromise he had to make by accepting a large deviation from the original %ROCK [Line 136]. Liron, focusing mostly on the center bin height was upset to find that it was not the mode [Line 130], which he rejected. This caused an increase of his uncertainty. He struggled to find a solution to this problem, and the interviewer tried to move his focus from the absolute height of the bins to their relative frequency:

144	Interviewer:	Shay talks about other things. He computes the percentages here.
145	Shay:	[These percentages are] the probability that it would turn out accurate and accurate enough. It is not sufficient that it is not accurate enough, it is also a probability that is not sufficiently high.
146	Shay:	Let us decide on the deviation that we agree to accept.
147	Interviewer:	How much deviation can you accept, Liron, from twenty?
148	Liron:	Two at the maximum. I mean two deviations.
149	Shay:	I would compromise on three.

In trying to explain his reasoning, Shay refined his articulations of the relations between probability, accuracy and certainty [Lines 145–146]. He realized that for decreasing uncertainty, there was a need for two related conditions: (1) Small range of values of statistics around the original %ROCK; and (2) High probability to get a sample statistic in that range. At the end of this stage, Shay determined that a sample size of 70 was "unequivocal not enough" [Line 168] and therefore increased the sample size to 100 to decrease his uncertainty.

Stage 6: Decrease Uncertainty by Increase of Sample Size. The students drew 100 Samples and created a sampling distribution for sample size 100 (Figure 3.16).

188	Liron:	I was right. I was right [observing Figure 3.16].
189	Shay:	That still doesn't mean that you were right.
192	Liron:	49 [%]. It's almost 50 [%]. It's good enough.

Their different perspectives are clear now: Observing the new sampling distribution based on samples size 100 (Figure 16), Liron increased his certainty since 49% of the samples were almost equal to the original %ROCK. Shay was unhappy with this small probability and kept experimenting by changing the bin width to one (Figure 3.16).

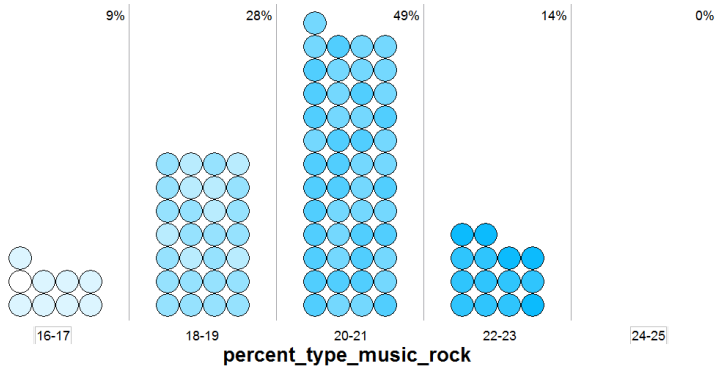


Figure 3.16. A sampling distribution of 100 random samples, size 100, bin width two.

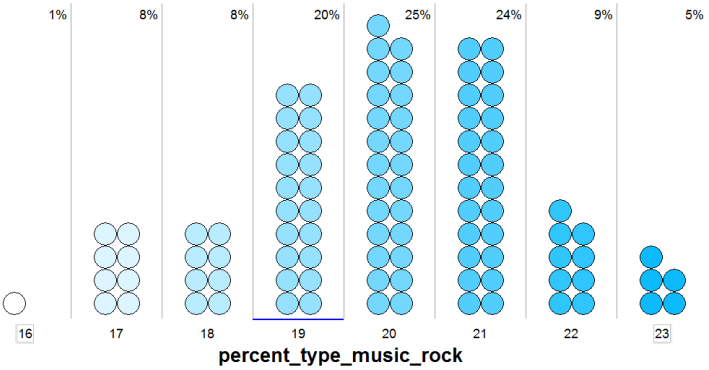


Figure 3.17. A sampling distribution with bin width one, sample size 100: The probability of the 18–22 range is 86%.

201	Shay:	We said that if it were in the range of two [deviations from 20%], then it would be accurate enough. So, I take all those that are in the range of two and add them to examine the chance that this will turn out accurate enough [Figure 3.17].
202	Liron:	Oh, we add from 18 to 22?
203	Shay:	Yes.

Shay explained why he changed the bin width in the graph to one [Line 201]: He wanted to find the probability to get a result with an inaccuracy of two deviations, which turned out to be 86%.

205	Interviewer:	Then what is this 86%?
206	Shay:	The chance.
207	Liron:	Which is much higher than 50% [49% in Figure 3.16, Line 192].
208	Interviewer:	The chance of what?
209	Shay:	It was 75%[4] previously [75% in Figure 3.15, Line 129].
210	Interviewer:	86% is the chance of what?
211	Shay:	That it will be accurate enough.
212	Liron:	As long as these 20% are the highest, I have proven what I have said.
213	Interviewer:	But you could receive a different sample. You could receive 23 or 21. How would you explain to me that it is more likely that I would receive 20? And very likely. 25% is not enough for me.

Shay and Liron responded differently to the interviewer's question about the meaning of the 86% in Figure 3.17. Shay noticed that the probability of the center bins increased from 75% to 86% as the sample size increased from 70 to 100 [Line 209]. This articulation is part of his ongoing and consistent effort to find a probability that would turn out accurate enough to make him more certain. Liron, for the first time in this episode, seems to refer to the relative frequency as a chance [Line 207], but immediately afterwards he returns to focus on the 20% as a mode [Line 212]. The interviewer tried to sway Liron from deterministic to probabilistic reasoning [Line 213]. As a result, and with the interviewer's mediation, Liron described a range of results he agreed to accept, exemplified how he would make a decision if he got a certain result of a real single sample, and accounted for chance in the sampling distribution.

228	Liron:	If we say that in a real sample, if we examine the children, and receive 19% children that like rock, then in my opinion it is all right. This is because it is amongst the highest chances.
229	Interviewer:	What does "all right" mean to you? What will you really believe?
230	Liron:	To 20 [the percentage of students whose favorite type of music is rock].

[4]Shay did not realize that on samples size 70, 75% was the probability of obtaining a statistic within a range of three deviations from the original `%ROCK` (17–23) and not of two deviations. On samples of size 100 he obtained a probability of 86% of getting a statistic within a range of two deviations from the original `%ROCK` (18–22).

Stage 7: Liron's and Shay's Conclusions: Quasi-Probabilistic vs. Probabilistic View. The interviewer asked each one of the students to summarize his conclusions. Liron explained his considerations about making decisions from a hypothetical single sample result.

252	Interviewer:	When will you say that your hypothesis was correct?
253	Liron:	I believe only this and this and this [point to 19–21 bins in Figure 3.17].
254	Interviewer:	And if 18% comes out, you wouldn't believe your hypothesis?
255	Liron:	I would sort of believe my hypothesis because it was close enough. As if, a range that, I, it is as if, uh . . . it is like saying maximum minimum. Okay?
257	Liron:	Minimum is 19 or 21. That means that my hypothesis is really correct. Not by a 100% but really correct.
258	Interviewer:	By what percent?
261	Liron:	I don't know. I have no idea. I am not so good at percentages. In my opinion, 18 or 21 is the maximum.
262	Interviewer:	That you would be ready to be wrong?
263	Liron:	Which means: My hypothesis is quite correct, but it was not really, it was not in the right direction. This is like 50%. I was half-right and half-wrong. A range of 17 or 23, I am not correct at all, I can already tell you that my hypothesis was wrong.

Liron explained how he would decide whether to accept or reject his hypothesis according to different real sample results while using phrases like "sort of believe" and "quite correct", which reflect uncertainty in his articulations. However, in his explanation, he refers only to the difference between the statistic's value and the original `%ROCK` and not to the probabilities or the frequencies of results based on the sampling distribution (like he started to do before). Therefore, we think that in the former stage he held a quasi-probabilistic view when he accounted for chance in the sampling distribution.

Shay explained his conclusions from the exploration of the sampling distributions.

266	Interviewer:	Shay, what do you think? When would you believe and what would you not believe?
267	Shay:	Okay, Let's say that I am ready to accept a deviation of two percent. This is a total of 86%, Okay? This leaves me with an error of 14%.
269	Shay:	14% [chance] that it would turn out incorrect. I refuse to take a 14% [error or risk].

270	Interviewer:	Is this too much?
271	Shay:	That wouldn't help! As hard as we may try, we must take all the data to be certain.

Shay described his probabilistic considerations about the ability to make conclusions from sample to population based on what he has learned from examining the two sampling distributions. Using rich and high level of expressions, Shay explained the reasons for his high uncertainty while taking into account accuracy and risk level and the connection between them.

3.5.6 Summary of Results

We summarize the results on Shay's and Liron's articulations of uncertainty separately (as shown also in Table 3.3) because most of the time they "walked" in parallel lines, with a few mutual influences.

Table 3.3
Summary of the Students' Articulations of Uncertainty

Stage	Stage Title	Shay's Position	Liron's Position
1	Accounting for uncertainty in sampling representativeness	Accounted for big uncertainty as before because sampling representativeness was obvious for him	Accounted for small uncertainty because of sampling representativeness
2	Accounting for uncertainty due to sampling variability	Accounted for big uncertainty resulting from sampling variability	Focused on the mode in the sampling distribution being equal to the original %ROCK, and therefore accounted for small uncertainty
3	Shay's discovery: Quantifying uncertainty	Quantified uncertainty using relative frequency of bins in sampling distribution	Found it difficult to focus on the signal in the sampling distribution

Table 3.3 – continued from previous page

Stage	Stage Title	Shay's Position	Liron's Position
4	The students' views of uncertainty collide	Observing likelihood of 22% in samples whose %ROCK was equal to the original %ROCK, decreased his confidence level about random samples of size 70 because it was only a likelihood of 22%	Observing likelihood of 22% in samples whose %ROCK was equal to the original %ROCK, increased his confidence level about random samples of size 70
5	Control of uncertainty: Better chance, but is it accurate enough?	Controlled uncertainty by determining different ranges around the original %ROCK and accounted for the probability that a sample will fall in that range; Quantified uncertainty by focusing on the probability to get "wrong" results and defined it as a risk	Increased his confidence level in samples of size 70 while observing graphs in which the center bin that included the value of the original %ROCK had large probability; Decreased his confidence level while observing graphs in which the center bin was not the mode
6	Decrease of uncertainty by increase of sample size	Increased his confidence level about samples of size 100 but his uncertainty remained too high because the probability to get "wrong" results was too big	Increased his confidence level about samples of size 70 because the likelihood for bins that included the value of the original %ROCK was bigger than 50%
7	Liron's and Shay's conclusions: Quasi-probabilistic vs. probabilistic view	Summarized his considerations about uncertainty taking a probabilistic view	Summarized his considerations about making decisions from a hypothetical single sample result referring only to the difference between the statistic's value and the original %ROCK

Shay's Articulation of Uncertainty: From Global to Probabilistic View. Shay demonstrated a global view of the sampling distribution in the first two stages. He was aware and certain of the signal in this distribution around 20%. But it was the noise he noticed in the distribution that caused him to feel uncertain about the ability to make conclusions based on random samples of size 70. Yet, he expressed his uncertainty only in general terms, for example that "it is not accurate enough." Only in the beginning of the third stage, motivated by his deep curiosity to understand uncertainty, he started viewing the sampling distribution probabilistically, which enabled him to quantify the uncertainty.

At the third and fourth stages, Shay interpreted the bin heights as probability statements about a sample statistic. At the end of Stage 5, he explained that he looked for the "probability that it [%ROCK in sample size 70] would turn out accurate and accurate enough." But he realized not only "that it is not accurate enough, it is also a probability that is not sufficiently high." Then he determined that a sample size of 70 was "unequivocal not enough" and increased the sample size to 100 to reduce uncertainty by getting smaller variability. For example, he realized that the probability of getting a sample statistic in the range of plus minus two deviations is 86% in sample size 100, larger than 75% in sample size 70. Although he did not remember that the probability of 75% on samples size 70 was calculated for getting a sample statistic in a larger range of plus minus three deviations, Shay was not surprised and even expected to get a larger probability on larger samples. Thus, it seems that he understood informally two key ideas regarding sampling distributions: (1) as the sample size gets larger, the variability of the sample means gets smaller; and (2) the bins' relative frequency represents the probability of the sample statistics (Garfield et al., 2005).

Furthermore, when Shay tried to control uncertainty in the fifth stage, he determined a range of statistic values with an "error" of three deviations from 20%, found the probability of obtaining a sample statistic outside of this range, and named it "risk." It seems that he thus described informally a measure of variability of sampling distribution, similar to the formal standard error of the mean. He also described the likelihood of different values of the sample %ROCK in order to quantify uncertainty. In order to control and decrease uncertainty, he described the probabilities that a sample will fall in different ranges around the original %ROCK. Therefore, we can claim that he was able to use his knowledge about sampling distributions to describe: a) the size of the standard error of the mean; and b) the likelihood of different values of the sample mean (Garfield et al., 2005). In some sense, Shay's "discovery" of how to control uncertainty by relating it to the probability of getting a certain statistical result can be viewed as a first step towards understanding the reasoning behind hypothesis testing.

Liron's Articulation of Uncertainty: From Deterministic to Quasi-Probabilistic View. Liron's articulations were characterized with a local view of uncertainty in the sampling distribution. He noticed from the beginning that most of the %ROCKs in the sampling distribution were equal to the original %ROCK and that the mean of the sampling distribution was very close to the original %ROCK. Focusing on these sig-

nals, Liron expressed a very high level of confidence most of the time and sometimes even an absolute certainty in samples of size 70. Liron's consideration of one of two possible conclusions (correct or incorrect) also demonstrates his deterministic view of uncertainty (Ben-Zvi et al., 2012). The shift in his view happened during the fifth and sixth stages: Following his discussions with Shay and observing Shay's actions and articulations, Liron widened his observations to an interval of results around the value of the original %ROCK. When the students began observing a sampling distribution of samples size 100, Liron referred to relative frequency in sampling distribution but still was focused on the mode. With the interviewer's mediation, he expressed a quasi-probabilistic view when he accounted for chance in the sampling distribution. But in the seventh stage, his decisions were based only on the values' difference from the original %ROCK and there was no reference again to probabilities or to frequencies.

3.6 Discussion and Implications

This chapter focuses on the question: How can students' articulations of uncertainty emerge while informally exploring sampling distributions using the integrated modeling approach? To address this question we analyzed Shay's and Liron's articulations of uncertainty in seven stages in which they explored sampling distributions in the model world in order to find the minimal sample size on which they could make ISIs in the data world. They struggled with the fundamental concept of statistical uncertainty in the process of making a statistical inference from a sample to population.

The study sheds light on how young students were able to engage with the complex idea of sampling distribution by encouraging them to articulate their uncertainty in the context of making ISIs. Even students with statistical knowledge about theoretical probability distributions find it difficult to make connections between theoretical models and empirical distributions (Noll & Shaughnessy, 2012). Both of the students understood that the exploration of the sampling distribution can help them decide on the minimal sample size needed to draw reliable conclusions about the population. Actually, their argumentation circled around the question of whether the sample size explored was large enough or not. That is, they began to connect between repeated samples that were drawn from a theoretical model and a single empirical sample that they were about to collect. This finding strengthens the argument that one needs to envision a process of repeated sampling to understand the logic behind ISI and the relationship between sample and population (Shaughnessy, 2007; Thompson et al., 2007).

We suggest that there was another important factor that helped the students to connect between repeated sampling and a single sample: the students' engagement with an authentic context (Edelson & Reiser, 2006) and in the data and model worlds. They explored sampling distributions that stemmed from an authentic and real motivation to study students' music preferences in their age group. The exploration of the sampling distribution came after they realized that they could not ask everyone,

but rather had to take a sample and decide on a sample size. Thus, their entrance to the model world and the resulting sampling distribution exploration was motivated by a real and an authentic goal.

We found two different views in the way the students observed and manipulated the sampling distributions: Shay moved from a global to probabilistic view and Liron from a local–deterministic to quasi-probabilistic view. These views shaped their articulations of uncertainty. Rubin, Hammerman, and Konold (2006) claimed that one needs to see the distribution as an aggregate to make conclusions from a distribution. In regards to the issue of sampling distribution, we suggest to broaden this claim: One needs to have an aggregate-probabilistic view of the sampling distribution to infer from a sampling distribution. Shay's awareness of the signal and noise of the sampling distribution increased his uncertainty in relation to samples of size 70, which motivated him to look for ways to control and quantify the uncertainty and motivated him to move to a probabilistic view. In Liron's case too, when he was encouraged by the interviewer or by Shay to consider a range or the frequency of results, he began to move from a local to global, quasi-probabilistic view.

Although Liron showed a deterministic view most of the time in this episode, he demonstrated in other activities of the *Connections Project* in grades 5 and 6 a probabilistic view in his articulations of distributions and fluently noticed signal and noise in data. We think that one reason for his local and deterministic view in this study was the type of statistic they explored. Exploring sampling distributions of possible percentages made it harder for him to conceive of the distribution as a whole. His recurrent descriptions of the simulation process and the meaning of the sampling distribution's data indicate that understanding the sampling distribution was not easy for him. It might explain partly his persistent focus on the signal. However, the glimpse of probabilistic views that Liron exposed may indicate an emergence of a change in his articulation and understanding of uncertainty. We suggest studying the conjecture that sampling distributions of means and many iterations between the data and model worlds may help students like Liron shift to a probabilistic view.

The IMA design and learning trajectory, which connects iteratively between the data and model worlds, seems to motivate students to consider, control and quantify uncertainties by exploring sampling distributions. In our case, they knew that after exploring the sampling distribution, they would have to make a decision about the sample size of the real data they will collect in order to make good conclusions about the population. Furthermore, before exploring sampling distributions, the students built the hypothetical model based on their context knowledge and were engaged in drawing and exploring samples from this model. We think that the process of building models and drawing samples from them contributed to their thinking about resampling and prepared them to the sampling distributions exploration. We suggest that the complex learning processes described above are strongly related to the IMA design, and take actions to experiment and study it further in different contexts and age levels.

We are well aware that Shay and Liron might not be representative of other students of a similar age since they benefited from their deep involvement in the *Connections Project* for two years before this study. Since we believe that the complex

issue of sampling distribution should be presented after some experience with EDA activities, we are currently conducting another study of three pairs of sixth grade students who had only partial involvement in the *Connections Project* (Manor & Ben-Zvi, in press) to test the idiosyncrasy of the case presented in this chapter.

As a result of this study, we have made some changes in the IMA learning trajectory in the way and timing which we present the sampling distribution (Manor, Ben-Zvi, & Aridor, 2014). We found it better to enable students to spend more time on exploring repeated samples and inventing methods to compare these random samples before the idea of sampling distributions is presented. No doubt, the innovative educational approach discussed in this study is exploratory and deserves significant additional design and research efforts in order to contribute to the growing body of research on promoting students' reasoning about uncertainty in the context of sampling distributions.

References

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, The Netherlands: CD-β Press, Center for Science and Mathematics Education.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning*, *2*(1 & 2), 127–155. doi: 10.1207/S15327833MTL0202_6

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, *45*, 35–65. doi: 10.1023/A:1013809201228

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—The International Journal on Mathematics Education*, *44*, 913–925. doi: 10.1007/s11858-012-0420-3

Ben-Zvi, D., Gil, E., & Apel, N. (2007). *What is hidden beyond the data? helping young students to reason and argue about some wider universe.* Paper presented at SRTL-5, University of Warwick, UK.

Biehler, R., Ben-Zvi, D., Bakker, A., & Maker, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643–690). New York: Springer.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, *1*(1). Retrieved from https://escholarship.org/uc/item/8sd2t4rr

Chinn, C. A., & Sherin, B. L. (2014). Microgenetic methods. In R. K. Sawyer (Ed.),

The Cambridge handbook of the learning sciences (2nd ed., pp. 171–190). New York: Cambridge University Press.

Edelson, D. C., & Reiser, B. J. (2006). Making authentic practices accessible to learners: Design challenges and strategies. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 335–354). New York: Cambridge University Press.

Franklin, C., & Garfield, J. (2006). The Guidelines for Assessment and Instruction in Statistics Education (GAISE) project: Developing statistics education guidelines for pre K–12 and college courses. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 345–375). Reston, VA: National Council of Teachers of Mathematics.

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47–78). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer.

Garfield, J., Chance, B., & Snell, J. L. (2000). Technology in college statistics courses. In D. Holton (Ed.), *The teaching and learning of mathematics at university level: An ICMI study* (pp. 357–370). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J., delMas, R., & Chance, B. (2005). *Tools for teaching and assessing statistical inference.* Retrieved from http://www.tc.umn.edu/~delma001/stat_tools/

Gil, E., & Ben-Zvi, D. (2011). Explanations and context in the emergence of students' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1–2), 87–108. doi: 10.1080/10986065.2011.538295

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics*, *34*(2), 69–81.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, *12*, 217–230. doi: 10.1007/s10758-007-9123-1

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1). Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1/

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., … Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning*, *13*(1-2), 68–86. doi: 10.1080/10986065.2011.538299

Konold, C., & Miller, C. (2011). *TinkerPlots™ 2.0 beta.* Amherst, MA:University of Massachusetts.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289. doi: 10.2307/749741

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage Publications.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173. doi: 10.1080/10986065.2011.538301

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 1–4. doi: 10.1080/10986065.2011.538291

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Manor, H., & Ben-Zvi, D. (in press). *Students' emergent articulations of models and modeling in making informal statistical inferences.* Paper will be presented at SRTL-9, Paderborn, Germany.

Manor, H., Ben-Zvi, D., & Aridor, K. (2014). Students' reasoning about uncertainty while making informal statistical inferences in an "integrated modeling approach". In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education: Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).*

Meira, L. (1998). Making sense of instructional devices: The emergence of transparency in mathematical activity. *Journal for Research in Mathematics Education*, *29*(2), 121–142. doi: 10.2307/749895

Moore, D. (2007). *The basic practice of statistics* (4th ed.). New York: W. H. Freeman and Company.

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education*, *43*(5), 509–556. doi: 10.5951/jresematheduc.43.5.0509

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, *7*(2), 83–106. Retrieved from http://www.stat.auckland.ac.nz/serj

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS) [CD-ROM].* Salvador, Bahia, Brazil.

Pfannkuch, M., Wild, C., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM—The International Journal on Mathematics Education*, *44*(7), 899–911. doi: 10.1007/s11858-012-0446-6

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education*, *31*(5), 602–625. doi: 10.2307/749889

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 3–4. Retrieved from http://www.stat.auckland.ac.nz/serj

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107–129. Retrieved from http://www.stat.auckland.ac.nz/serj

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics (icots-3)* (Vol. 1, pp. 314–319).

Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.ime.usp.br/~abe/ICOTS7/Proceedings/index.html

Saldanha, L., & McAllister, M. (2014). Using re-sampling and sampling variability in an applied context as a basis for making statistical inference with confidence. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education: Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).* Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L., & Thompson, P. (2002). Conception of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270. doi: 10.1023/A:1023692604014

Schoenfeld, A. H. (2007). Method. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69–107). Charlotte, NC: Information Age Publishing.

Shaughnessy, M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on the teaching and learning of mathematics* (Vol. 2, pp. 957–1009). Charlotte, NC: Information Age Publishing.

Thompson, P. W., Liu, Y., & Saldanha, L. A. (2007). Intracacies of statistical inference and teachers' understanding of them. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207–231). New York: Lawrence Erlbaum Associates.

CHAPTER 4

EXPERIMENT-TO-CAUSATION INFERENCE: UNDERSTANDING CAUSALITY IN A PROBABILISTIC SETTING

MAXINE PFANNKUCH[1], STEPHANIE BUDGETT[1], AND PIP ARNOLD[2]

[1]The University of Auckland, New Zealand
[2]Cognition Education Limited and The University of Auckland, New Zealand

Abstract

Research on students' understanding of *experiment-to-causation* inference is limited despite the randomized experiment being prevalent in high school and introductory statistics courses. Using design research we: Determined conceptual foundations, created a two-lesson learning trajectory incorporating dynamic visualization software for the randomization test, implemented the trajectory in large introductory statistics classes ($n \approx 450$) and a workplace class, and analyzed student data from pretests and posttests and interviews to ascertain their reasoning processes in order to inform future teaching and learning approaches. In this chapter we have mainly focused on six students to explore their reasoning processes as they moved from the observed data and randomization test to making an *experiment-to-causation* inference. Our findings suggested that the dynamic visualization software assisted students to recall and understand the processes underpinning the randomization test. Student inference argumentation, however, needed further development.

Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning.
By A. Zieffler and E. Fry (Eds.)

We identified twelve elements within uncertainty and causality where the reasoning of students needed to be developed in instruction to enable them to appreciate more fully the argumentation and concepts underpinning the designed experiment and the randomization test.

Keywords: Randomization test; Introductory statistics students; Randomized experiment; Dynamic visualizations; Causality and uncertainty; Inference argumentation

4.1 Overview

In this chapter we focus on the randomized experiment and understanding causality. Since causality is established within a probabilistic setting, we aim to explicate the notions within uncertainty underpinning *experiment-to-causation* inference using the randomization test. We discuss six main interconnected uncertainty ideas that underpin a two-lesson learning trajectory designed using the dynamic *Visual Inference Tools* (VIT: http://www.stat.auckland.ac.nz/~wild/VIT). We then explore ideas of uncertainty prevalent in students' reasoning processes as they progress from thinking about the observed data, recalling the randomization test with the VIT software, to making a claim about the data. We identify twelve notions of uncertainty that instruction may need to address when developing students' ideas in the realm of *experiment-to-causation* inference.

The study is part of a large project, which aimed to understand how to introduce school- and tertiary-level students to inferential ideas using bootstrapping and randomization methods. The research reported in this chapter focuses on the pre- and post-instruction tests and interviews of six introductory university and workplace statistics students. Occasionally we refer to the test responses of the other students ($n \approx 800$). The study was conducted within the classroom setting for the university students ($\text{class sizes} \approx 450$) and a professional development workshop setting for the workplace students ($n \approx 20$).

4.2 Problem

Research on statistical inference has largely focused on *sample-to-population* inference and students' understanding of significance testing including the p-value. Apart from the work of Madden (2008a, 2008b, 2011) there seems to be little research that focuses on *experiment-to-causation* inference. With experimental design and causal inference included in the introductory statistics curricula at both the secondary and tertiary levels (e.g., College Board, 2010; Common Core State Standards Initiative, 2010; Franklin et al., 2007; Ministry of Education, 2007), there is a need to explore students' reasoning about causality. Frameworks for understanding students' reasoning, conceptualizations, and misconceptions together with researched learning trajectories need to be further developed to inform the teaching of causal inference. Hence, it is useful to study *experiment-to-causation* inference in order to understand

the reasoning processes that students use regarding causality and uncertainty when learning the randomization test. Using that knowledge we should be able to construct better learning trajectories. Although we acknowledge that the randomization test is a formal inferential method, our approach could be classified as partial informal inference, as students are not introduced to formal ideas of the null hypothesis, p-values and significance.

4.3 Literature and Background

For some time, statisticians (e.g., Pearl, 1996) and educators (e.g., Wild & Pfannkuch, 1999) have questioned why statistics has neglected causality. Wild and Pfannkuch (1999, p. 238) suggest that the looking for causation should be at the forefront in education:

> Statistics education should really be telling students something every scientist knows, "The quest for causes is the most important game in town." It should be saying "Here is how statistics helps you in that quest. Here are several strategies and some pitfalls to beware of along the way…"

Since the search for causes is of fundamental importance, they believe that a goal of the introductory statistics curriculum should be to move students from association to causation, and that there is a need to provide accessible material to teachers to meet this goal. They point out that correlation, the objective measure of linking one variable to another, along with the mantra, "correlation does not imply causation" has dominated statistics and statistics education. Pearl (1997) believes the field of statistics has not addressed causal inference, apart from the randomized experiment, because the language of statistics is ensconced in the language of probability. For the field to move forward in the area of causality he has invented mathematical constructs for thinking about causal pathways in observational studies. Rubin (2004) has proposed a similar, albeit different, framework. Both Pearl and Rubin have opined that introductory statistics courses need to better address statistical inference and causality, especially related to observational studies. In fact, Pearl has instigated an award, the American Statistical Association's *Causality in Statistics Education Prize*, to encourage the teaching of causal inference in introductory statistics.

Currently, within conventional statistics courses, testing for a causal relationship is limited to Fisher's randomized experiment, where there is an intervention and random assignment of units into groups (e.g., treatment and control). The randomized experiment, to date, has been the primary path to causal inference in statistics. Fisher's insight, which enabled causal inference, was replacing the link between the explanatory and response variables with a random coin toss, that is, random re-assignment (Pearl, 1996). In this situation, probability modeling can be used to determine whether the treatment is effective. This juxtaposition of uncertainty and causal inference within the context of the randomization test may be problematic for students when first encountered.

4.3.1 Uncertainty, Modeling, and Technology

Uncertainty in statistical inference lies at the nexus of statistical and probabilistic reasoning. Formally, in statistical inference, uncertainty is embodied in concepts such as confidence intervals and significance testing, the understanding of which is based on big ideas such as random behavior, independence, variation, distribution, the Law of Large Numbers and sampling distributions (Gal, 2005; Konold & Kazak, 2008; Pratt, 2005). In this formality, *statistical inference* is the process through which uncertainty is quantified. The broader process of inference, however, combines this quantification with two other unquantifiable types of uncertainty, *data quality* and *data validity*.

Data quality is the uncertainty related to the quality of the design of the study, non-sampling errors, and the measures, data, and information gathered that are used in making inferences; the unquantifiable sources of variation, which researchers attempt to minimize in the conduct of a study, that give researchers reason to hesitate in making claims. *Data validity* is the uncertainty about whether the right data were collected, whether the right questions were asked of the data, whether confounding variables explain the findings, whether the process that generated the data has changed over time and hence applications of any findings are no longer valid, and whether the researchers' mental model of the world matches reality. This type of uncertainty also includes doubt from the knowledge that findings are based on current knowledge and that findings can be overturned in the future in the face of new evidence (e.g., Scarf, Imuta, Colombo, & Hayne, 2012) leading to the realization that all knowledge is uncertain, which can lead to skepticism about any evidence. In effect, when students make an inference or claim, they need to consider or weigh the evidence on these three types of uncertainty. A major question is how to untangle these three types of uncertainty when developing students' reasoning about making judgments from data with respect to *experiment-to-causation* inference.

Researchers have examined how people make judgments under uncertainty (Kahneman, 2011), at what age students understand the construct of uncertainty (Langrall & Mooney, 2005), and how young students articulate uncertainty (Ben-Zvi, Aridor, Makar, & Bakker, 2012). They have been surprised at the deep-rooted cultural bias towards deterministic thinking, which seems to interfere with developing students' ability to reason probabilistically (Fischbein, 1975). To conceptualize the world non-deterministically requires long-term experiences and reflection upon probabilistic situations including an emphasis on modeling random behavior (Garfield, delMas, & Zieffler, 2012; Greer & Mukhopadhyay, 2005). The purpose of such modeling is to mimic random behavior in a real world system in an effort to understand the behavior of the real-world system, to answer questions about that system, and to predict future outcomes in the real-world system (Pfannkuch & Ziedins, 2013).

Modeling random behavior underpins the quantification of uncertainty using formal methods for statistical inference (e.g., confidence intervals and significance testing). We believe that introductory statistics students should be introduced to the quantification of uncertainty via bootstrapping and randomization methods rather

than through the conventional parametric approaches for inference that rely on a mathematical formalization (e.g., t-test, ANOVA). In line with Cobb (2007), we think the logic of inference, and the "big ideas" and concepts underpinning inference are more transparent to students using these methods, and are transferable to a wider range of situations. Also, these methods are becoming more prevalent in statistical practice (Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2009). Moreover, the research of Madden (2008a, 2008b, 2011), Garfield et al. (2012) and Tintle, Topliff, Vanderstoep, Holmes, and Swanson (2012) point to positive outcomes in students' statistical inferential reasoning when using randomization methods to teach inference for probabilistic situations.

The bootstrap and randomization methods can also be mediated through visual representations, which allow some concepts to become more accessible to students. Technology helps students link multiple representations—visual, symbolic, and numeric—and enhances their understanding through promotion of a visualization approach to learning (Sacristan et al., 2010). Dynamic software can allow students to analyze directly the behavior of a phenomenon, to visualize statistical processes in ways that were not previously possible, such as viewing a process as it develops rather than analyzing it from the end result. Such representational infrastructure allows access to statistical concepts previously considered too advanced for students. As Wood (2005, p. 9) states, simulation approaches "offer the promise of liberating statistics from the shackles of the symbolic arguments that many people find so difficult."

4.3.2 Our Approach to Experiments and Inference

The experiments we refer to henceforth are comparative experiments that have both an intervention and random allocation to groups. The random allocation is performed in an attempt to make the group comparisons "fair"; a design that can facilitate causal inferences about the effects of an intervention. To assist introductory statistics students in making a direct conceptual connection, we adopted as a basic principle that the "inferential method should mirror the process of data production" (Wild, Pfannkuch, Regan, & Parsonage, 2013, p. 9). That is, the data is produced by random allocation to treatment groups and therefore the inference method should be based on random re-allocation to treatment groups. As (Teague, 2006, p. 169) stated:

> The experimenter must always pay careful attention to the design of an experiment, since the method of analysis is determined by the manner in which the experimental units are randomized to treatments. The way you randomize is the way you analyze.

To enable students to make an experiment-to-causation inference we expect them undertake three actions: (1) thinking about the data obtained from an experiment; (2) conducting the randomization test by modeling random behavior; and (3) making a claim about the data. All these actions involve drawing on the underpinning ideas about uncertainty in making inferences.

4.3.3 Theoretical Framework: Six Interconnected Underpinning Ideas

Since inference lies between statistical and probabilistic reasoning, we draw on the work of Konold and Kazak (2008) who established four main ideas that are at the heart of connecting data and chance activities: model fit, distribution, signal–noise, and the Law of Large Numbers. Building on their work, which was for *sample-to-population* inference, we have redefined and interpreted these four ideas for *experiment-to-causation* inference. To this end, we have included two additional ideas, inference argumentation and principles of experimental design and causation, and modified Konold and Kazak's idea of the Law of Large Numbers to include random process and independence. Altogether, we have identified six inter-connected main ideas related to uncertainty that seem to underlie the ways of thinking about experimental data when attempting to make an inference (see Table 4.1). Below we briefly describe these six main ideas.

Table 4.1
Framework of Underpinning Ideas for Three Actions when Thinking about Uncertainty in Experiment-to-Causation Inference

Action 1: Thinking about Observed Data	**Action 2: Modeling Random Behavior**	**Action 3: Making a Claim about the Data**
Explanations for observed difference • Model Fit • Signal–Noise • Principles of Experimental Design and Causation	Testing observed difference against chance alone • Model Fit • Distribution • Signal–Noise • Law of Large Numbers, Random Process, and Independence	Argument for observed difference • Signal–Noise • Inference Argumentation • Principles of Experimental Design and Causation

The VIT module we use for the randomization test has one dynamically linked vertical screen (see Figure 4.1). The module shows the original data in the top plot. The middle plot represents the possible differences in centers when randomly reallocating under chance alone. The re-randomization distribution[1], which is then used to find the likelihood of observed difference or greater under chance alone, is dynamically built and displayed in the bottom plot.

Model Fit. When examining plots of observed data from a comparative experiment, researchers typically have prior contextual expectations about the direction of the difference (cf. Arnold, 2013). Since an experiment is often conducted based on prior research, there would be an expectation that the treatment would show some effect.

[1]We deliberately used this language to convey the underlying idea.

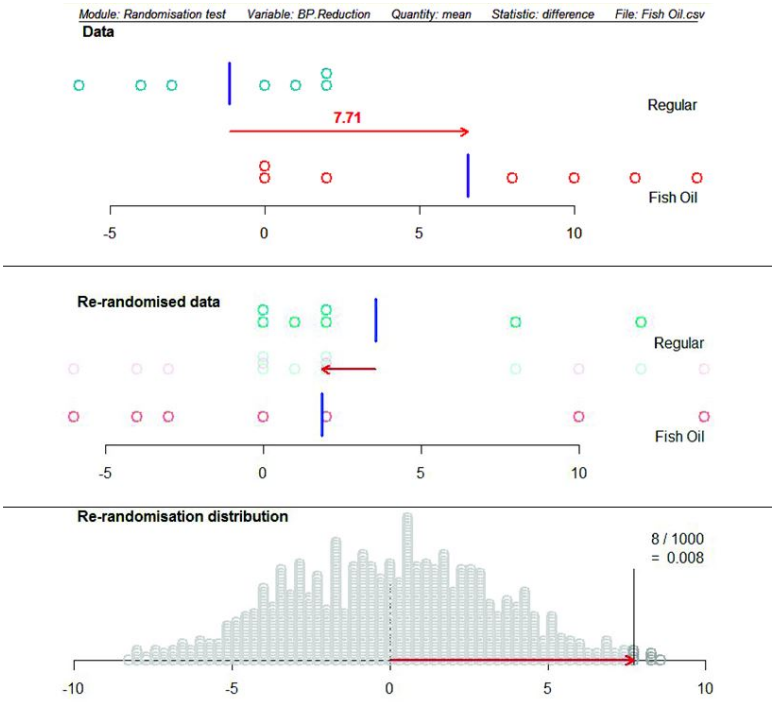


Figure 4.1. Screenshots of three dynamically linked plots within one vertical screen for VIT randomization test.

Students also bring their own general knowledge to the observed data and hence biological mechanisms inherent in the treatment, explanations about one group of experimental units compared to the other, and confounding variables can be proffered for any observed difference (Pfannkuch, Wild, & Regan, 2013). If the distributions observed are not as expected then further investigation is warranted including how the experiment was conducted. Therefore, features of the observed data are evaluated against students' "models" (e.g., expectation or conjecture; Konold & Kazak, 2008).

Similarly, when building the re-randomization distribution for the observed difference in centers, students have, or develop, expectations about features of this distribution. They typically expect to see a unimodal, symmetric distribution that is centered at zero. Another expectation that we have of the students is that they will draw on their experience of the *randomization variation module*, which they were shown during instruction. The module demonstrates chance acting alone where there is no treatment, only random assignment of units to two groups. Hence the probability model that is created in the randomization test can also be evaluated against the students' "models" (expectation of distributional features and expectation of the possible range of chance variation alone).

Signal–Noise. For experiments, we see the ideas of signal–noise showing up at three distinct points in the inferential process. The first point is during the examination of the observed data, which includes noise—chance effects of just who happened to be assigned to each group—and a potential signal—if the effect of the treatment is present. That is, chance acting alone may be the cause for observed differences (complete noise) or the differences may be due to both chance and treatment acting (noise and signal). Students must consider both of these explanations when thinking about why there is an observed difference. The second point at which signal–noise needs to be considered is during examination of the re-randomization distribution. At this point, students need to be able to reason about whether a treatment effect (signal) would be detectable under the obscuring effects of chance variation (noise) (Pfannkuch et al., 2013). Lastly, students need to consider signal–noise when they are interpreting the tail proportion to determine (1) whether the treatment is effective—a composite of signal and noise or (2) that they do not know whether the treatment is effective because the observed difference could be due to only noise or could be due to noise and signal—in which case there is a failure to detect the signal within the noise.

Principles of Experimental Design and Causation. For students to understand and interpret the observed results from a randomized comparative experiment, they need a good understanding of the principles behind the design. Apart from understanding the rationale for the use of a control group, blinding, and placebo, they also need to be aware that the researcher controls one or more of the explanatory variables through the use of random assignment of the experimental units to the two groups. In particular students need to realize that random assignment is the method for making the comparison fair with respect to: (1) eliminating bias that may result if the researcher assigns the units; (2) balancing the groups on variables that are known to affect the response; and (3) balancing the groups on confounding variables that may be unknown to the researchers (Agresti & Franklin, 2007). Another key idea behind randomized comparative experiments is that causality can "be established if the values of the explanatory variable are randomly assigned to the units" (Lock, Lock, Lock, Lock, & Lock, 2013, p. 36). Another facet of experiments is that volunteers are often used which means being cautious about or not generalizing results to a broader population.

Law of Large Numbers, Random Process and Independence. Many researchers have noted that learning to reason probabilistically includes developing notions of a repeatable process, random behavior, the Law of Large Numbers and independence. Yet, linking the idea of randomness and independence can be challenging for students (Watson, 2005). In the VIT dynamic creation of the re-randomization distribution, which is a probability model, students can visually see:

- The independence of each trial as the original data with the group label removed is brought together and then randomly re-allocated to the two groups;
- Random behavior in operation as the red arrows, the re-randomized differences in centers, constantly change; and

- The Law of Large Numbers in operation as the re-randomization distribution builds up and stabilizes as the random re-allocation process is repeated 1000 times.

Although we do not formally refer to the Law of Large Numbers, assumptions of a random process, and independence in the learning trajectory used, students can still potentially experience these concepts in this visualization of the process.

Distribution. A distribution is formed from a collection of individual data values into a conceptual entity that has its own characteristics and properties such as center, shape, variation, and density, which are inextricably linked to the context of the situation (Arnold & Pfannkuch, 2012; Bakker & Gravemeijer, 2004). Part of interpreting distributions is the visual decoding of the display such as the units and variables on the axes and what each data value represents (Friel, Curcio, & Bright, 2001). When confronted with the observed data from an experiment, students need to visually decode the display along with understanding the background of the study. They also need to make comparisons of the distributional properties of the two groups rather than individual cases. This can be especially difficult for novices (e.g., Konold, Higgins, Russell, & Khalil, 2004). However, there is a shuttling between observing distributions as an entity and observing individual data values. For example, the notion of *tendency* can be observed through comparing the property of distributional shift. The reasoning is centered on the idea that the group taken as a whole may benefit from the treatment but it may not be the case for every unit.

Similarly, when students observe the dynamic creation of the re-randomization distribution, or a static display, they need to visually decode the variables on the axes and what each data value represents. For the re-randomization distribution, or probability model, the property of interest is the tail proportion, which is visually represented, and therefore students need to understand what this represents in terms of all the other "data" (statistics) generated. From the perspective of Liu and Thompson (2007), such a dynamic visual display should be orienting students towards a stochastic conception of probability, since they are being provided with an image of possible outcomes.

Inference Argumentation. The nature of inference argumentation is based on understanding the logic of an indirect argument that uses probabilistic statements. This is often difficult for students to grasp, and hence, confusion about the argument can result in misinterpretation of the p-values (Nickerson, 2004). Many researchers (e.g., Falk & Greenbaum, 1995; Nickerson, 2004) have documented misconceptions related to the interpretation of the p-value and the consequent inferences made. Two examples of these misconceptions are:

- Accepting the null hypothesis if the p-value is considered to be large
- Considering a p-value as the probability that the null hypothesis is true given the data, rather than the probability of the data assuming that the null hypothesis is true.

In our teaching approach we purposefully have the students use language that we hope will partially convey particular concepts. For p-value, we use "tail proportion", for the null hypothesis we use "chance is acting alone". However, the dynamic visualizations that we use only show the tail proportion, so we expect inference argumentation to remain difficult for students.

In addition to reasoning from the tail proportion, students also have to consider the aforementioned ideas related to causation. Causal, or deterministic, thinking is the predominant mode of thinking within society, with most people not willing to accept the role of chance. Biehler (2011) used an inference example cited in Makar, Bakker, and Ben-Zvi (2011) to point out that the use of probabilistic thinking may lead to other issues. The example, which compared the physical fitness for fifth and sixth graders used the following language: "From these two samples, we infer that the physical fitness in sixth grade is probably better than in seventh grade?" (p. 152). Biehler (p. 6) observed: "'Probably' better expresses uncertainty. However what have we exactly gained? All our knowledge is uncertain. We can add this to every sentence we say." Hence, if people have the point of view that all knowledge is uncertain then they may be unwilling to use causal language, even to express inferences from experiments.

4.3.4 Research Questions

Since there seems to be little research in the area of *experiment-to-causation* inference with regard to conceptions of uncertainty, and since the VIT software is new and untested with respect to students' reasoning processes, we believe an exploration into students' concepts of uncertainty when using inference may contribute to the existing knowledge base. To examine students' reasoning processes regarding causality and uncertainty in the context of making partially informal *experiment-to-causation* inferences, we will focus on the following specific research questions:

1. What reasoning processes do students use when thinking about the observed data from an experiment (Action 1)?
2. What ideas and reasoning processes do students use when recalling the randomization test (Action 2)?
3. What argumentations do students use when making a claim about data from an experiment (Action 3)?

4.4 Subjects and Methods

The findings presented in this chapter come from a collaborative research project involving 33 team members and over 2700 students. The research team was comprised of a statistical software conceptual developer, an international advisor, two education researchers, two resource developers, five professional development facilitators, eight university lecturers, and 14 secondary school teachers. Using prin-

ciples of design research (Hjalmarson & Lesh, 2008), the development process involved two research cycles, each consisting of four phases: (1) from an identified problematic situation, understanding and defining the conceptual foundations of inference; (2) development of new resource materials and dynamic visualization software (VIT); (3) implementation with Year 13, introductory university, and workplace statistics students; and (4) retrospective analysis followed by modification and supplementation of resource materials. The focus of design research is to support and engineer new types of reasoning and thinking in response to problematic situations. As well as being pragmatic through producing an educational product that can be used by teachers, design research can also lead to new educational theories and areas of research (Bakker, 2004).

4.4.1 Participants and Procedure

The research reported in this chapter focuses on the pre- and post-instruction written responses and interviews of six introductory university and workplace students. Seventeen university students (randomly sampled from 200 volunteers in a population of $n = 2553$) and nine workplace volunteers (sampled from $n = 14$) participated in the interview process. Eleven of these 26 students were randomly allocated the randomization posttest in class (others completed a bootstrapping posttest). We chose to concentrate on the responses of six of these students (S1 to S6) because they were interviewed by the same research assistant, and were able to articulate their ideas. These six students' prior experience of statistics would be fairly representative of about 60% of the university and workplace cohorts. Occasionally we refer to the written responses of the wider cohorts to give an indication of the prevalence of the reasoning under discussion.

None of the participants had any experience with *experiment-to-causation* inference or the randomization test. All students experienced the same learning trajectory of two 50-minute lectures for the randomization test, which incorporated hands-on activities, attention to language and verbalizations, and VIT dynamic visualizations. Learning occurred within the classroom setting for the university students ($\text{class sizes} \approx 450$) and a professional development workshop setting for the workplace students ($n \approx 20$). An assignment component allowed students to use the VIT software to perform the randomization test that was demonstrated as part of the teaching sessions. For a detailed description of the teaching sequence, see Budgett, Pfannkuch, Regan, and Wild (2013).

4.4.2 Assessment Items and Data Analysis

Test and interview items that are discussed in this chapter are provided in the Appendix. Data from the tests were entered into spreadsheets. The first two authors of this chapter initially developed either hierarchical or non-hierarchical descriptors and coding frameworks for each assessment item based on the student data. The decision as to whether a hierarchical or non-hierarchical descriptor was necessary depended on the type of assessment item. At least 200 student responses were independently

coded until the descriptors and coding frameworks no longer needed to be modified and a consensus was reached on interpretation of the descriptors and the rating of student responses was synchronized. Two research assistants then independently coded the remainder of the data. For the interviews, which were audio- and video-taped and transcribed, familiarization with the data was initially conducted in cognizance of the findings and frameworks developed for the written responses. Based on this qualitative analysis of the interviews, a framework of the elements for describing the reasoning ideas within each of the three actions was developed (see Table 4.2). We then interpreted and discussed the interview data until a consensus was reached on interpreting how the students were thinking, identifying the gaps in the students' thinking, and the main elements emerging from the data that should be incorporated into the framework.

4.5 Analysis and Results

From the theoretical framework (see Table 4.1) and the student data we empirically developed a framework (see Table 4.2) for elements within uncertainty that are operationalized for *experiment-to-causation* inference. These notions are ones that we have identified in an introductory environment and are not an exhaustive list. They are simply an initial attempt towards a framework that could be used for understanding students' reasoning about causality within a probabilistic setting. A description of the reasoning and ideas within each element will be given, discussed and illustrated with student examples to demonstrate issues that students need to consider and grapple with as they learn about causal inference.

Table 4.2
Framework of Elements within Uncertainty Activated for Experiment-to-Causation Inference for Each of the Three Actions

Action 1: Thinking about Observed Data	**Action 2: Modeling Random Behavior**	**Action 3: Making a Claim about the Data**
Explanations for observed difference • Treatment is effective (treatment acting alongside chance) • Chance is acting alone • Experiment design and other issues	Testing observed difference against chance alone • Purpose of test • Simulating random re-allocation, uncertainty • Measuring uncertainty • Distribution of possible measures under uncertainty • Quantification of uncertainty	Argument for observed difference • Interpretation of tail proportion • Rare occurrence • Causal evidence • Tendency • Generalization • Experiment design and other issues

4.5.1 Action 1: Thinking about the Observed Data

For Action 1, we consider the pretest and posttest responses of S1, S2, S3, S4, and S5 (see Table 4.3) along with the responses from the wider cohorts to Question 1 (see Appendix). Briefly, students were given an experimental scenario, accompanying plots, and summary statistics. Then they were asked for two explanations for the observed difference. The explanations that we were seeking were: (1) the treatment is effective, and (2) chance is acting alone.

Table 4.3
Selected Students' Pretest and Posttest Responses to Question 1 (see Appendix) which Asked for Two Main Possible Explanations for the Observed Difference

Student	Pretest Responses	Posttest Responses
S1	(1) Fish oil replacing regular oil really reduces blood pressure. (2) Samples are too small and the observed difference is a result of biased sampling or big sampling error.	(1) The observed difference happened by chance. (2) Consumption of fish oil reduces blood pressure.
S2	(1) A fish oil diet does reduce blood pressure more than a regular diet. (2) Blood pressure is reduced when participants know they are on a fish oil diet.	(1) Chance is acting alone. (2) Something other than chance is acting. This may be the manipulation of the treatment type.
S3	(1) The fish oil diet is effective in lowering blood pressure. (2) External factors are influencing the results due to a small sample size.	(1) The fish oil diet is effective in reducing blood pressure. (2) Chance.
S4	(1) Those in the fish oil group had higher blood pressure than those in the regular oil group to begin with. (2) Those in the fish oil group had a greater range of blood pressure reductions than those in the regular oil group.	(1) The observed difference is not due to the effectiveness of the fish oil diet (i.e., it is chance that caused this difference). (2) The observed difference is due to a combination of chance, as well as the fish oil diet, to some extent.
S5	(1) That the fish oil is lowering the amount of mercury and hence blood pressure. (2) That regular oil is raising mercury levels and blood pressure.	(1) That the fish oil is lowering mercury levels in the blood and therefore lowering blood pressure. *not provided*

Of the 1,886 students who responded to this question in the pretest, 50% were able to state that one explanation for the observed difference between the two groups was that the fish oil was effective in reducing blood pressure, with written statements such as those shown by S1 to S3 in Table 4.3. This increased to 69% ($n = 868$) in the posttest. Thus many students responding noted the fact that the study was specifically designed to investigate the effectiveness of fish oil when compared with regular oil. However, in their search for explanations for the observed difference, some students reasoned beyond the experiment, using their contextual knowledge.

This was evident in both the pretest and posttest responses provided by S5 (see Table 4.3) and responses from the wider cohort such as:

> *That the Omega 3 contained in fish oil has a positive effect in reducing blood pressure values.*
> *Fish oil group may contain vitamins and minerals needed to reduce blood pressure.*
> *Fish oil helps lower cholesterol more than a regular oil diet would.*

While these statements may be true, the experiment was not designed to test the effectiveness of cholesterol, Omega 3, Mercury content, vitamins or minerals. These students appear to have proceeded to the next stage and, rather than thinking *if* the treatment is effective, they are wondering *why* it is effective. Such inferences are beyond the scope of the experiment.

Of the 1,487 students who provided a second explanation in the pretest, very few (5%) mentioned chance. An analysis of responses indicated that students were searching for other reasons for the observed difference. When asked what he meant by "external factors influencing the results," S3 stated, "they might have gone and done a whole lot of exercise." Other suggestions included the fish oil group being fitter, having a different lifestyle or, as suggested by S4, having higher blood pressure than the regular oil group to begin with. These confounding variables may indeed be explanations for the observed difference between the two groups. However, given the experimental design of the study and the fact that participants were randomly allocated to treatment groups, if those on the fish oil diet happened to be fitter, happened to have a different lifestyle, or happened have higher blood pressure to begin with, then these are what we would classify as *chance explanations*. Some students, including S2, had their own beliefs about blood pressure: "Blood pressure does reduce over time naturally and so the fish oil is actually not doing anything. . . the regular oil is causing people's blood pressure to remain constant." Such a belief suggests contextual knowledge information or misinformation is used to explain observed differences. S2 also suggested that blood pressure would reduce more in the fish oil group since they knew they were being treated: "I made the assumption that people who were on either diets knew that they were on the fish oil diet or the regular oil diet so they had that knowledge," which meant he was considering the placebo effect. Information about design issues such as double blinding, and whether the fish oil and regular oil diets were given as tablets would normally be given to students but in a time-restricted test this was not possible resulting in some students focusing on these issues as explanations for the observed difference.

Given that most students had some previous experience of *sample-to-population* inference, it did not surprise us that some students raised concern about representativeness of the groups. When asked to explain her pretest response, S1 commented:

> *To make accurate conclusions about the whole influence of fish oil we need to assess how these 14 people. . . are representative of the whole male population with high blood pressure. . . if they are not like typical people with high blood pressure we may have got biased results.*

Another commonly held belief, again perhaps attributable to prior experience of *sample-to-population* inference, was that the group sizes were too small for any

meaningful dialogue about the observed difference. Examples can be seen in the comments made by S1 and S3 in the pretest (see Table 4.3). In the posttest, S1 reflected on her pretest uneasiness with the group size and realized:

> *My concern was mainly about the sample size, just the small... yeah so I didn't think about the difference between the two groups just by chance.*

When asked if she now had more of an understanding of the chance acting alone concept, she responded:

> *Yeah, definitely. I would say that the whole like this, hypothesis, this would happen by chance... we need some statements that we can test. So one of them is whether this could happen by chance or not.*

Thus it would appear that S1 is now recognizing that chance may contribute to the difference observed between the two groups and is beginning to consider how to test for that.

As anticipated, many more students (61%, $n = 810$) suggested chance as an alternative explanation in the posttest. Interpreting brief written responses from the wider cohort such as "chance", and "chance is acting alone", it is difficult to know precisely how these students are now reasoning. However, their responses suggest that chance is now part of their thinking and vocabulary. When asked what he meant by chance, S3 responded by saying:

> *Just it could randomly occur. Pretty much it's possible for results to be just randomly different, I guess.*

While S3 acknowledges that randomness may be responsible for the observed difference, it is unclear if he has a sound grasp of the notion of chance acting alone.

S4's written posttest response (see Table 4.3) indicates that she believes that chance is present in both of her explanations for the observed difference. She reasons that even if the fish oil is effective, chance is also operating and therefore a chance component contributes, at least partially, to the observed difference between the two groups. This is a well-reasoned response, acknowledging that chance is always acting, even if the treatment is effective.

In summary, we reflect on the research question for Action 1 about the reasoning processes used by these students when thinking about the observed data from an experiment, making reference to the development of the underpinning ideas identified in Action 1 in Table 4.1. Reflecting on the pretest responses, we believe that the natural instinct of beginning students with no direct experience of the randomization test or principles of experimental design, when searching for explanations for the observed difference, was to rely on their prior inferential knowledge (*sample-to-population*) or to search for causes with which they felt comfortable, even if these causes were beyond the scope of the experiment. Within the *model fit* idea, the students seemed to bring their own experience to the observed data in a quest to offer explanations for the difference. Many were unable to access underpinning ideas of *experimental design and causation* since these did not form part of their prior knowledge. Given the lack of chance explanations in the pretest, it would appear that the

students did not attend to the *signal–noise* idea and that their natural reasoning processes did not entertain effects of randomness and chance.

In the posttest, responses from students conveyed more of an appreciation of the ideas encompassed within the notion of *experimental design and causation* with fewer concerns about representativeness, group size or confounding variables within Action 1. However, the *signal–noise* idea still needs to be further developed. Most of the chance explanations provided by these students and the wider cohort did not appear to convey the understanding that chance is always acting. Instead, the overriding impression was that most students had the underlying notion that either chance is acting alone, or chance is not acting at all. Such a problem is not surprising given that students only had two hours of instruction. However, we need to be aware that further instruction should address these two apparent conceptions:

- If the evidence favors a chance alone explanation, it excludes the possibility that the treatment may be effective.
- If the evidence favors a treatment explanation, it excludes the possibility that chance is acting alongside treatment.

Thus the idea that "treatment is effective" comprises both treatment and chance components, and the idea that "chance is acting alone" does not rule out the effectiveness of the treatment, is a learning issue that needs to be addressed. We believe that such conceptions can partly be attributed to the logic of the indirect argumentation associated with making a claim as a result of *experiment-to-causation* inference.

4.5.2 Action 2: Modeling Random Behavior

For Action 2 the students referred to Questions 1 and 2 (see Appendix). They were asked about the purpose of the test and to recall the randomization test in order to determine their understanding of how the distribution in Question 2 was formed. Note that Questions 1 and 2 were not presented in the format and with the representations used in the dynamic visualizations (for a comparison see Figure 4.1). Hence, the students needed to decode the representations given *and* recall the re-randomization process. To quantify the uncertainty on the distribution given in Question 2, the students needed to take the difference in means given in the table and plot it. Unlike the VIT software which gives the tail proportion visually and numerically when a button is activated, they had to recall the observed difference being plotted, the tail proportion being shaded in and then work out that they had to roughly count the number of differences equal or greater than 7.71. Before we elaborate on the student responses, an overall summary of the elements of reasoning and ideas that we were looking for is given in Table 4.4 along with student examples, codes, and descriptors for each of the elements.

Focusing on the responses of S1 to S4, we use the codes T1 to T5 (see Table 4.4) to illustrate how the randomization test was promoting ideas of uncertainty and where more development in students' reasoning appears to be needed.

S2 could state that the randomization test was determining "the chance of getting the result that we did in the circumstances where chance could be acting alone (T1)" and that there was a "mixing up of the conditions with the observations if everything was due to chance (T2)." He knew that each dot in the distribution represented a difference between the means (T3), that the process was repeated 1000 times and that a distribution developed (T4). However, his reasoning within element T4 faltered as he failed to connect that chance acting alone was visually represented by the distribution and he said he was "confused." He was also unable to obtain the tail proportion (T5). The interviewer asked, "so then what would be the observed difference if you were to plot that from this graph [points to graph in Question 1]." He responded, "I assume 7.71, oh right, it would be about there [he locates 7.71 on the distribution and puts a box around the tail proportion], yeah I've got it now." At that moment he connected the steps in the procedure for the randomization step, found the tail proportion and quantified the uncertainty (T5). Thus we conjecture S2 had a fragmented understanding of the randomization test process. He is developing ideas of a repeated chance process forming a distribution but is not yet fully connecting the underlying concepts.

S3 was able to succinctly describe the purpose of the randomization test: "You have a measurement of the difference and with the randomization test you measure how likely it is that chance alone will produce the same difference (T1)." He then followed with a description of the randomization test process.

> *They separated the results from the group and then just randomly assigned them in a resample* (T2), *and then they took the mean difference of that* (T3) *and then repeated that process a 1000 times in this case, and it's got a distribution of what was possible by chance alone* (T4), *and then compared the result that they got from the actual test with the distribution, to get a tail of how likely it was* (T5), *if it was just chance.*

For S3 one of the dots in the distribution "would represent chance." It is "just one difference between the means for a re-sampling." Hence unlike S2, he seems to make the connection between the notion of chance alone and that the distribution is a visual representation of chance alone. He was also able to quantify the uncertainty by putting 7.71 on the distribution and calculating the tail proportion, as did S1 and S4.

Table 4.4
Summary of Elements of Reasoning within Action 2

Element	**Description of Ideas**	**Student example**[†]
Purpose of test (T1)	Test observed difference if chance is acting alone (Assumption treatment has no effect)	So you say if everything was due to chance what would it look like and what would our probability be of getting the same result if we did? What would be the chance of getting the result that we did in the circumstances where chance would be acting alone? (S1)
Simulating random reallocation, uncertainty (T2)	Notion of randomness of who gets into which group	You take all the results and instead of having them split into two groups, you take them into one group and then split them into two randomly. (S6)
Measuring under uncertainty (T3)	Record and interpret differences in center	Well when you re-randomize them and take the difference which is what this one does [refers to imagined plot], it gives you that, it just shows you what possible values you would get for the difference if it were purely chance. (S6)
Distribution of possible measures under uncertainty (probability model; T4)	Repeat T2 and T3 many times (repeatable process) Build distribution of a statistic (interpret what is measured)	And then repeated that process a thousand times in this case, and it's got a distribution of what was possible by chance alone. (S3) *So if we just take one dot here, what is that?* (I) That is just one difference between the means, from a resampling [re-randomization]. (S3)
Quantification of uncertainty (tail proportion; T5)	Purpose of putting observed difference on re-randomization distribution Read the tail proportion	You would be able to see how big the observed difference was between the two and how it compares to the first sample. (S4) *Where is the observed difference located on this plot?* (I)

[†] I = Interviewer

Table 4.4 – continued from previous page

Element	Description of Ideas	Student example[†]
		I think I thought the observed difference was just there for me. Yeah 7.7. The tail proportion [refers to 7/1000 she has given on posttest] is the probability that the observed difference occurs. (S4)

[†] I = Interviewer

S1 gave a good indication she was recalling the visual images of the VIT software because when asked about what happened in the middle plot she said: "just click on re-sampling and then the sample was rearranged and re-sampled and then a new difference was shown. [The difference was shown as a] red arrow." Note that three of the students S1, S3, and S2 (not shown) used the term re-sample rather than re-randomization. In the instruction we used re-sample for the bootstrap method and re-randomization for the randomization test, but this careful distinction in terminology to reinforce the difference between *sample-to-population* inference and experiment-to-causation inference bypassed many students.

In summary we reflect on the research question for Action 2 about the ideas and reasoning processes about uncertainty that these four students used when recalling the randomization test (Figure 4.1) and their development of the underpinning ideas identified in Action 2 in Figure 4.1. We conjecture that these students seemed to have the notion the observed difference is tested against chance alone. Within the *model fit* idea the students seemed to be expecting the distribution given in Question 2, as they did not query it, and all knew the *distribution* had been generated from a *random process* that was repeated many times suggesting a notion of the *Law of Large Numbers*. Their references to the data being mixed up and re-allocated to the two groups suggest unarticulated ideas of *independence*. The notion of a chance distribution generated from the recording of the differences in means for each re-randomization was also recognized by the students. However, only three of these four students seemed to understand the purpose of representing the tail proportion in that distribution and hence were able to quantify the uncertainty, the tail proportion. S2 failed to cognitively integrate that the distribution was an image of possible chance outcomes (noise), against which the original observed difference is tested or signal is detected. The data suggest that this idea of *signal–noise* within Action 2 is one of the more difficult concepts for students to grasp.

We believe that the dynamic imagery of the VIT software facilitated the ability of students to recall many of the ideas underpinning the randomization test and, hence, was assisting them in developing concepts associated with uncertainty in the context of making partially informal inferences. The notions of uncertainty that need to be developed within this test are many-fold and multi-faceted. There are ideas of testing against uncertainty and simulating, mimicking and measuring or quantifying uncertainty with a chance alone distribution. The generation of this distribution allows a

quantification of uncertainty for a causal inference, an uncertainty idea that is quite different from the uncertainty ideas for the development of the distribution. Hence a causal inference is predicated on quantifying the uncertainty of an observed difference or greater from the quantification of an "uncertainty" distribution that has been generated from a random process.

4.5.3 Action 3: Making a Claim about the Data

Following Question 2 (Appendix, posttest) students were asked during the interview for their interpretation of the tail proportion; if it was small or large. (Note: For these introductory students the guideline for determining whether chance was not acting alone that was used in instruction was less than 10%.) They were also asked to elaborate on what they were thinking when they wrote their responses to Question 3 (see Table 4.5). Using the responses of S1, S4, S5, and S6 we explore the claims they are prepared to make about the data, and the reasoning underpinning their argument.

Table 4.5
Examples of Some Student Responses to Question 3 (see Appendix)

Student	**Written Response**
S1	"A study, conducted on a control group, showed that, most probably, fish oil diet reduces blood pressure for those with a high blood pressure." The original statement concerned all people, while the study was conducted (aimed at) for people with high blood pressure.
S4	The study was based on a sample so we cannot claim that a fish oil diet lowers blood pressure for the whole population. Also an interval should be given to show there is uncertainty. The study ruled out that chance is acting alone but this means that the lowered blood pressure could be a combination of chance and the fish oil diet. "It is a fairly safe bet that the mean reduction in blood pressure of those on a fish oil diet is higher than the mean reduction in blood pressure of those on a regular oil diet."
S5	The line is too definite. "People have a good chance of lowering their blood pressure with a fish oil diet" would be more accurate and acceptable
S6	The test shows that blood pressure was lowered in comparison to the regular oil diet, and there has only been a causal relationship drawn as only chance has been ruled out by re-randomization. A more accurate statement would be "Fish oil diet probably lowers blood pressure in comparison to regular oil diet. More research needed."

S5's interpretation of a small tail proportion was, "chance is not acting alone", which he later changed to "chance is probably not acting alone", because he did not want to definitely say the treatment was effective, although there was a "good chance" fish oil lowered blood pressure (see Table 4.5). Sample size was his grounds

for not being definite; "they only used 14 people." S6's interpretation of a small tail proportion was that:

> *It's very unlikely chance is acting alone, something else is acting... it might not be the fish oil; it probably is... more research is needed. The fish oil diet probably lowers blood pressure.*

Although he recognized that a causal relationship could be drawn as chance had been ruled out (see Table 4.5), his grounds for not making a definite statement were based on possible confounding variables because "people have all sorts of different diets."

S1 stated a small tail proportion suggested that for the observed difference:

> *It would be assumed that it is really unlikely that it happened by chance. It is highly probable that there exists a causal relationship between the fish oil diet and reduction in blood pressure.*

Her quantification of uncertainty about a causal inference could be a step towards considering that the observed difference may be a rare occurrence and in fact she may be drawing an incorrect conclusion. She also recognized that generalization was a notion that needed attention when formulating an inference argument (see Table 4.5). The influence of *sample-to-population* inference, however, led S1 to reason that there was uncertainty about the treatment being effective "because of sampling, not to make a statement directly about all people... [there] is sampling error... quality of our sample group." Similarly S4's reasoning was based on *sample-to-population* inference (see Table 4.5) but she did not want to make an inference from a sample suggesting some possible misconceptions within this arena. She explained her argument in terms of chance is not acting alone and "there is evidence that it is just not chance but it could also be a combination of some things... it could be fish oil or not." Although she could correctly verbalize that "treatment is effective" is comprised of two components, chance and something else acting alongside chance such as treatment, she stated, "we can't say concretely that it is the fish oil diet that worked." She said her use of the language

> *It's a fairly safe bet (see Table 4.5) includes the uncertainties, so you are not committing yourself to saying it does lower blood pressure... you are talking about the average and not something else... it gives the idea that it is based on a sample as well and not for everyone in general.*

Note that her reasoning about uncertainty is expressed through referring to the average, a tendency idea that some other students more explicitly expressed (e.g., "We can say it can help lower blood pressure, but not definitely work on each person who used a fish oil diet") and a *sample-to-population* inference notion.

For a small tail proportion drawing a causal inference has not yet crystallized, as the students' arguments show the influence of their prior knowledge such as sample size and *sample-to-population* inferences and contextual knowledge of other factors that could affect blood pressure. They express their uncertainties in probability language, suggesting they are attaching a likelihood to the treatment effectiveness, a facet which is further illustrated in their interpretation of a large tail proportion.

For a large tail proportion, for example 30%, S5 said, "chance is acting alone" which he changed to "chance probably is acting alone" because 30% to him was

the probability that chance is acting alone; "it's kind of like in the middle. You'd be leaning towards chance is acting alone but not definite." Thus, there was a chance that "fish oil could lower blood pressure." S6 provided a similar explanation for a large tail proportion:

> *The researchers could not conclude anything. The tail proportion shows evidence that chance is acting alone, but it does not prove if it is. As a result, no conclusion could be drawn, as there is a 30% chance of chance acting alone.*

S6 is correct that no conclusion can be drawn but his reasoning is incorrect. He further elaborated on his reasoning processes for small and large tail proportions, which shows his thinking about how the tail proportion gives the probability of chance acting alone.

> *Well, five out of 1000 is tiny. You wouldn't see that sort of thing happening often but a 30% chance is one in three, it's quite likely, will probably happen. If you had a choice between a poison that kills you 0.5% of the time and one that kills you 30% of the time, you'd drink the first poison.*

S1 also indicated her reasoning was along the same line as S5 and S6, as she reported, "the probability is quite high that it happened by chance," and that chance was responsible, "just a chance that people were spread into groups like this." She also stated this one test was insufficient to determine whether the treatment was effective; "this test wouldn't be a good base to conclude anything about the real influence of fish oil on blood pressure," indicating that she did not have a sound grasp of the principles behind experimental design or she was searching for a biological mechanism to explain why fish oil could be effective in reducing blood pressure. S4 explained her argument as "chance is acting alone" and "that difference would be because of chance and not because of diet." Thus she is reasoning from the position of chance is acting alone versus chance is not acting alone.

From the student interviews and other student responses ($n = 695$) to Question 3 we proposed four notions within uncertainty that needed to be attended to when making a claim statement, which we named: Rare occurrence (R), Causal evidence (C), Tendency (T), and Generalization (G) (see Table 4.6 for an indication of reasoning and ideas behind each of these proposed elements). Hence, we hoped to see language in students' statements alluding to these notions, such as:

> *We are pretty sure* (R) *that a fish oil diet causes* (C) *males with high blood pressure* (G) *to tend* (T) *to have a higher reduction in blood pressure than those on a regular oil diet. We need to be careful about generalizing beyond the group in the study* (G).

Many students changed the word "will" in the original Question 3 statement to "may", or to something similar, but we were unable to tell from many of their explanations which one of the identified uncertainty notions they were using in their reasoning. It also may have been that they were using other notions of uncertainty (e.g., believing that all knowledge is uncertain; "Nothing is 100% certain. You cannot state for a definite fact that fish oil will lower blood pressure."), or using their *sample-to-population* knowledge and believing the group sizes were too small. From the student interviews already discussed there is evidence that these notions of uncertainty, as well as others not listed here, were being invoked.

In summary we reflect on the research question for Action 3 about the argumentations students use when making a claim about data from an experiment. Within the *signal–noise* idea the students did not seem to have yet grasped that the reasoning is about whether a treatment effect is detectable under chance variation, not about whether the treatment is effective or not, or chance is acting alone or not. By using the latter two reasoning processes, they appeared to invoke a plethora of notions about uncertainty and misconceptions related to comprehension of principles of *experimental design and causation* and *inference argumentation*. For example, confounding variables and group size seemed to become the rationale for a reluctance to consider causality; the tail proportion becomes the probability of the treatment being effective; and other issues, such as the tendency notion, generalization, and all knowledge is uncertain influence their argumentation.

Students' argumentation in Action 3 is not surprising since they only experienced a two-lecture introduction to *experiment-to-causation* inference and the VIT tools only give a visual image of the tail proportion within a distribution. Hence, they express misconceptions previously identified with p-values (e.g., Nickerson, 2004). Our findings, however, uncover a wider range of thinking within uncertainty that is invoked with *experiment-to-causation* inference and that will need to be addressed in further instruction.

4.5.4 Summary of Student Notions about Uncertainty

Our aim in this chapter was to uncover new considerations about students' reasoning processes regarding causality and uncertainty in the context of making partially informal inferences. From our analysis of student reasoning processes within each of the three actions that occur towards making an *experiment-to-causation* inference, we have uncovered twelve elements within uncertainty (Table 4.6) that we think need to be addressed in instruction to enable students to appreciate and grasp more fully the thinking and argumentation underpinning the designed experiment and the randomization test.

A two-lecture introduction was insufficient, as was the case in this study, for students to understand the implications of experimental design, and to overcome prior knowledge, such as *sample-to-population* inference. As is prevalent within society, these students seemed to engage in deterministic thinking as they sought other causes beyond the cause set up by the experiment. Hence, to understand causality in a probabilistic setting, our students needed more time for exploration and experience, which they did obtain in a further four weeks of instruction later on in the course.

Table 4.6
Summary of Elements within Uncertainty that May Need to be Addressed in Instruction for Experiment-to-Causation inference

Element	**Description of reasoning and ideas**
Causal evidence	Understanding that in a properly executed randomized comparative experiment causality can be established if the values of the explanatory variable (treatment) are randomly assigned to the units.
Randomization Test	Understanding the purpose of the test and reasoning and ideas underpinning the quantification of uncertainty towards experiment-to-causation inference (see Figure 5).
Tail Proportion	Understanding that the aim is to detect a signal, the treatment effect, under the obscuring effects of noise or chance variation. A small tail proportion indicates a signal has been detected, while a large tail proportion indicates a signal has not been detected suggesting that noise could be obscuring the signal or there could be no signal, just noise, implying that a claim cannot be made. (See discussion section on this metaphor.)
Treatment is effective	Understanding that the treatment is effective element is composed of a chance component and a treatment effect component.
Rare occurrence	Realizing the possibility, although small, that a difference in centers at least as large as that observed could happen by chance alone. That is, the observed difference may be a rare occurrence and the wrong inference may have been made (Type 1 error—not covered in our two-lecture introductory instruction).
Generalization	Understanding that care must be taken with any generalization to a wider group than those in the study who were volunteers with particular characteristics (e.g., male, high blood pressure). The population is all those who participated in the experiment. Inappropriate to think about a wider population.
Tendency	Understanding that the inference is about the tendency of the treatment group as a whole to improve, not every individual.
Confounding variables	Understanding that unknown or potential confounding variables can be treated as chance explanations, which are accounted for in the method of random assignment and in the re-randomization distribution.
Design issues (e.g., group size)	Realizing a design issue such as group size is not a problem. Understanding that smaller group sizes require a large observed difference in centers in order to detect whether the treatment is effective under the obscuring effects of chance variation compared to larger group sizes.
Sample-to-population inference	Realizing that a designed experiment uses volunteers, does not take a sample from the population, and does not aim to make an inference about a population; rather it aims to make an inference about an intervention.

Table 4.6 – continued from previous page

Element	Description of reasoning and ideas
Contextual knowledge	Understanding that claims are based on the data in hand and that contextual knowledge, for example, about possible biological mechanisms for the observed difference in centers is used for the next stage of an investigation. Realizing that one's own contextual knowledge and beliefs can bias perceptions or leads one's thinking astray.
All knowledge is uncertain	Acknowledging that there are other sources of uncertainty such as quantification of uncertainty for statistical inference as well as the uncertainty about current knowledge being overturned in the future.

4.6 Discussion and Implications

Research has largely focused on *sample-to-population* inference and has consistently documented a tendency for students to think deterministically or causally and to not take sample size into account (e.g., Kahneman, 2011; Meletiou-Mavrotheris, Lee, & Fouladi, 2007). Students, however, when first introduced to *experiment-to-causation* inference do not seem to be willing to use causal thinking from the designed experiment. Rather, they tend to focus on many considerations of uncertainty and causality, such as sources of variation within the study design, the idea that all knowledge is uncertain, the group size is too small (not a concern in this situation), not every individual case benefits from the treatment, the observed difference might be a rare occurrence, the group of people on whom the experiment was conducted were volunteers, wondering about the biological mechanism behind the treatment, chance explanations, and chance is acting alone. The idea that chance and treatment act alongside one another also needs to be addressed in instruction. Making judgments from data, therefore, involves students in untangling considerations regarding uncertainty in the realms of statistical inference, data quality, and data validity.

Cognitively coordinating, attending to, and building conceptions of uncertainty for *experiment-to-causation* inference requires a teaching sequence that gradually develops more sophisticated notions of uncertainty and causality which addresses the elements of uncertainty identified in Table 4.6. The integrated textbook, *Core-Plus Mathematics* (Hirsh, Fey, Hart, Schoen, & Watkins, 2008), has learning trajectories that address experiments and causation using randomization tests with *CPMP-Tools* (Keller, 2006). Hart, Hirsch, and Keller (2007) believe these tools provide cognitive amplification, resulting in a conceptual understanding of inference. Also of note, is Madden's (2008) research, which, although addressing different research questions, was able to demonstrate that high school mathematics teachers who participated in a four-day professional development course could successfully compare distributions using the randomization test with *CPMP-Tools* and *Fathom*™ (Finzer, 2005). Similar to the findings of many researchers about developing students' probabilistic reasoning (e.g., Garfield et al., 2012; Konold & Kazak, 2008), however, we conjecture that developing students' understanding of causality in a probabilistic setting will require multiple experiences over several years.

The six underpinning interconnected ideas (model fit; signal–noise, principles of experimental design; the Law of Large Numbers, random process and independence; distribution; and inference argumentation) change subtly as students move from thinking about the observed data, to conducting a randomization test by modeling the random behavior of reallocating units to two groups, to making a claim about the data. These interconnected ideas involve uncertainty related to statistical inference for experiments. Grappling with understanding the concepts underpinning the quantification of statistical inference, and interacting with many considerations about uncertainty are learning experiences that these students will need in further instruction. The dynamic visual imagery, resources, verbalizations and teacher–student discussion used in this study are, we believe, a small step in the right direction for students to begin to appreciate uncertainty in its many guises.

Inference argumentation, however, as our findings suggest, requires further development and is in accordance with other research on interpretation of the p-value (e.g., Nickerson, 2004). Such a finding is not surprising given the limitation of a two-lesson introduction to the randomization test and the nature of the argumentation, which is incumbent on reasoning about detecting a signal within the abstract notion of a chance alone distribution. To improve students' inference argumentation we suggest that future research explore new metaphors, visual imagery, and verbalizations. For example, we used the language "chance is acting alone" as an explanation of the observed difference in centers, and for arguing from the re-randomization distribution. Chris Wild (personal communication, November 28, 2013) suggests using language such as, "can randomization do this?" and "compare what we have got with what randomization alone can deliver", where randomization must be understood as random assignment of units to two groups. Although such language may be less abstract than "chance is acting alone", the downside is that it does not immediately lead to more universal ideas further along conceptual and learning trajectories for teaching inference. Also the signal–noise metaphor for inference argumentation (see tail proportion element in Table 4.6) may provide students with better visual imagery, such as an association with detecting signals in outer space to learn about the argument, rather than chance is acting alone imagery

Causal inference also needs attention. Its long history is intertwined with philosophical argumentation as perspectives and ideas have changed; from the attribution of causes to gods or people to physical objects; from making causal inferences based on correlation ideas to using the randomized experiment, and more recently, from observational studies (Pearl, 1996). Therefore, teaching approaches need to acknowledge students' intuitive reasoning, prior knowledge and general philosophical approach to argumentation. With students' propensity to think deterministically, and not to appreciate the role of chance ((Fischbein, 1975), it appears from our findings that it may take time for students to grasp the idea of a causal inference and the role of chance in evaluating evidence of causality.

While more research is needed to modify the theoretical frameworks for statistical inferential thinking, the conceptual pathways in the curriculum, and to understand students' reasoning about uncertainty, teaching in a completely new way through using the randomization test and focusing on *experiment-to-causation* inference has

opened up and revealed the depth of thinking that is needed to grasp more fully the issues surrounding conceptions of uncertainty and causality.

Acknowledgements

This study is supported in part by a grant from the *Teaching and Learning Research Initiative* (http://www.tlri.org.nz). We also thank Chris Wild for his comments on this chapter and our three research assistants, Kate Aloisio, Kimberley Eccles, and Savannah Post.

References

Agresti, A., & Franklin, C. (2007). *Statistics: The art and science of learning from data*. Upper Saddle River, NJ: Pearson Prentice Hall.

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering questions from existing data* Unpublished doctoral dissertation. The University of Auckland.

Arnold, P., & Pfannkuch, M. (2012). The language of shape. In *Proceedings of the 12th International Congress on Mathematics Education, Topic Study Group 12, 8–15 July, Seoul, Korea* (pp. 2446—2455). Seoul, Korea: ICME-12. Retrieved from http://icme12.org/

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, The Netherlands: CD-β Press, Center for Science and Mathematics Education.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer Academic Publishers. doi: 10.1007/1-4020-2278-6

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—The International Journal on Mathematics Education*, *44*, 913–925. doi: 10.1007/s11858-012-0420-3

Biehler, R. (2011). *Five questions on curricular issues concerning the stepwise development of reasoning from samples.* Presentation at the Seventh International Forum on Statistical Reasoning, Thinking and Literacy, 17–23 July 2011, Texel, The Netherlands.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. (2013). Dynamic visualizations and the randomization test. *Technology Innovations in Statistics Education*, *7*(2), 1–21. Retrieved from http://escholarship.org/uc/item/9dg6h7wb

Cobb, G. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1), 1–15. Retrieved from http://escholarship.org/uc/item/6hb3k0nz

College Board. (2010). *Advanced Placement statistics course description*. Author. Retrieved from http://media.collegeboard.com/digitalServices/pdf/ap/ap-statistics-course-description.pdf

Common Core State Standards Initiative. (2010). *Common Core state standards for mathematics. common Core state standards (college- and career-readiness standards and K12 standards in english language arts and math)*. Washington, D.C.: National Governors Association Center for Best Practices and the Council of Chief State School Officers. Retrieved from http://www.corestandards.org

Falk, R., & Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*(1), 75–98. doi: 10.1177/0959354395051004

Finzer, W. (2005). *Fathom™ dynamic statistics software (version 2.0).* Emeryville, CA: Key Curriculum Press.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, The Netherlands: Reidel.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

Friel, S., Curcio, F., & Bright, G. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–159. doi: 10.2307/749671

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 39–63). New York: Kluwer/Springer Academic Publishers. doi: 10.1007/0-387-24530-8_3

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM—The International Journal on Mathematics Education*, *44*(7), 883-898. doi: 10.1007/s11858-012-0447-5

Greer, B., & Mukhopadhyay, S. (2005). Teaching and learning the mathematization of uncertainty: Historical, cultural, social and political contexts. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 297–324). New York: Kluwer/Springer Academic Publishers. doi: 10.1007/0-387-24530-8_13

Hart, E., Hirsch, C., & Keller, S. (2007). Amplifying student learning in mathematics using curriculum-embedded, Java-based software. In W. G. Martin & M. E. Strutchens (Eds.), *The learning of mathematics: Sixty-ninth yearbook* (pp. 175–202). Reston, VA: National Council of Teachers of Mathematics.

Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., & Epstein, R. (2009). Bootstrap methods and permutation tests. In D. Moore, G. McCabe, & B. Craig (Eds.), *Introduction to the practice of statistics* (6th ed., pp. 16-1–16-60). New York: Freeman.

Hirsh, C., Fey, J., Hart, E., Schoen, H., & Watkins, A. (2008). *Core-Plus mathematics: Contemporary mathematics in context*. New York: Glencoe McGraw-Hill.

Hjalmarson, M., & Lesh, R. (2008). Engineering and design research: Intersections for education research and design. In A. Kelly, R. Lesh, & K. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 96–110). New York: Routledge.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux.

Keller, S. (2006). *CPMP-Tools.* East Lansing, MI: Michigan State University and Core-Plus Mathematics Project.

Knapp, H., & FitzGerald, G. (1989). The antihypertensive effects of fish oil. a controlled study of polyunsaturated fatty acid supplements in essential hypertension. *New England Journal of Medicine*, *321*(23), 1610–1611. doi: 10.1056/NEJM198904203201603

Konold, C., Higgins, T., Russell, S., & Khalil, K. (2004). *Data seen through different lenses.* (Unpublished manuscript)

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1). Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1/

Langrall, C., & Mooney, E. (2005). Characteristics of elementary school students' probabilistic reasoning. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 95–119). New York: Kluwer/Springer Academic Publishers. doi: 10.1007/0-387-24530-8_5

Liu, Y., & Thompson, P. (2007). Teachers' understandings of probability. *Cognition and Instruction*, *25*(2), 113–160. doi: 10.1080/07370000701301117

Lock, R., Lock, P., Lock, K., Lock, E., & Lock, D. (2013). *Statistics: Unlocking the power of data.* Hoboken, NJ: John Wiley & Sons, Inc.

Madden, S. R. (2008a). *Dynamic technology scaffolding: A design principle with potential for supporting statistical conceptual understanding.* Paper presented at the International Congress on Mathematics Education (ICME-11), Monterrey, Mexico. Retrieved from http://tsg.icme11.org/document/get/479

Madden, S. R. (2008b). *High school mathematics teachers' evolving knowledge of comparing distributions* Doctoral dissertationKalamazoo, MIWestern Michigan University. Retrieved from http://iase-web.org/documents/dissertations/08.Madden.Dissertation.pdf

Madden, S. R. (2011). Statistically, technologically, and contextually provocative tasks: Supporting teachers' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1–2), 109–131. doi: 10.1080/10986065.2011.539078

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173. doi: 10.1080/10986065.2011.538301

Meletiou-Mavrotheris, M., Lee, C., & Fouladi, R. (2007). Introductory statistics, college student attitudes and knowledge—a qualitative analysis of the

impact of technology-based instruction. *International Journal of Mathematical Education in Science and Technology*, *38*(1), 65–83. doi: 10.1080/00207390601002765

Ministry of Education. (2007). *The New Zealand curriculum*. Wellington, New Zealand: Learning Media.

Nickerson, R. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pearl, J. (1996). The art and science of cause and effect. Public lecture delivered at UCLA 81st Faculty Research Lecture Series. In J. Pearl (Ed.), *Causality: Models, reasoning, and inference* (2nd ed., pp. 401–428). New York: Cambridge University Press.

Pearl, J. (1997). *The new challenge: From a century of statistics to the age of causation.* Paper presented at the International Association for Statistical Computing Second World Congress, Pasadena, CA.

Pfannkuch, M., Wild, C. J., & Regan, M. (2013). Students' difficulties in practicing computer-supported statistical inference: Some hypothetical generalizations from a study. In T. Wassong, D. Frischemeier, P. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit werkzeugen mathematik und stochastik lernen [Using tools for learning mathematics and statistics]* (pp. 393–403). Wiesbaden, Germany: Springer Spektrum. doi: 10.1007/978-3-658-03104-6

Pfannkuch, M., & Ziedins, I. (2013). A modeling perspective on probability. In E. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 101–116). New York: Springer. doi: 10.1007/978-94-007-7155-0

Pratt, D. (2005). How do teachers foster students' understanding of probability? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 171–189). New York: Kluwer/Springer Academic Publishers. doi: 10.1007/0-387-24530-8_8

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29*(3), 343–367. doi: 10.3102/10769986029003343

Sacristan, A., Calder, N., Rojano, T., Santos-Trigo, M., Friedlander, A., & Meissner, H. (2010). The influence and shaping of digital technologies on the learning—and learning trajectories—of mathematical concepts. In C. Hoyles & J. Lagrange (Eds.), *Mathematics education and technology—rethinking the terrain: The 17th ICMI study* (pp. 179–226). New York: Springer.

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? simple associations may explain moral reasoning in infants. *PLOS One*, *7*(8), 1–4. doi: 10.1371/journal.pone.0042698

Teague, D. (2006). Experimental design: Learning to manage variability. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 151–169). Reston, VA: National Council of Teachers of Mathematics.

Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory

statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21–40. Retrieved from http://www.stat.auckland.ac.nz/serj

Watson, J. (2005). The probabilistic reasoning of middle school students. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 145–169). New York: Kluwer/Springer Academic Publishers. doi: 10 .1007/0-387-24530-8_7

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. doi: 10.1111/j.1751-5823 .1999.tb00442.x

Wild, C. J., Pfannkuch, M., Regan, M., & Parsonage, R. (2013). *Next steps in accessible conceptions of statistical inference: Pulling ourselves up by our bootstraps.* (Unpublished manuscript)

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education*, *13*(3), 1–11. Retrieved from http://www.amstat.org/publications/jse/

Appendix: Appendix: Assessment Items from the Pretest and Posttest

Pretest and posttest scenario: (Knapp & FitzGerald, 1989)

Fish Oil and Blood Pressure Study
Prior to conducting this study the researchers conjectured that those on a fish oil diet would tend to experience greater reductions in blood pressure than those on a regular oil diet. Researchers randomly assigned 14 male volunteers with high blood pressure to one of two four-week diets: a fish oil diet and a regular oil diet. Therefore the treatment is the fish oil diet while the regular oil diet is the control.

Each participant's blood pressure was measured at the beginning and end of the study, and the reduction was recorded. The resulting reductions in blood pressure, in millimetres of mercury, were:

Fish oil diet:	8	12	10	14	2	0	0
Regular oil diet:	−6	0	1	2	−3	−4	2

Plots of the data and summary statistics are:

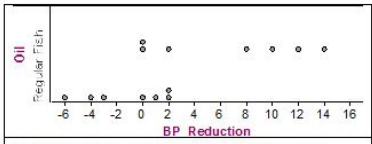


Figure 2. Dotplots of reductions in blood pressure

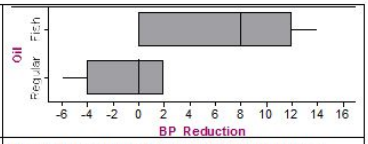


Figure 3. Box plots of reductions in blood pressure

Variable	Oil	N	Mean	SD
BP Reduction	Fish	7	6.57	5.86
	Regular	7	−1.14	3.18
Difference			7.71	

Table 1. Summary statistics comparing blood pressure reductions of the fish oil and regular diet groups

Pretest and posttest questions (NB: In the posttest "statistical test" in Question 1, part B was changed to "randomization test")

The observed data in Figures 2 and 3 show that the reduction in blood pressure values for the fish oil group tends to be greater than those for the regular oil group. Write down the TWO MAIN possible explanations for this **observed difference** as shown **in Figures 2 and 3**.

A. The two main possible explanations for this observed difference are:

i. _____

_____

ii. _____

_____

B. Which **ONE** of your possible explanations (i. or ii.) would the researchers test using a *statistical test*?

_____

Posttest Question 2

After looking at plots of their data and summary statistics the researchers conducted a randomisation test. The distribution of the differences in the means arising from 1000 re-randomizations is shown in Figure 4.

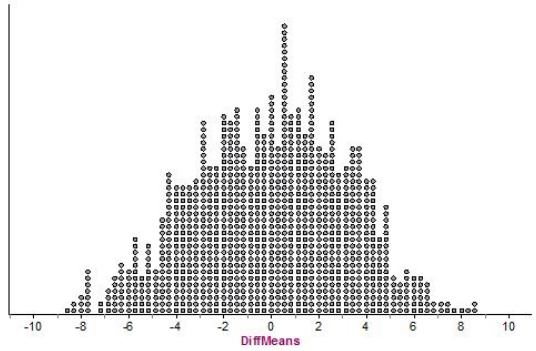


Figure 4. Distribution of the difference in means arising from 1000 re-randomizations

Approximately, what is the tail proportion as a fraction in the re-randomization distribution in Figure 4?

Answer: The tail proportion as a fraction is about _____

Posttest Question 3

In reporting the findings of this study a newspaper stated:

People will lower their blood pressure with a fish oil diet.

This statement is statistically incorrect. Explain why and rewrite the statement correctly.

CHAPTER 5

PRESERVICE TEACHERS' REASONING ABOUT UNCERTAINTY IN THE CONTEXT OF RANDOMIZATION TESTS

Rolf Biehler, Daniel Frischemeier, and Susanne Podworny

Institute of Mathematics, University of Paderborn

Abstract

We investigate the reasoning of preservice teachers about uncertainty in the context of randomization tests facilitated by TinkerPlots™. This method of hypothesis testing is a widely used method to look beyond the comparison of two groups and to generalize findings beyond a sample. To support preservice teachers while they conducted these tests, we developed two courses to lead them to randomization tests: One course moves from data analysis to randomization tests; the other from simulations to randomization tests. A video study which includes the observation of preservice teachers while conducting randomization tests with TinkerPlots™ placed at the end of both courses will be the focus of this chapter. Finally, this chapter will give insights into how the process of randomization tests can be supported for learners and will outline a group of German preservice teachers' encounters with randomization tests.

Keywords: Informal inferential reasoning; Randomization tests; preservice teachers; TinkerPlots™

5.1 Overview

The education of preservice teachers in mathematics at the University of Paderborn consists of three domains: mathematics, didactics of mathematics, and pedagogy. An obligatory mathematics course is called *Elementary Statistics and Probability Theory*, in which the participants are introduced to the basics of statistics and probability. In addition, preservice teachers can attend a seminar, which deepens this course in a succeeding term. The authors of this chapter have designed two of these seminars, incorporating 15 sessions into each (Frischemeier & Biehler, 2012; Podworny, 2013). The first course deals with exploratory data analysis with TinkerPlots™ and is called *Developing Statistical Reasoning Using the TinkerPlots™ Software*. The second course, *Applied Stochastics—Understanding and Solving Complex Problems with Simulations*, is about simulation with TinkerPlots™. With each course we address reasoning about uncertainty in the context of exploring real data. This finishes with randomization, which is a topic highly related to *Reasoning about Uncertainty in the Context of Making Informal Statistical Inferences*. A common final goal in both courses is supporting the participants in learning to draw conclusions from data via randomization tests. For information concerning randomization tests, see Edgington and Onghena (2007), Ernst (2004), and Zieffler, Harring, and Long (2011). Both courses have different approaches and routes but both finally lead to randomization tests. The first course's route is from data analysis over group comparisons to randomization tests and the second course's route is from simulation to randomization tests. Note that due to organizational reasons there were no participants who attended both courses. Randomization tests are the final topic in both courses and we want to investigate the reasoning of preservice teachers about uncertainty in this context. After the preservice teachers gained experience with informal methods of drawing conclusions from data we introduced null hypothesis testing, where p-values were used as way of quantifying uncertainty. Drawing conclusions from data was a fundamental aspect in both courses, but conducting a randomization test turned out to be a very difficult statistical task on many levels; for example developing an adequate null hypothesis, the calculation of the possible arrangements and permutations of the data, the computation of the p-value, and the correct conclusions from computed p-values with regard to the null hypothesis. The simulation method supported by the use of adequate software can make the process much easier and can support learners in their modeling process. Educational software, in particular TinkerPlots™, enables the demonstration of the randomization process itself: The random assignment in the sampler is visible.

To analyze preservice teachers' thinking and reasoning while conducting a randomization test with TinkerPlots™, we conducted a video study after each course. In this study, the preservice teachers were recorded while doing a randomization test with TinkerPlots™. Results of this study are presented in this chapter. We will further present a scheme representing a whole cycle of testing and further material which is supposed to support learners while doing these tests.

Note, that we did not carry out an experimental comparative study with random assignment of subjects to the two courses so that we could have made a causal in-

ference from the two different treatments to the quality of preservice teachers' reasoning. Nevertheless the comparison of the two groups will provide us with some insights and hypotheses about the possible effects of the two different routes to randomization tests.

5.2 Problem

In Germany, randomization tests are widely unknown in textbooks at the secondary and tertiary level. Consequently, there is no research concerning German preservice teachers' coping with randomization tests. In the German school curriculum even the "testing" of hypotheses with p-values does not appear very often. In grade 11 or 12, formal hypothesis testing with pre-defined significance levels is a topic in several federal states, but often not obligatory for the final examinations. In general it is reported that pupils, students and preservice teachers have many misconceptions about hypothesis testing independent of the kind of test procedure. In our courses we do not teach how to compute p-values by using probability distributions, but use the software TinkerPlots™ for estimating p-values via simulation. This is a more informal approach, taken from the informal inferential reasoning discussion (see e.g., Zieffler, Garfield, delMas, & Reading, 2008). We see many advantages in using a software capable of simulation (e.g., Biehler & Maxara, 2007; Konold, Harradine, & Kazak, 2007; Meyfarth, 2008). Most participants in our courses have no or very little knowledge about hypothesis testing, since this is not part of the obligatory course, *Elementary Statistics and Probability*.

Similar to delMas, Zieffler, and Brown (2013), who developed an introductory course with the goals of developing students' informal and formal statistical inferences, we developed two courses leading to randomization tests (see Section 3.4). In this article we focus on German preservice teachers working on a randomization test task (see Section 3.4). Because they have never previously experienced the idea of randomization, we want to address the following research questions:

1. How do the preservice teachers accomplish the steps of a randomization test?

2. How well are the preservice teachers able to model a randomization test experiment with TinkerPlots™ ? What is the role of TinkerPlots™ in their thinking?

3. How do the preservice teachers interpret the results of the randomization test?

5.3 Literature and Background

P-values and their interpretations present many difficulties for learners. For an overview of (mis-)conceptions regarding hypothesis testing with p-values, see Garfield and Ben-Zvi (2008, p. 270). These authors report problems with questions like the one mentioned previously concerning students generalizing a result found in a sample.

An opportunity for emergent inferential reasoning, especially in connection with results from group comparisons, is a randomization test. Ernst (2004) describes a detailed and formal introduction into randomization tests but the question arises as to how this kind of reasoning can be implemented in mathematics teaching at earlier levels? Rossman (2008) recommends starting inferential reasoning with randomization tests. His introduction to randomization tests uses the example "Dolphin Therapy". Thirty people aged 18–65 years with a diagnosis of depression were randomly assigned to either a treatment with common medical methods, or a treatment of a special dolphin therapy. At the end of the experiment, the researchers noticed that the proportion of patients with severe depression in the dolphin therapy group was smaller than the proportion of patients with severe depression in the control group. The question arises: Is it possible to infer that the dolphin therapy is a more effective treatment of depression than the common medical therapy? Or did the observed difference occur by chance?

The key concept, and the key question, is that of "chance variability". A big advantage according to Rossman (2008), and a convincing argument for a first step into informal inferential reasoning via randomization tests, is that "... this procedure for introducing introductory students to the reasoning process of statistical inference is that it makes clear the connection between the random assignment in the design of the study and the inference procedure" (p. 10). Some difficulties concerning the interpretation of p-values can be reduced by using randomization tests: "... [a randomization test] also helps to emphasize the interpretation of a p-value as the long term proportion of times that a result at least as extreme as in the actual data would have occurred by chance alone under the null model" (Rossman, 2008, p. 10). To make the argument clearer, it is easy to imagine that if the random assignment to groups is repeated many times, the p-value gives the relative proportion of the repetitions, where we get a value at least as extreme as the observed value. Cobb (2007) emphasizes that—by using randomization tests in introductory courses—students have a better opportunity to understand the "core logic of inference" (p. 11). There are no mathematical derivations from probability distributions to be done, which can lower the understanding. In more detail Cobb refers to the 3R's for a randomization test with software: Randomize data production; Repeat by simulation to see what's typical (and what's not); Reject any model that puts your data in its tail. Rossman (2008) provides examples of how such a randomization-based approach might be implemented at the secondary and tertiary level.

The use of randomization tests is also discussed by Edgington and Onghena (2007). They point out that, "... a randomization test is valid for any kind of sample, regardless of how the sample is selected. This is an extremely important property because the use of non random samples is common in experimentation, and parametric statistical tables ... are not valid for such samples" (p. 6). These authors do, however, point out that random assignment is a necessary condition for using randomization tests. However, these differentiations are partly controversial among statisticians. First, we will describe the "process-approach" by Konold (1994), and second, we will distinguish several scenarios that may arise when "making inferences."

Konold (1994) describes an approach where data is produced from a random assignment process (called "process-approach"). Within the process-approach (for details, see Konold, 1994; Konold & Pollatsek, 2002) data is seen as produced by a possibly hypothetical process.

Zieffler et al. (2011, p. 119) go into more detail and distinguish between four types/scenarios of "making inferences": Scenario 1 is characterized as a random sample and a non-random assignment, which they call "generalizable research". This may lead to conclusions about the population from which the sample was drawn but generally does not allow causal inferences. Scenario 2 describes a situation of a non-random sample and a random assignment to an experimental and a control group which is meant to be "randomized experimental research". In this case a generalization to a wider population is not possible but causal inferences related to the treatment in the experimental group. Scenario 3 refers to a random sample and a random assignment and is called "generalizable, randomized experimental research." Here are two types of possible inferences: Drawing conclusions about the population on the one hand and inferences on causality on the other hand. Finally, Scenario 4 covers a non-random sample and a non-random assignment which is understood as "non-generalizable, nonexperimental research", in this case conclusions have to be drawn very carefully. A method like bootstrapping would have been more adequate in the case of a random sample. This would mean sampling with replacement. (However, from a didactic point of view, resampling with replacement is less intuitive and more difficult to explain to students.) The idea of a randomization test procedure is that the random allocation of the variables can be imagined as a 1:1 mapping. In the special case of doing randomization tests with TinkerPlots™, this process can be made explicitly visible in the sampler.

Let us take a look at hypothesis tests in general. Which misconceptions occur when students perform and interpret a hypothesis test? Vallecillos (1994) reports that many students think similarly to a deductive process and do not appreciate the uncertainty in the reasoning process. Another typical misconception regarding the p-value is that the p-value is the probability that the null hypothesis is true, given the observed data (Garfield & Ben-Zvi, 2008, p. 270). Liu and Thompson (2009) conducted a study with eight high school teachers and point out that they had the interpretation: If the null hypothesis is rejected, than the statement must be false. Related to this, Harradine, Batanero, and Rossman (2011, p. 12) state that instructors need to "critical[ly] evaluat[e] ... the use of alternative methods (e.g., randomization tests) when first introducing statistical inference. Great care should be taken in this area given the widespread and long-term use of classical statistical inference".

Liu and Thompson (2009) propose to do randomization tests with TinkerPlots™. However, not much empirical research related to learning trajectories and learners' misconceptions while doing a randomization test has been published. Frischemeier and Biehler (2013) found that when analyzing written statistical projects which included a randomization test, preservice teachers seem to lack conceptual knowledge. For example, they can have problems generating an adequate null-hypothesis or drawing conclusions from a calculated p-value, while having good procedural knowledge regarding the use of TinkerPlots™ while doing randomization tests.

Based on this research, we developed a framework for clarifying the different steps that are necessary when conducting a randomization test. Adapted from this framework we elaborated a randomization test scheme with guidelines for our participants with the steps they are supposed to carry out for doing a simulation based randomization test with TinkerPlots™.

5.4 Subjects and Methods

Our first course takes students directly from group comparisons using descriptive statistics to randomization tests with a minimum of probability elements (e.g., Konold, 1994). In contrast, our second course routes students from probability modeling to randomization tests, similar to what is proposed in CATALST (Zieffler & Catalysts for Change, 2013). More specifically, the course began with students using simulation to compare models to real data and then transitioned to randomization tests using a minimum of elements of data analysis.

5.4.1 Topics of Course 1: Developing Statistical Reasoning Using the TinkerPlots™ Software

In this course the preservice teachers go through the entire PPDAC-cycle (Wild & Pfannkuch, 1999) which includes elements such as generating statistical questions and hypotheses, constructing a questionnaire, collecting data[1], analyzing data with the TinkerPlots™ software (Konold & Miller, 2011), and writing down findings in a statistical report. The preservice teachers learn how to write a statistical report, make group comparisons[2], discover and describe a relationship between two numerical variables, and make conclusions from a sample to a wider population in the form of informal inferential reasoning (IIR; for a definition, see Zieffler et al., 2008). At the end of the course, they, in teams of two, completed a statistical project with topics of their choice related to a data set concerning the leisure time activities of first semester preservice teachers. Making comparisons of distributions of numerical variables (i.e., group comparisons) was a fundamental goal of the course. In addition to describing and interpreting single distributions and exploring differences between them we wanted the preservice teachers to draw wider conclusions and generalize their findings. A typical task associated with group comparisons was: "Is there a real difference regarding the variable `Time_Reading` (time spent on reading books or magazines in hours per week) between the boys and the girls, or did that difference occur by chance?" At the end of the course we introduced the preservice teachers to

[1]The dataset primary used in this course was collected in a first semester mathematics course for preservice teachers and the self-created questionnaire involved items concerning their leisure time activities, intentions of becoming a math-teacher and their first impressions of studying at the University of Paderborn.

[2]The comparison between two or more distributions of a numerical variable.

randomization tests. Details about how randomization tests were introduced in the course are available in Frischemeier and Biehler (2013).

5.4.2 Topics of Course 2: Applied Stochastics—Understanding and Solving Complex Problems with Simulations

The course covers five topics: (1) Data analysis with TinkerPlots; (2) Basic simulations with TinkerPlots™; (3) Precision of simulations; (4) Independence and dependence; and (5) Hypothesis testing with p-values and randomization tests.These topics covered different time spans in the course. The main focus for the class was on the second and fifth topics. Each of these spanned four class sessions. Preservice teachers witness demonstrations of and short introductions to the software, and work in pairs on pre-designed learning trajectories. Preservice teachers learn to model different "real-world" situations using the TinkerPlots™ sampler, compute and display outcomes from random experiments and interpret the results. This process is supported with a "graphical simulation scheme", which can be used to plan, structure, or document a simulation with TinkerPlots™. We developed this tool, based on experiences we had teaching with Fathom® and adapted it for use with TinkerPlots™. The preservice teachers were required to put together a portfolio with some selected tasks and with reflections on every topic in order to support their understanding (Stratmann, Preußler, & Kerres, 2009). A major goal of this course was to have preservice teachers experience a more informal method of solving probability and statistics problems in addition to the formal computations that are prevalent in the culture of the German schools. With simulations, the preservice teachers were able to examine problems that could not have been solved formulaically at their mathematical level.

5.4.3 Study Participants

The study participants were preservice mathematics teachers at the primary and secondary school level and were either enrolled in Course 1 ($n = 12$) or Course 2 ($n = 24$). Each had previously completed the course, *Elementary Statistics and Probability Theory*, which does not cover inferential statistics. The preservice teachers all attended the university after their school day is finished, and did not have any prior practical experience in teaching.

Two months after the completion of each course, the study participants were interviewed in teams of two ($n = 6$ pairs in Course 1 and $n = 12$ pairs in Course 2). The interviews were designed to reveal the cognitive processes of learners conducting a randomization test. The participants were given the *VSE-task*, which will be described below. Although the participants were also given other tasks, the results of these tasks will not be reported in this chapter. The study, which was digitally recorded, was two-phased, with a working phase at the beginning and a "stimulated-recall" phase afterwards (see Busse & Borromeo Ferri, 2003). In the working phase, the preservice teachers worked independently on the VSE task and were required to communicate with each other (along the lines of "thinking aloud"; Bromme, 1981) without any input or interruption by the interviewer. This phase was videotaped with

the screen-capture software Camtasia. In the second phase, "stimulated-recall", the participants were shown the recording of their working phase and prompted by the interviewer to express their thoughts and aims at several stages. The interviewer interrupted the video during selected situations and posed questions such as, "what did you think at this moment?", "why did you do it that way?", "can you explain your intention on this aspect?", etc. in the form of basic, direct questions as suggested by Leiss (2007).

5.4.4 Data

The data collected included the randomization test schemes, exercise sheets, Camtasia recordings and TinkerPlots™ files from the working phase (phase 1) and the Camtasia recordings from the "stimulated-recall" phase.

The Randomization Test Task for the Participants. We wanted to observe in which way our preservice teachers could deal with a "typical" randomization test task consisting of a group comparison and a randomization test. In this section, our focus will be on the procedure of carrying out a randomization test. For our study, we took a dataset exported from the website of the German Bureau of Statistics[3] which contains 861 cases sampled at random from 60,552 interviewed German employees. The data included variables such as gender, wage per month, kind of employment agreement, and so forth. The task was to compare the monthly wages of males ($n = 477$) and females ($n = 384$). Figure 5.1 suggests that male employees tend to have higher monthly wages than female employees in the sample.

Is there a difference with respect to a variable between the two groups or could that difference have occurred at random due to the selection of our sample? This kind of question may lead us to a randomization test that can be carried out with software such as TinkerPlots™ (Konold & Miller, 2011).

The results of conducting a randomization test using TinkerPlots™ are shown in Figure 5.3. In the VSE sample dataset there is an observed difference of 833€ between the average income of women and men. The null hypothesis (that is to be rejected) is that there is no difference between the average income of women and men in the larger population. Using the results from the randomization test in TinkerPlots™, the estimated probability that the difference between women and men is 833€ or larger in a sample of 861 people, under the assumption that the null hypothesis is true, is 0.000.

In the sampler, the two variables "gender" and "salary" (from the original VSE data set) are represented as two separate devices (boxes). "Gender" is represented in a stack device that can take the values of "female" or "male", and "salary" is represented as a mixer with 861 cases labelled with the values of the employees'

[3]The `VSE_2006` dataset contains anonymous data for research and teaching, generated from the 2006 Earning Survey data. The 2006 Earning Survey was conducted as a stratified sample survey of nearly 28,700 companies with 10 or more employees. The companies employ around 1.8 million employees nationwide. It is available at https://www.destatis.de/DE/Publikationen/Thematisch/VerdienstеArbeitskosten/VerdiensteBerufe/VerdienstenachBerufe.html (retrieved on June 30, 2013).

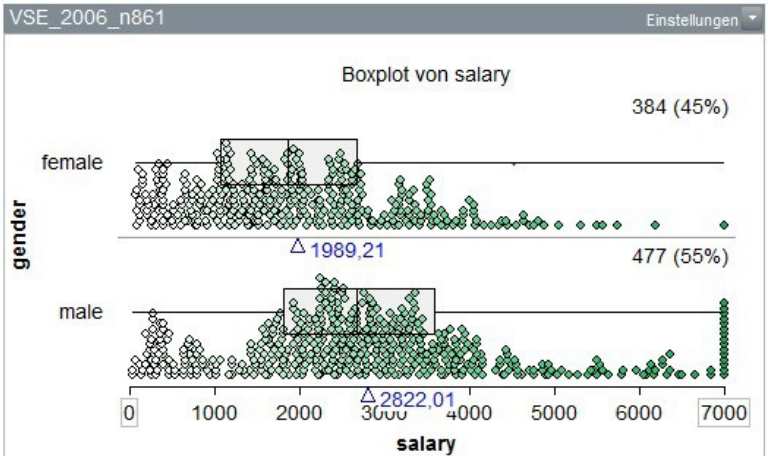


Figure 5.1. Group comparison for the VSE task.

In the dataset (see above) you can see the monthly salaries of 861 women and men of the year 2006. The display suggests that women are way behind men concerning their salary. Someone argues against the result of the group comparison between women and men that only 861 employees were asked. Therefore the differences could have emerged due to the selection of our sample.

YOUR TASK: Now check if there is evidence against the assumption that there is no difference between women and men in the population with regard to their average salary. (This would mean that we can expect similar differences for all employees)

Figure 5.2. Excerpt of VSE task handed out to the participants of the interview study.

monthly salaries. Note, that the variable "gender" is not uniformly distributed in this device (there are 477 males and 384 females in the sample). The null hypothesis is modeled by independently sampling (without replacement) a gender and salary from the two devices. In the results table (labelled "Results of Sampler 1") we see a subset of the results from a single sample. The difference of the average income for the two groups from this sample is computed and displayed in the stacked histograms of the "Results of Sampler 1". In this random draw, 'women' earn 212€ more than 'men'.

This random sampling and computation of the mean difference is carried out 1000 times. The difference in the means is collected and shown in the table in the bottom left-hand corner of Figure 5.3. These differences are displayed in the plot "History of Results of Sampler 1" in the bottom right-hand corner of Figure 5.3. This distribution represents the sample mean differences one would expect if there was no mean difference in the population. None of these simulated mean salary differences are as extreme as the observed difference of 833€ seen in the VSE data. So the p-value is 0.0000. What can we conclude from these results? Due to this very low p-value there

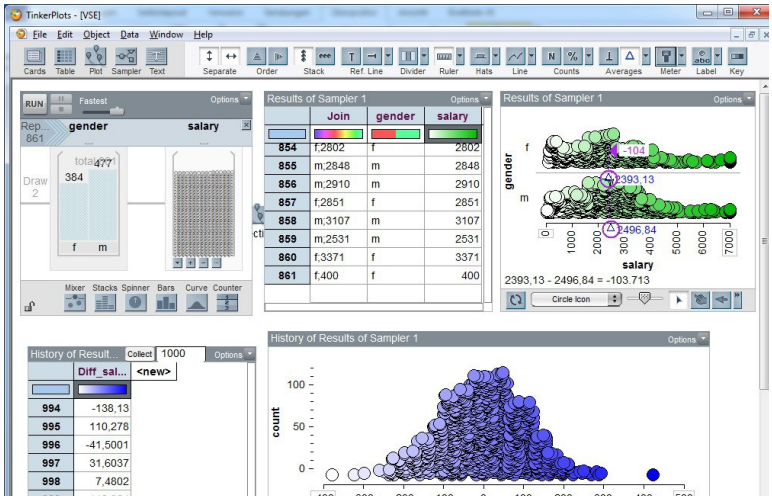


Figure 5.3. Screenshot from TinkerPlots™ displaying the results of a randomization test.

is strong evidence against the null hypothesis of no-difference-in-means of gender and salary in the population.

We note that the observed mean difference in salary of 833€ does not appear on the scale where the simulated values are plotted; the highest value from the simulation is around 350€. This is a difficult for the participants to understand. Although this seemingly produces a p-value of 0, it requires learners doing the VSE task to recognize that the actual p-value is not zero, but needs to be interpreted as a very small p-value.

Framework for Randomization Tests with TinkerPlots™. Biehler (1997, p. 175) describes a cycle for computer-supported statistical problem solving which consists of four phases: "Statistical problem", "problem for the software", "results of software", and "interpretation of results in statistics". We extended this model to a framework that describes the cycle of conducting a randomization test (and can also be seen as a cycle for conducting chance experiments via simulation in general) using software (see Figure 5.4).

Here we distinguish between three worlds: The contextual world, the statistical world, and the world of software, each of which is embedded within the other. For doing a randomization test using software (here using TinkerPlots™), six steps are needed, two from each world. The starting and ending points are in the world of context. The intermediate steps are located in the statistical world and also in the world of the software. These steps will be explained further and illustrated by the VSE task in italics.

Real problem: The starting point for the cycle is a task with a question. Normally this is provided in the task. The group difference in the VSE task is: *Men earn 833€*

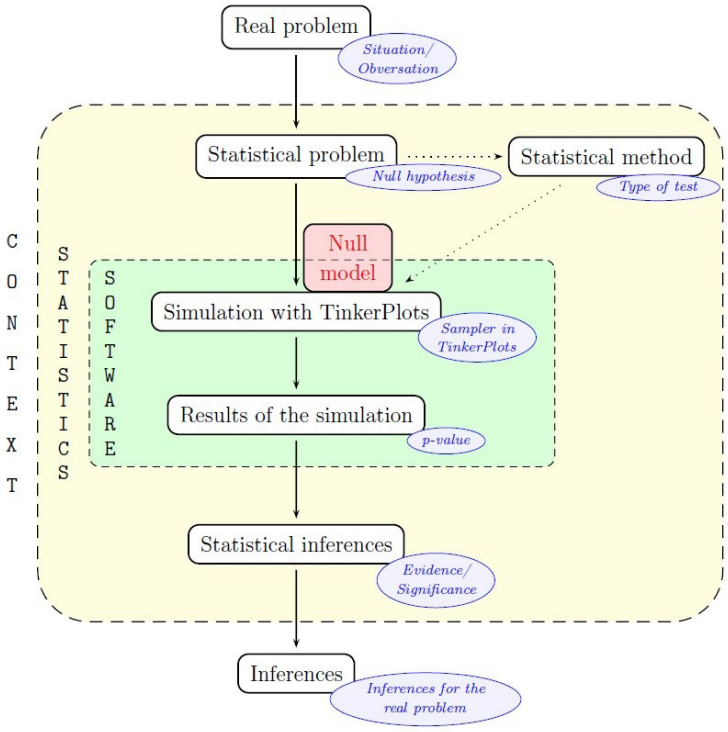


Figure 5.4. Framework for randomization testing.

more per month on average than women in the sample. What can you infer with regard to the population? In this situation the real problem is given in the task. Note that there might also be tasks which require the formulation of a real problem.

Statistical problem: At this point, when the null hypothesis is generated, we move from the "contextual world" (real world) to the "statistical world". It is notable that at this stage it is necessary to formulate an adequate null hypothesis, which can be "tested". For the VSE task an adequate null hypothesis might be: *There is no difference between the average salary of men and women in the population.*

Statistical method: After formulating an adequate null hypothesis, an appropriate statistical test has to be chosen. In this case, a randomization test would be appropriate.

Simulation with TinkerPlots™: As a basis for a simulation in TinkerPlots™, the previously constructed null model links the statistical and the software world. The null model of no difference can be modeled using the sampler in TinkerPlots™ (similar to that shown in Figure 5.5).

The first mixer contains 861 cases, 384 of which are women and 477 of which are men. This mimics the number of men and women in the sample data. The second

mixer contains one case for each of the 861 salaries of the sample data. A case is randomly selected from each of the two mixers, a gender and a salary. This process is repeated 861 times (without replacement) to produce a new, randomized sample. The difference in mean salary between males and females for this new sample drawn under the assumption that *gender is independent of salary* is computed and recorded. This entire re-randomization process is then repeated many times.

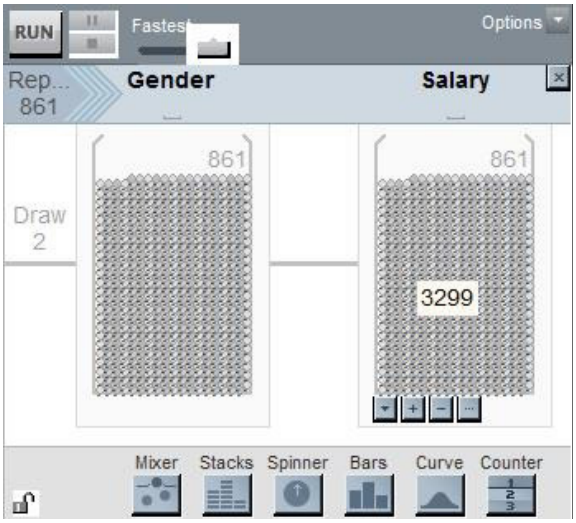


Figure 5.5. TinkerPlots™ sampler with two mixers.

Results of the simulation: In order for the results to be valid, the simulation in TinkerPlots™ has several technical steps that need to be followed. These steps will be explained later in the Results section. The technical steps mentioned above completely take place in the software world. At this stage, the results of the simulation have to be interpreted and documented. In the VSE task, a p-value would be computed to indicate *the probability of obtaining a result at least as extreme as the mean difference observed in the original data.* In our case, the p-value is less than 0.0001.

Statistical Inferences: When making statistical inferences, we re-enter the statistical world. The results produced by the software have to be transformed into statistical terms: Evidence, significance. For the VSE-task this is in short that *a p-value of less than 0.0001 leads to very strong evidence against the null hypothesis.*

Inferences: Finally we come back to the context world, where we interpret what the statistical inferences (done in the step above) imply about generalizations and conclusions we can make about the real problem. For example, in the VSE task, *we can make the conclusion that salary is not independent of gender in the population—there are mean differences in the salaries between females and males.*

Supporting Material for Conducting Randomization Tests with Software. A handout (see Figure 5.6) to support the preservice teachers in performing the com-

plex randomization test procedure was developed based on the proposed framework. It can also serve as a means of documentation of the several steps done by the participants.

Task:	
1. **Observation** What is the difference in the dataset?	
2. **Null hypothesis** Formulate the null hypothesis, which will be assumed as true.	
3. **Simulation of the null hypothesis** Describe the modeling process, modeled with the sampler of TinkerPlots.	
4. **Test statistic** Define the test statistic.	
5. **P-Value** Read off the *p-value*.	
6. **Conclusions** Draw conclusions concerning your null hypothesis with the *p-value*.	

Figure 5.6. Randomization test scheme.

As drawing adequate conclusions from p-values is difficult for learners, we also provided guidelines to support learners when using statistical evidence to draw conclusions (see Figure 5.7). These guidelines were given to preservice teachers in both courses.

5.5 Analysis and Results

The data analysis took place in two parts. In the first part, we analyze the randomization test schemes and the TinkerPlots™ files. In the second part, we take a look at chosen excerpts of the videos.

- We have weak evidence against the null hypothesis, if $p \leq 10\%$
- We have medium evidence against the null hypothesis, if $p \leq 5\%$
- We have strong evidence against the null hypothesis, if $p \leq 1\%$
- We have very strong evidence against the null hypothesis, if $p \leq 0.1\%$
- We have no evidence against the null hypothesis, if $p > 10\%$

Figure 5.7. Guidelines for supporting learners while drawing conclusions from p-values.

5.5.1 Analysis One: Worksheets and TinkerPlots™ Files with Regard to Statistical Steps

The data are initially analyzed with regard to the statistical steps taken to conduct a randomization test. These data comprise the preservice teachers' completed test schemes and TinkerPlots™ files. The written material (e.g., test schemes) can give us insight into the verbal skills and the understanding of the preservice teachers in a compressed form. The TinkerPlots™ files, which we analyze in a second step, offer insights into preservice teacher's TinkerPlots™ skills and knowledge, as well as their approach to a simulation task. Both data sources can reveal preservice teachers' difficulties in certain steps.

As an example of a filled out randomization test scheme, we present the case of Sara and Maggie (see Figure 5.8). This is an example of a sufficiently completed randomization test scheme.

For our analysis process we take into account the completed randomization test schemes and the TinkerPlots™ files. We distinguish between "statistical steps" (the six major steps necessary to perform the task; see Table 1) and "TinkerPlots™ steps" (the seven major steps we believe are crucial to conducting a randomization test with TinkerPlots™; see Table 4). According to our cycle (Figure 5.4) and the randomization test scheme (Figure 5.8) we base our analysis on six major "statistical" steps.

In the following paragraphs, we will describe "successful" performance for each of the "statistical" and "TinkerPlots™" steps. "Successful performance" is based on an expected solution at each step. We will also document typical mistakes that were seen in the data. Finally, we will summarize the results by recording the frequency of successfully accomplished steps and presenting the success rate[4] for each step. This will be presented for each course and as an overall measure.

[4]The success rate tells us the percentage of the participants who have accomplished the step.

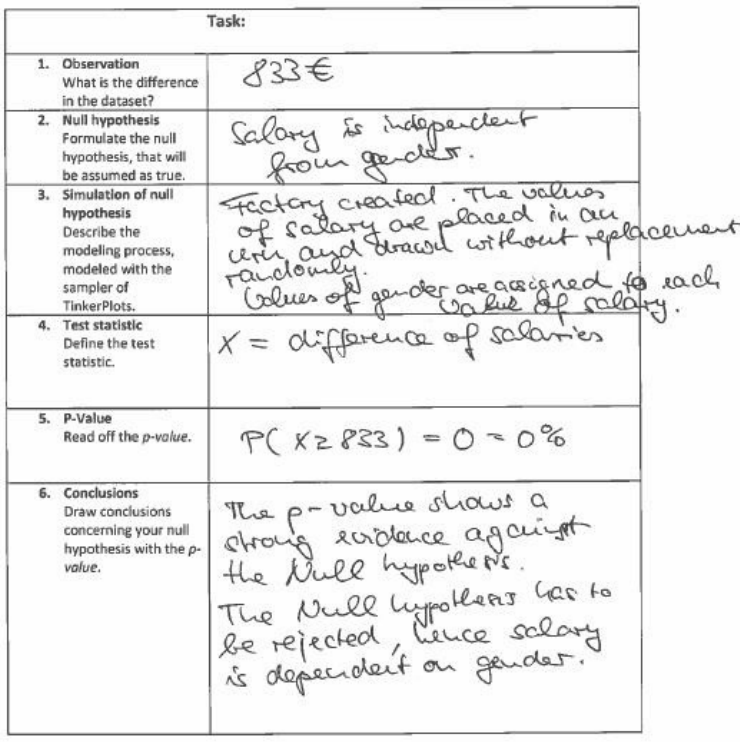


Figure 5.8. Randomization test scheme of Sara and Maggie.

Table 5.1
Six Major "Statistical" Steps when Conducting a Randomization Test

Step 1	Reading off the difference of the means of the groups in the dataset
Step 2	Formulating an adequate null hypothesis
Step 3	Describing the null model
Step 4	Formulating the test statistic
Step 5	Determining the p-value
Step 6	Drawing conclusions from the p-value

Step 1: Reading off the difference of the means of the groups in the dataset: In a first step the participants have to read off the difference between the conditional means, namely the mean difference between women's salaries and men's salaries. This idea is displayed in Figure 5.9. An expected solution at this step could be, "Women earn on average 833€ less than men. This difference is 29% of the male average." This value is the observed value of the test statistic employed in the hy-

pothesis test and should be included in the scheme. Successful performance on this step would be that the preservice teacher measured and recorded this value correctly in their scheme.

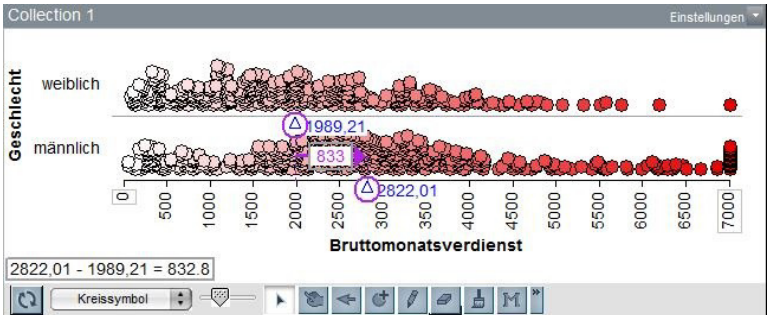


Figure 5.9. Reading off the difference of means between men (männlich) and women (weiblich) concerning the average monthly salary (Bruttomonatsverdienst).

Step 2: Formulating an adequate null hypothesis: The starting point of the randomization test is formulating an adequate null hypothesis, for example, "The average salary is independent of gender". Successful performance on this step would be that the preservice teacher generated an adequate null hypothesis.

Step 3: Describing the null model: The null hypothesis then has to be translated into a simulation model in TinkerPlots™. The simulation model needs to include a sampling device (e.g., mixer) for each attribute (gender and salary). The sampling from each device needs to be independent and without replacement. In this step, the preservice teachers should have prepared the simulation and described these aspects of the simulation model. We will judge the quality of the simulation model they actually built when we look at the TinkerPlots™ files. Here we judge the quality of the statistical description of the null model. In Table 5.2 we see some examples of null model descriptions distinguished by Course 1 and 2.

Table 5.2
Examples Regarding the Description of the Null Model

Course 1	**Course 2**
Urn with gender and salary. New assignment (thrown together).	Factory created. The values of salary are placed in an urn and drawn randomly without replacement. Values of gender are assigned to each value of salary.
861 pairs are drawn from two urns. Attribute 1: Gender and Attribute 2: Salary. These pairs are collected in a table.	Sampler. Two mixers: Salary and gender. Data placed in urn. Simulation without replacement.

Most participants did not provide a clear description of the null model. In Course 2 the "factory" metaphor was introduced and used very often, but only one pair used this metaphor in their description (see Table 5.2, top right).

We did not rate the descriptions as successful or not, because this was difficult to decide. Too many different formulations and use of vocabulary made it impossible to categorize the answers. We will have a look in chosen transcripts in the sense of reconstructing the cognitive process of the participants. It is clear that more emphasis has to be put on this step in a revised teaching sequence.

Step 4: Formulating the test statistic: In this step the participants were supposed to describe the test statistic in statistical terms, or at least, in their own words. The test statistic can be described as an expected solution for "the difference between the average monthly salary of men and women". In a solution, to be rated correct, the words "difference", and "averages of salary" must have occurred in a meaningful combination. Only two phrases were rated as correct; the remainder were rated as incomplete or wrong. In Table 5.3 we present test statistics rated as "correct", "incomplete", or "incorrect" for each of the courses.

Table 5.3
Written Statements Regarding the Formulation of the Test Statistic

Course	**Correct (3x)**	**Incomplete (9x)**	**Incorrect (6x)**
Course 1	Difference of averages of salary	Differences of salary (2x)	Salary
	Difference of average salary of men and women (2x)	Differences of salary of men and women (5x)	Difference is 833€ or more
Course 2		Difference of averages (2x)	Mean of salary
			No description (3x)

We rated "Difference of averages of salary" (see Table 5.3, "Correct") as correct, because there the component "difference" and the idea of the average of the variable "salary" was included, whereas "Differences of salary" and "differences of averages" (Table 5.3, "Incomplete") would be rated as incomplete because it lacks either "average" or "salary". Single expressions like "salary" or "mean of salary" were coded as incorrect.

All in all, we have only three correctly formulated test statistics, all from preservice teachers enrolled in Course 2. The verbal formulation of the test statistic in statistical terms was a problem for almost all of the participants: Fifteen of the eighteen descriptions were incomplete or incorrect.

Step 5: Determining the p-value: The p-value, the relative frequency that the difference in the average salary between women and men generated by the simulation is equal to 832.80€ or larger than 832.80€, is approximately 0% ($p < 0.0001$). An expected solution might be, "The p-value is approx. 0 % here. This is the probability

of getting a value as extreme or more extreme than the observed one of 833€, under assumption that the null hypothesis is true." An additional verbalization like that in the second sentence is desired, but not necessary to be rated as correct. None of the preservice teachers offered an additional verbalization of the p-value in any form.

Step 6a: Drawing correct statistical conclusions: A p-value of 0% should lead to very strong doubts against the null hypothesis. The observed value 833€ is very extreme compared to the simulated distribution. A solution could be: "A p-value smaller than 0.001 shows very strong evidence against the null hypothesis. There are two possibilities: either the null hypothesis is true, and something very uncommon has happened, or the null hypothesis is not true and should be rejected." Successful performance on this step would be assigned if the p-value is interpreted correctly, either with the "evidence" terminology or with the interpretation given in the second sentence.

Step 6b: Drawing correct contextual conclusions: Did the preservice teachers connect the context of the task with the interpretation of the p-value? To reject the null hypothesis, a difference must be established between the average salary for men and women. An expected solution might be, "With a p-value of approximately 0%, we can argue against the null hypothesis and infer that there is an effect of gender, in which case, men earn, on average, more than women." Successful performance on this step would be if they refer to the task in the sense of connecting interpretation with context.

Overview of Results. Figure 5.10 displays the proportion of successfully accomplished steps overall (left) and separated by course (right). Steps 3 and 4, where the statistical issues should be formulated, could not be rated as correct or incorrect and therefore do not appear in Figure 5.10.

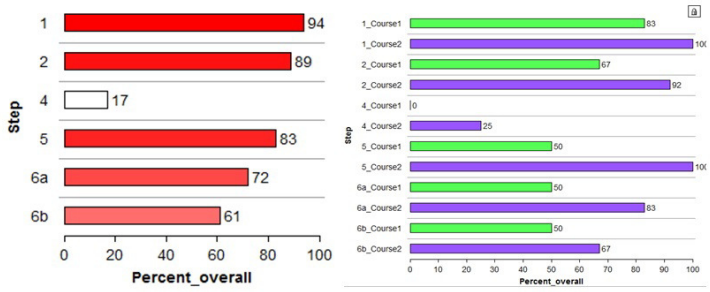


Figure 5.10. Success rates overall and success rates according to courses for Step 1, 2, 4, 5, 6a, and 6b.

In Course 2, the procedure of conducting a hypothesis test was taught explicitly, so the better results are not surprising. At least half of the preservice teachers in Course 1 were able to conduct a randomization test. One difficulty of the task, mentioned previously, was that the p-value obtained from the simulation was very close to zero. The participants from Course 1 never saw such an extreme p-value in their

coursework, which may explain the lower success rate for Steps 5–6b. In general, the reference to context was also a problem for the participants. This is where the transition from the "statistical world" to the "contextual world" takes place in the software cycle (see Figure 5.4).

Overview of Mistakes that Occurred in the Statistical Steps. *At Step 2:* The error was to formulate the alternative hypothesis as the null hypothesis. "Null hypotheses" we rated as "non-adequate" were: "Men earn more than women on average," which occurred twice out of 18 pairs. One pair did not generate a null hypothesis at all.

At Step 5: Those who were not rated as correct at this step were not able to find the p-value at all. Our interpretation is that this was due to the extreme observed value, which did not appear in the graph of the simulated distribution. Those who could not solve this step came from Course 1, where they had less experience with p-values and never saw a p-value near zero.

At Step 6a: Why is the success rate of Course 1 so low? One reason might be that in Course 1, terminology related to evidence was not commonly used. (The "evidence" language was only used in Course 1 one time.). In Course 2, the participants were trained to use the "evidence" terminology for making inferences about the null hypothesis.

5.5.2 Analysis One: Worksheets and TinkerPlots™ Files

Let us now have a look at the steps which have to be done in TinkerPlots™ (partly parallel to the statistical Steps 3, 4, and 5) when conducting the simulation of the null model (Table 5.4).

Table 5.4
TinkerPlots™ (TP) Steps

Step	Description
TP1	Populating the mixers with the correct labels/values to mimic the original sample.
TP2	Setting the number of repetitions (how many cases should be randomly selected from each mixer) to the original sample size.
TP3	Setting the number of repetitions (how many cases should be randomly selected from each mixer) to the original sample size.
TP4	Plotting the new, randomized sample and depicting the measure of deviation from the null hypothesis (e.g., mean difference) in the plot.
TP5	Collecting the chosen measure from many different re-randomizations using "Collect Statistic" and the history function.
TP6	Plotting the collected statistics to examine the distribution of the "test statistic".
TP7	Computing the p-value

Steps TP1–TP7: Modeling the simulation in TinkerPlots™—summary of technical aspects: We identified crucial steps for the simulation process and will describe them below.

Step TP1: Populating the mixers with the correct labels/values to mimic the original sample: Mixers can be used as devices (see Figure 5.11) or stacks or a combination of both, because they are the only devices with the option to sample "without replacement". The devices have to be filled with the values of the two attributes "Gender" and "Salary" of the dataset. It is not necessary to rename the attributes as in Figure 5.11. Successful performance on this step is to complete the settings correctly.

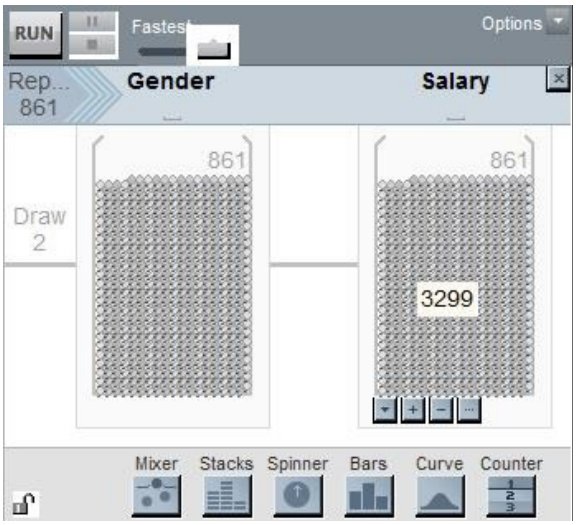


Figure 5.11. Example of a sampler to model the VSE task.

Step TP2: Setting the number of repetitions (how many cases should be randomly selected from each mixer) to the original sample size: For a correct simulation of a randomization test the sample size has to be 861. There must be as much repetitions as there is data in the sampler. The original dataset consists of 861 cases, so in this case, 861 is the number that has to be chosen for the number of repetitions. Successful performance on this step is to do the simulation with 861 repetitions.

Step TP3: Set the number of repetitions (how many cases should be randomly selected from each mixer) to the original sample size: A fundamental aspect in doing a randomization test is drawing without replacement when simulating the experiment. This option has to be adjusted for both devices. Successful performance on this step is assigned if drawing was done without replacement.

Step TP4: Plotting the new, randomized sample and depicting the measure of deviation from the null hypothesis (e.g., mean difference) in the plot: The distribution of the results of one run (that is, 861 repetitions of the three draws) of the sampler has to be displayed in a plot. A crucial further step is the identification of the av-

erage salary of women and men. Successful performance on this step is the correct calculation (e.g. with the ruler) of the difference between the means.

Step TP5: Collecting the chosen measure from many different re-randomizations using "collect statistic" and the history function: Successful performance on this step consists in collecting the differences of means as "history".

Step TP6: Plotting the collected statistics to examine the distribution of the "test statistic": The collected values lead to the distribution of the test statistic. This distribution has to be shown in a plot in a way that allows further analysis. For displaying the bell-shaped distribution of collected measures the plot should be separated completely (see Figure 5.12). This shows that the observed value of 833€ does not appear in the simulated distribution. Successful performance on this step is assigned if this has been done.

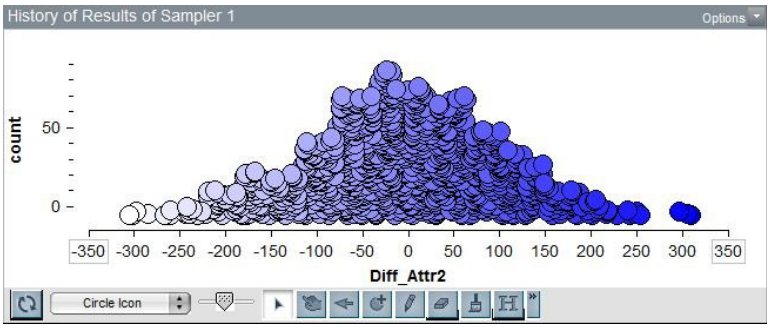


Figure 5.12. Example of a simulated distribution based on the null hypothesis.

Step TP7: Computing the p-value: A crucial point is to identify the p-value in the distribution of the test statistic (see Step 4). The distribution ranges from about -350 to $+350$, so the observed value of 833 (Category 1) does not appear on the scale. The p-value is the probability of obtaining the observed value of 832 or more extreme (this means larger) values, under the assumption that the null hypothesis is true. It should be clear that this probability is very near zero, because there is no result as extreme as that. This can be concluded from the display without any use of TinkerPlots™ functions. However, many pairs used the TinkerPlots™' divider tool. Successful performance on this step is to calculate the p-value with the divider in TinkerPlots™ correctly. Figure 5.13 displays the overall success rates of accomplishing TinkerPlots™ steps.

Even if Step TP1 was done incorrectly with an equal distribution for gender, we considered this not as crucial in regard to the evaluation of the following steps—nonetheless, the step TP1 was coded as "not successful" if it was done incorrectly with an equal distribution of gender. This mistake does not have an effect on the coding of the succeeding steps. This is the same for step TP3 (it should be drawn without replacement). If this was not done, it indicates the need to do a bootstrap test: We did not talk about this in class, but it could also be a solution. Nonetheless, the

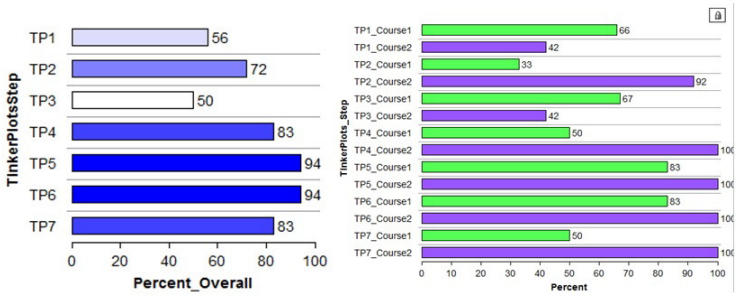


Figure 5.13. Overall success rates of TinkerPlots™ steps.

step TP3 was coded as "not successful" if it was done incorrectly with replacement. This mistake does not have an effect on the coding of the upcoming steps.

Having a look at Figure 5.13, we conclude that the participants have very few problems conducting a simulation of a randomization test in TinkerPlots™. Overall, we can conclude that the steps were done satisfactorily. More than half of the pairs did a good job at conducting a randomization test with TinkerPlots™ and at interpreting the results. We rated their performance "good" if the statistical steps 1, 2 and 6 (6a or 6b) and the TinkerPlots™ steps TP2, TP4, TP5, TP6, and TP7 (see above) are correctly done.

Mistakes and Difficulties in the TP Steps. *Step TP1:* As can be seen in Figure 5.13, Step TP1 had the lowest success rate. One difficulty in Step TP1 was that many participants selected the proportion of women and men used in the sampling device to be equal, rather than basing it on the proportions in the original sample. This was done by 44% of the studied participants. Of these, 50% used a mixer with only two cases, one labeled "male" and another "female", and 50% used a spinner device with an equal distribution for gender (Figure 14a and 14b).

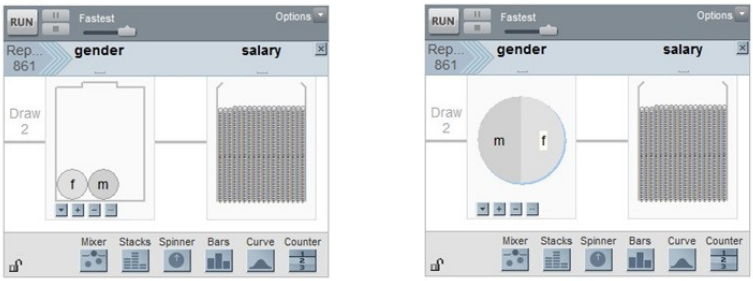


Figure 5.14. Equal distribution for gender with either a mixer or a spinner.

Step TP3: Another common difficulty, which occurred at Step TP3, was that the cases of the mixer were drawn with replacement—the default setting of the Tinker-

Plots™ sampler. All 44% who chose the equal distribution of cases in step TP1 drew with replacement. If the spinner device was chosen, this problem is exacerbated, as spinners always sample with replacement. With regard to statistical understanding this aspect is important since we are conducting a randomization test (which has to be done without replacement) and not a bootstrap test (which is meant to be done with replacement).

A fundamental "contradiction" which arises is that all the TinkerPlots™ steps and most of the statistical steps (1, 2, 5, and 6) were correctly done, but the statistical steps 3 (describing the null model) and 4 (describing the test statistics) were poorly articulated in the randomization test schemes.

5.5.3 Analysis Two: Re-Analysis of Selected Steps on the Basis of the Video Data

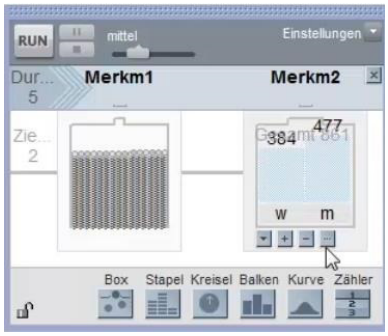
Although informative, the "level one" analysis does not give enough insight into the cognitive processes of the preservice teachers, specifically those used while setting up the null model. To further understand those processes, we examine the video data. We use selected transcriptions and base our analysis of the written material by means of "crucial episodes" in the sense of Voigt (1984). The selection of these crucial episodes is subjective. We chose episodes that reveal interesting insights into the cognitive processes. Here, we examine two pairs of preservice teachers, Sara and Maggie (Course 2) and Conrad and Maria (Course 1) as they carry out a randomization test. We focus on the episodes related to the construction of the null model.

The pair of participants from Course 2, Sara and Maggie, worked together during the entire course. They were also homework partners. Based on their contributions in class and the evaluation of their homework, we believe they have a good understanding of the theory and also that they can successfully use TinkerPlots™. During class they often discussed their work. Sara and Maggie are the only pair of participants that used the "factory" metaphor of the sampler in their description of the null model. Because of these observations during the course, we chose to study Sara and Maggie as participants who would exhibit exemplary responses.

Conrad and Maria, the participants from Course 1, were chosen because they represent the more "typical" pair of participants in the study. Similar to Sara and Maggie, we observed many positive interactions between Conrad and Maria in class. In contrast to Sara and Maggie, however, we did not judge the pair to have the same amount of statistical understanding—Conrad seemed to have a better understanding than Maria, and also did most of the TinkerPlots™ work in class. Maria seemed a bit shy, which might explain some of the differences we observed between her and Conrad.

Sara and Maggie: The Null Model. Prior to the point at which this transcript begins, Sara and Maggie read off the difference between the means of the two groups in the dataset and have generated an adequate null hypothesis ("salary is independent from gender", see Figure 5.8). Now they begin discussing how to fill the sampler in TinkerPlots™.

72	Sara:	No I mean for the simulation in a moment.
73	Maggie:	Do it as you may suppose.
74	Sara:	For the simulation, we have to paste all salaries which are there.
75	Maggie:	Aha!
76	Sara:	No "Zufall" [random], is it right there?
	[TinkerPlots™]	They drag a sampler [German label: random].
77	Interviewer:	Yes.
78	Sara:	Ok. Then all salaries there, how did it work?
79	Maggie:	Yes right. You have to copy, wait I'll try it.
80	Sara:	And then, oh wait, we have to delete this [they delete the balls in the sampler]. Ok, again actually with men and women or in the second?
81	Maggie:	Oh you want to have a second factory, ok.
82	Sara:	Ok, then we take. Or do you want to have another?
83	Maggie:	Do it. Now again, do you want only the percentages?
84	Sara:	Oh yes it's right. We need eh. But there are as many men as women, aren't there?
85	Maggie:	Look here, you can't do this, no I think not. We can try it, because otherwise we can't use the mixer. [. . .]
95	Maggie:	Yes. I fill in the scheme.
96	Sara:	Eh. What's this?
97	Maggie:	It's ok, what have we done there?
98	Sara:	Appropriate a.
99	Maggie:	Factory.
100	Sara:	Yes. (laughing)
	[TinkerPlots™]	The students constructed this sampler.



101	Maggie:	So a factory.

In this excerpt we get to know some of Sara and Maggie's thoughts as they are setting up the sampler. They describe what has to be done in several steps: "We have to paste all salaries which are there" (Line 74); "You have to copy" (Line 79). In Line 81 they describe the device of the sampler as a factory. This is an important observation, since this describes the meaning of the null model metaphorically and exemplifies Konold et al.'s (2007) idea of "understanding distributions through modeling them" using the factory metaphor. The "factory" metaphor was introduced in Course 2 and was used during most course activities. These two participants used the "factory" metaphor for the data-production process in TinkerPlots™. Not only this, but they could articulate and describe this process as suggested by the video transcript and in their written files. They also documented this idea in the randomization test scheme, as seen in Figure 5.15.

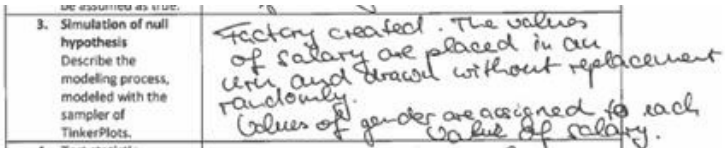


Figure 5.15. Excerpt of Sara and Maggie's randomization test scheme.

We interpret that Sara and Maggie seem to see the sampler as a kind of (data) factory, which produces data under certain conditions and having certain properties. This pair demonstrates a good understanding of the underlying process for setting up and populating the sampler. Their language reveals they have made the connection between the statistical world and the software world. As we mentioned previously, most participants in our study do not make this connection. All in all, Sara and Maggie solved the VSE task very well.

Conrad and Maria: The Null Hypothesis and the Null Model. Conrad and Maria also read off the differences of means between the two groups in the VSE dataset. The first part of the transcript begins as the pair discuss the generation of the null hypothesis.

09	Conrad:	I don't know, if the H-null hypothesis—is it then, that so, that women earn less or is it called that it's refuted—so the other way round?
10	Maria:	I don't know, no
11	Conrad:	Women earn on average less than men
12	Maria:	It doesn't matter, right?
13	Conrad:	Well
14	Maria:	We write it soon (U)
	[TinkerPlots™]	On the randomization test scheme they write down the following null hypothesis:

Women earn on average less than men.

This excerpt can be interpreted as a typical discussion the participants engage in at the often choose a "wrong" hypothesis, such as, "women earn on average less than men" (see Figure 5.15). Conrad and Maria articulated what they expected for the difference of average of salaries in the population, the research hypothesis, rather than the null hypothesis. Evidence from the other participants' test schemes suggest that this is a typical problem for participants when setting up a null hypothesis—the inability to distinguish (or confusion between) between the null hypothesis and the research hypothesis.

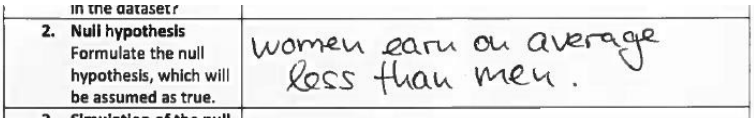
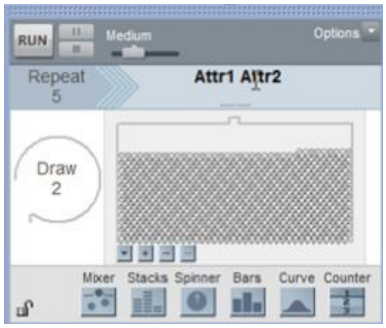


Figure 5.16. Excerpt of randomization test scheme of Conrad and Maria after generating the null hypothesis and before setting up the null model.

The next excerpt from Conrad and Maria's transcript gives us insight into how they constructed their model in TinkerPlots™ (null model).

15	Conrad:	Yes, we want to create two urns now, where on the one side—well, that we can allocate the genders, right?! That it is independent of the gross earnings, i.e., on one side the quantity of the gender and on the other side, no on one side all salaries and on the other side how many men and women, the parts. And then it's thrown together.
16	Maria:	Right.
17	Conrad:	That the allocation of the gross monthly earnings is independent of the gender, that is just like that
18	Maria:	Yes that's what we want



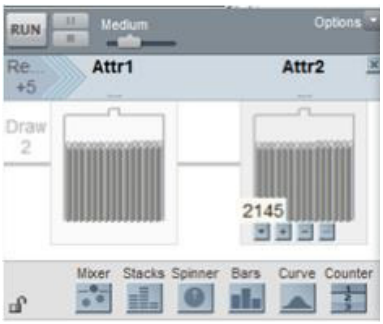
[TinkerPlots™]

19	Conrad:	*(Conrad and Maria laughing)*

Yeah . . . now we need a second urn (U) like—what do we have to click that we get a second urn, a second box? . . . Do you know what I mean?

20 Maria: Yes, I know what you mean.

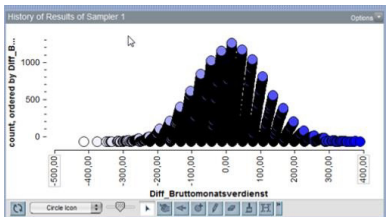
Without speaking they copy the values of the second attribute (salary) in a second mixer.



In contrast to their null hypothesis, they construct the sampler using independent sampling devices (mixers; see excerpt of transcript). We assume that they had the correct conception in mind and knew that they have to reject their null model, which assumes that salary and gender are independent. After setting up their null model, Conrad and Maria carried out the remaining steps in TinkerPlots™ correctly. While they had probably internalized a "scheme" consisting of the required steps for carrying out a randomization test, their understanding or connection to the statistical steps was missing here. We also observed this phenomenon in the classroom during both courses.

After running the simulation, Conrad and Maria then looked for conclusions. At this stage the transcript continues:

[TinkerPlots™] *Conrad and Maria fill the devices of the sampler correctly. The are drawing 861 times without replacement and plot the results of the sampler. Then they calculate the difference of means of both distributions and collect 5000 "measures" via the history function. Finally, they plot the collected measures (as seen in the plot below)*



After collecting the measures there have doubts whether the null hypothesis is correct:

45 Conrad: No (negating), and now we have to . . . now we should this—what is it called? So this, what measured this, so, what percentage it is . . . about

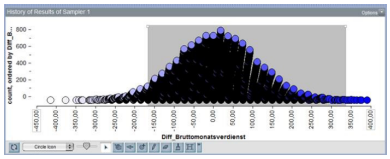
52 Conrad: That couldn't be, right? Then it would be 0% . . . yes, you know—we've referenced to the old thing, right?!

53 Conrad: (meanwhile) Yes, I know what you mean

54 Conrad: And then here once, we should go into minus and even into plus, the [. . .] but— . . . or had we done anything wrong? With this 832 or is it—

56 Conrad: Yes, it would be so, then we register it, I don't know differently, no, then . . . it is, there are 0%

[TinkerPlots™]



They create a divider.

68 Conrad: In other words, we

69 Maria: Don't have a difference

From the transcript, it is clear that Conrad is the discussion leader at this stage. Not surprisingly, he also seems to have a better understanding of the whole process than Maria. After looking at their collected history of the differences between the average salary of men and women, they start the discussion about the null hypothesis.

70 Conrad: Well, right. And then (. . .) we got our—I think, our null hypothesis is wrong, right? Because we should reject it, because it doesn't fit . . . because we should stay under 5% . . . and then we should reject the null hypothesis

71 Maria: (interrupts C) Additionally we add a "not" (They add "not" in Step 2 of their randomization test scheme) *On the randomization test scheme they register finally the following null hypothesis: Women earn on average not less than men.*

72 Conrad: Well, our null hypothesis is that the women do not earn less than men on average, we are below 5% and and therefore the null hypothesis is rejected and so we can say that women earn less than men on average.

They interpret the distribution and the p-value correctly in the sense of the task, albeit not for their null hypothesis. This leads them to some cognitive conflict regarding their initial null hypothesis. Because they were convinced that they carried out the simulation correctly, they changed their written null hypothesis by adding the word "not" to their randomization test scheme (see Figure 5.16). This confirmed our interpretation that they initially had in mind that the assumption, "there is no difference in the average salary of men and women" was supposed to be rejected. That is probably why they stated, "women earn on average less than men" in their hypothesis. They seem to know they should reject the null hypothesis with a small p-value and these are the reasons why their null hypothesis must be wrong.

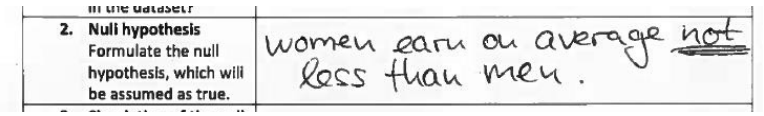


Figure 5.17. Excerpt of randomization test scheme of Conrad and Maria—final stage.

The addition of the "not" to the null hypothesis, which happened almost at the end of their work, makes the whole randomization test process finally correct. While Conrad and Maria did not correctly write a null hypothesis at the beginning of the randomization test scheme, and did not immediately connect it with the simulation, they, at least, noticed the discrepancy between their initial writing and their interpretation of the simulation results. Conrad reveals his understanding of this discrepancy as he immediately realized that he had to change the null hypothesis. Moreover, he recognized that he should not change the simulation nor think that he misinterpreted the distribution and p-value. This is an example of learning during/from the use of software, and shows a connection between the statistical and software world.

We conclude that Conrad and Maria solved the task correctly. They have, like others in our courses, a good understanding of the technical aspects of a simulation, and in particular of those aspects of the simulation of a randomization test. However, they only showed a limited understanding of the statistical background of a hypothesis test. Further analysis, not reported in this chapter, suggests that this result may be generalized to many participants from Course 1, but only for some participants from Course 2. The extended discussions and realization of several hypothesis tests with large and small p-values over a period of four weeks in Course 2 certainly helped to deepen the understanding of participants taking the second course.

5.6 Discussion and Implications

In contrast to Rossman (2008), we do not start informal statistical inferences with randomization tests. In both courses (data-centered or probability-centered), we teach randomization tests at the end of our courses. The reason for this is that we want preservice teachers to be proficient using the software for simulating chance exper-

iments before being exposed to the complex topic of randomization tests. Also, we want preservice teachers to gain experience in making informal inferences before making formal inferences either in the context of probability models or that of data analysis. Because of these learning trajectories, randomization tests are introduced at the end of our courses.

5.6.1 Summary of Some Findings

We have distinguished two domains of knowledge, the statistical knowledge needed to fully understand the randomization test procedure, and the software knowledge needed carry out the randomization test simulation using TinkerPlots™. Many preservice teachers in our study were able to conduct the majority of steps to carry out a randomization test using TinkerPlots™. This is a pleasing outcome since the participants had not previously been exposed to randomization tests and given the limited amount of time in the courses to teach these methods.

The results of this study also suggested that the participants had gaps in the statistical knowledge underlying the randomization test. For example, participants had a difficult time generating an adequate null hypothesis, setting up the null model, and interpreting a p-value. It is, perhaps, not surprising that these struggles have been previously documented in the literature (e.g., Garfield & Ben-Zvi, 2008; Vallecillos, 1994). These findings have implications for the re-design of the learning trajectories for both courses, and we will address this at the end of this section. Before this, however, we will summarize what we found concerning our research questions.

How well are the preservice teachers able to model a randomization test experiment with TinkerPlots™? What is the role of TinkerPlots™ in their thinking? The data suggests that the participants in Course 2 (which emphasizes probability and only covers minimal data analysis) have more statistical knowledge. This provides some evidence that a course where simulations of chance experiments and hypothesis testing are taught explicitly before entering a learning trajectory to randomization tests might lead to a better understanding of randomization tests as compared to a course where randomization tests are immediately preceded by data analysis.

The technical features of TinkerPlots™ do not seem to be problematic for the preservice teachers. Problems only occur at the interface of the software and statistical world (TP Steps 1 and 3). When TP Step 1 was performed incorrectly, it was because participants populated the sampler using an equal proportion of women and men instead of reproducing the sample proportions. When TP Step 3 was performed incorrectly, participants' mistake was in sampling "with replacement." The data from Conrad and Maria's transcript suggests that TinkerPlots™ can help support students' reasoning, at least in the sense of refining their null hypothesis see also delMas et al. (2013). The crucial point seems to be the transition between the statistical and the software level (Figure 5.18), particularly the construction of the null model.

In which way do the preservice teachers accomplish the steps of a randomization test? How do the preservice teachers interpret the results of the randomization test? Most of our preservice teachers were able to conduct the steps

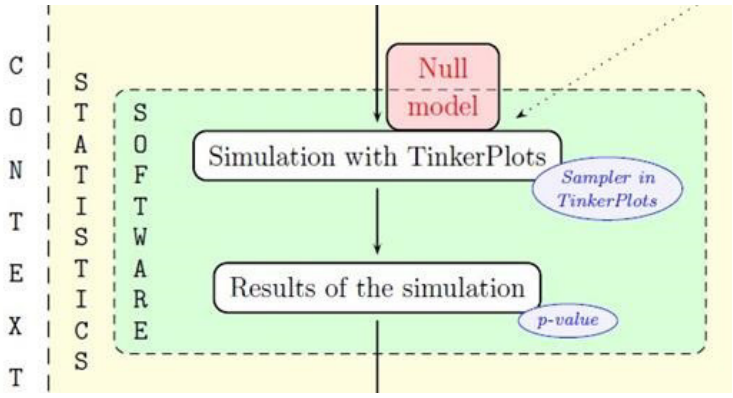


Figure 5.18. Excerpt from "software cycle when conducting chance experiments".

of a randomization test with TinkerPlots™ when supported by the randomization test scheme. Some of the participants struggle at typical crucial points, like not being able to formulate an adequate hypothesis (similar to results of the study of Liu and Thompson (2009). We also observed common difficulties when interpreting p-values, as Garfield and Ben-Zvi (2008) reported, for example.

Despite these difficulties and gaps in their statistical knowledge, the preservice teachers were able to make inferences about group comparisons. But, doubts remain about how deep their conceptual understanding of hypothesis testing actually is. Since we see (c.f. Figure 5.13) that the participants of Course 2 performed better in several steps and in several aspects, the approach with minimal data analysis and an emphasized probability component might be better suited for learners approaching the randomization test method. This conclusion is not more than a suggestion as we did not do a randomized comparative experiment. Furthermore since participants taking Course 2 were more experienced in simulating chance experiments and making conclusions from given p-values, this might suggest the need to simulate several chance experiments before introducing hypothesis testing—randomization tests in particular. For getting a better understanding of the randomization process itself, it might be helpful to add a hands-on activity such as that proposed by Arnold, Budgett, and Pfannkuch (2013).

A further important finding is that courses need to put more emphasis on relating the statistical and the contextual world (see Figure 5.4). Since we identify typical mistakes such as the false reproduction of the sample (in the sense of a number of draws unequal to the number of cases in the sample) and not drawing without replacement (as it is necessary, when doing a randomization test), we recommend that the null model should be discussed in detail before simulating a chance experiment with software. This might include a discussion of suitable null models for different situations. One specific redesign of the learning trajectory might be trying to improve connections between generating the null hypothesis and conducting a TinkerPlots™ simulation by explicitly discussing the construction of the null model. It might also

be helpful to formulate the results of the simulation in both "everyday" language and using statistical language.

Furthermore, students need to learn how to interpret the p-value with regard to the "real-world" situation. A statistically correct answer, like, "I reject the null hypothesis because the p-value is smaller than 0.001", cannot be rated as a satisfactory answer, because the connection to the context (the gender difference in salaries) is missing. There is a need to discuss the meaning of what a small p-value indicates about the real problem, in this case, gender difference in salaries.

Our findings have made us aware of some limitations of the study. The VSE data exhibit such a large gender difference in the variable salary that the p-value in the randomization test is calculated as 0 and does not evoke a discussion about uncertainty related to p-values.

A major further implication that arises for us is that the support of learners with regard to the structural aspects in form of a randomization test scheme is crucial. Observations in the video-study and in the two courses made it evident for us that the participants made substantial use of the randomization test scheme when structuring their activities. Nonetheless, there are aspects and misconceptions that cannot be addressed by the use of a randomization test scheme. The development of conceptual knowledge of the participants has to be improved on "generating an adequate and testable null hypothesis" and "drawing conclusions from a given p-value", so revising the global learning trajectories of Course 1 and Course 2—less than Course 1, is an implication. Discussions about inferences of small and very small p-values, but also about possible inferences when large p-values occur should be implemented in the teaching. Also the diagram "framework for randomization testing" (see Figure 5.4) might be useful as an explicit tool in the teaching process.

References

Arnold, P., Budgett, S., & Pfannkuch, M. (2013). Experiment-to-causation inference: The emergence of new considerations regarding uncertainty. In A. Zieffler & E. Fry (Eds.), *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 119–146). Two Harbors, MN: University of Minnesota.

Biehler, R. (1997). Students' difficulties in practising computer supported data analysis—some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169–190). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R., & Maxara, C. (2007). Integration von stochastischer simulation in den stochastikunterricht mit hilfe von werkzeugsoftware. *Der Mathematikunterricht*, *53*(3), 45–62.

Bromme, R. (1981). *Das denken von lehrern bei der unterrichtsvorbereitung. eine empirische untersuchung zu kognitiven prozessen von mathematiklehrern*. Weinheim, Basel: Beltz.

Busse, A., & Borromeo Ferri, R. (2003). Methodological reflections on a three-step-design combining observation, stimulated recall and interview. *ZDM—The International Journal on Mathematics Education*, *35*(6), 257–264. doi: 10.1007/BF02656690

Cobb, G. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1), 1–15. Retrieved from http://escholarship.org/uc/item/6hb3k0nz

delMas, R., Zieffler, A., & Brown, E. (2013). Students' emerging reasoning with uncertainty in a randomization-based first course in statistics at the tertiary level. In *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking and Literacy (SRTL-8)* (p. 147-164).

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman and Hall/CRC.

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, *19*, 676–685. doi: 10.1214/088342304000000396

Frischemeier, D., & Biehler, R. (2012). Statistisch denken und forschen lernen mit der software TinkerPlots. In *Beiträge zum mathematikunterricht 2012.* Münster: WTM.

Frischemeier, D., & Biehler, R. (2013). Design and exploratory evaluation of a learning trajectory leading to do randomization tests facilitated by TinkerPlots. In B. Ubuz, C. Haser, & M. A. Mariotti (Eds.), *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (pp. 799–809).

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (p. 235-246). Springer Science+Business Media B.V.

Konold, C. (1994). Understanding probability and statistical inference through resampling. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the IASE* (pp. 199–211). Perugia, Italy: University of Perugia.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, *12*, 217–230. doi: 10.1007/s10758-007-9123-1

Konold, C., & Miller, C. (2011). *TinkerPlots™ 2.0 beta.* Amherst, MA:University of Massachusetts.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289. doi: 10.2307/749741

Leiss, D. (2007). *Hilf mir es selbst zu tun—lehrerinterventionen beim mathematischen modellieren*. Franzbecker: Hildesheim.

Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings

of proto-hypothesis testing. *Pedagogies*, *4*(2), 126–138. doi: 10.1080/15544800902741564

Meyfarth, T. (2008). *Die konzeption, durchführung und analyse eines simulationsintensiven einstiegs in das kurshalbjahr stochastik der gymnasialen oberstufe—Eine explorative entwicklungsstudie*. Franzbecker und Kasseler Online-Schriften zur Didaktik der Stochastik 6: Hildesheim. Retrieved from http://nbn-resolving.de/urn:nbn:de:hebis:34-2006100414792

Podworny, S. (2013). Mit TinkerPlots vom einfachen simulieren zum informellen hypothesentesten. In *Beiträge zum mathematikunterricht 2013*. Münster: WTM.

Rossman, A. (2008). Reasoning about informal statistical inference: A statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19. Retrieved from http://www.stat.auckland.ac.nz/serj

Stratmann, J., Preußler, A., & Kerres, M. (2009). Lernerfolg und kompetenz: Didaktische potenziale der portfolio-methode im hochschulstudium. *Zeitschrift für Hochschulentwicklung*, *4*(1), 90–103.

Vallecillos, A. (1994). *Theoretical and experimental study on errors and conceptions about hypothesis testing in university students* Unpublished doctoral dissertation. University of Granada, Spain.

Voigt, J. (1984). *Interaktionsmuster und routinen im mathematikunterricht—Theoretische grundlagen und mikro-ethnographische falluntersuchungen*. Beltz: Weinheim.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. doi: 10.1111/j.1751-5823.1999.tb00442.x

Zieffler, A., & Catalysts for Change. (2013). *Statistical thinking: A simulation approach to uncertainty* (2nd ed.). Minneapolis, MN: Catalyst Press.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. Retrieved from http://www.stat.auckland.ac.nz/serj

Zieffler, A., Harring, J., & Long, J. (2011). *Comparing groups: Randomization and bootstrap methods using R*. New York: Wiley.

CHAPTER 6

EXPLORING TEACHERS' IDEAS OF UNCERTAINTY

LUCIA ZAPATA-CARDONA

Universidad de Antioquia, Colombia

Abstract

This chapter reports research that studied the ideas of uncertainty held by teachers while working in activities designed to promote informal inferential reasoning. The present study was done within a professional development program for in-service statistics teachers. The program was one semester long and the participants were ten statistics teachers from public schools in Medellin, Colombia. The teachers engaged in the program bringing tasks, teaching materials and class videos to the weekly meetings to promote discussion and reflection. The data for the present report come from teacher's discussions and reflections solving two statistical tasks that took teachers throughout an investigative cycle. The findings reveal that teachers attributed important value to perceptual beliefs and placed less trust in probabilistic reasoning. Additionally, the teachers's use of probabilistic language to quantify uncertainty moved from the extremes of telling everything or nothing to telling something.

Keywords: Uncertainty, Teacher education, Statistics education, Probability tasks

6.1 Overview

There are different meanings of uncertainty. In daily language, it is common to hear expressions like "the results of the game are uncertain" or "it is uncertain whether it will rain today". Uncertainty suggests a measurement of probability that can be the result of a formal or informal process of inference. In the field of statistics, uncertainty is used in subtly different ways. One way of referring to uncertainty is related to hypothesis testing and confidence intervals, which are tools to help the researcher in the process of decision-making. Another way is directly linked to the components of the study design. Several components such as reliability of the measurements, dependability of data management and non-sampling error must be considered when designing a study. There is yet another type of uncertainty that arises when the appropriateness of the data collected to solve the situation of interest is in question, or when confounding variables need to be considered (Arnold, Budgett, & Pfannkuch, 2013).

For the present work, uncertainty is considered a fundamental concept underlying informal statistical inference. In making inferences about the population, we do not really know what the ultimate result is, but we use the outcomes of a representative sample to make an inference to a larger set of data. We looked at how teachers use the ideas of uncertainty to guide the process of informal inferential reasoning. The aspect of uncertainty addressed in this chapter is related to the ideas of uncertainty held by in-service teachers in situations in which they are required to make informal statistical inferences. We studied the ideas of uncertainty held by the teachers while working on tasks designed to promote informal inferential reasoning.

This research was accomplished within the setting of a professional development program. Our main interest was probabilistic language since it constitutes an important tool to look at the ideas of uncertainty. This study followed a qualitative research paradigm and gathered data from the statistics teachers' discussions and reflections after solving statistical tasks. The professional development program was carried out with ten in-service statistics teachers with a wide range of experience who taught at different school levels (elementary, middle and high school). The program allowed teachers to bring, share and discuss instructional tasks with their colleagues but also allowed them to experience some tasks designed by the research team which followed the investigative cycle suggested by Wild and Pfannkuch (1999). The analysis focused on some episodes of teachers solving two statistical tasks and paid attention to the language used. We finished with some implications for research and instruction

6.2 Problem

Ideas of uncertainty are important in developing reasoning about statistical inference at the school-level. While formal statistical inference (e.g., hypothesis testing) is typically taught at the secondary or tertiary level, children at a very young age can use statistical inference informally to make decisions about observed patterns and draw conclusions without formally running a statistical analysis (Ben-Zvi, Gil, & Apel,

2007; Makar & Rubin, 2009). Consequently, teachers should be prepared to deal with uncertainty when it emerges in classroom discussions in relation to informal statistical inference.

There is evidence that statistics teachers often overlook great opportunities in which they can focus on ideas students bring to the class to orient, clarify or discuss statistical concepts (Makar & Rubin, 2009; Zapata-Cardona & Rocha, 2012). The majority of the time, teachers do not neglect those opportunities on purpose, but rather because of their own limitations in articulating students' ideas from classroom discussions to informal statistical inference. One of the reasons for this problem, particularly in the Colombian educational system, is teachers' lack of training in the discipline of statistics and in the teaching of statistics. Colombian teacher education programs in mathematics generally only require one basic statistics course and professional development programs are optional. Once a teacher has a college degree (s)he is certified to teach statistics even though (s)he may never have taken a professional development course. Research has suggested that this degree of preparation is not enough to help teachers successfully develop solid foundations of statistical reasoning in their students (Zapata-Cardona & Rocha, 2011).

One way for teachers to help develop students' ideas of statistical inference is to expose them to problems and situations in which elements of uncertainty are essential. For example, teachers might promote tasks that require students to make predictions including probabilistic language that articulates the level of confidence (Ben-Zvi, Aridor, Makar, & Bakker, 2012). This can be a challenging task for teachers who have not had the proper training.

Many professional development programs for teachers are taught based on the principle that teachers need to be told what to do in the classroom (Arnaus, 1999). According to this idea, a trainer can instruct a group of teachers about how to teach. However, more recent research has suggested that professional development programs should instead be oriented around the circumstances and events that teachers face daily in their practice (Kirkwood & Christie, 2006). The programs should also focus on a better understanding of the relationship between theoretical and experiential knowledge within particular contexts. This approach to professional development can promote teachers' reflection and critical inquiries about their own practice (Humes, 2001).

The purpose of this research is to study the development of ideas about uncertainty of in-service teachers within a professional development program designed to address situations that teachers face on a day-to-day basis in their own classroom. In addition to examining the ideas of uncertainty held by in-service statistics teachers, this work will also study the expressions teachers use to quantify uncertainty. Because of the limited research on teachers' ideas related to inference, this research is based on results from previous research with students to guide a professional development program.

6.3 Literature and Background

Research on informal statistical reasoning has been carried out with students from different levels of the educational system (Arnold et al., 2013; Bakker, Ben-Zvi, Makar, & Kurvers, 2013; Ben-Zvi et al., 2012; Garfield & Ben-Zvi, 2008; Manor, Ben-Zvi, & Aridor, 2013; Metz, 1998; Pfannkuch, 2011). However, informal statistical reasoning seems to be a topic of little interest within teachers' professional development programs. This lack of interest could be grounded in the belief that teachers are already able to guide experiences on informal statistical inference in their teaching.

In the field of cognitive development, the work of Piaget and Inhelder (1975) has strongly influenced the research on uncertainty. They presented children from 5 to 14 years old with a sequence of different physical tasks. The results of their extended study showed that ideas about uncertainty begin to appear in children around seven years old, who prior to this age assume a deterministic causality. In contrast, other researchers (Fay & Klahr, 1996; Kuzmak & Gelman, 1986) found that preschool age children have an understanding about uncertainty.

The literature on decision-making has also contributed to the understanding of people's reasoning about uncertainty. This field of research has documented the difficulties adults face making probabilistic judgments. Usually, a typical adult fails to make any probabilistic distinction between determinacy and indeterminacy (Konold, 1991) and most of the time assigns deterministic behaviors to phenomena that are regulated by chance (Kahneman & Tversky, 1982). Consequently, the adults fail to recognize the extent to which chance contributes to what they experience about the world.

Educational studies have provided an interesting view on students' reasoning. A particular study focused on the development of students' expressions of uncertainty in reasoning from samples (Ben-Zvi et al., 2012). The researchers were able to show the evolution of fifth graders' probabilistic language. Another influential study (Makar & Rubin, 2009) developed a theoretical framework for how people reason about informal statistical inference. In the proposed framework, the authors highlight three principles that appear to be essential in informal statistical inference: Generalization, use of data as evidence and use of probabilistic language. The study highlighted that the correct use of probabilistic language is important to avoid deterministic claims. However, in the classes the authors studied, little attention was paid to probabilistic language.

Many research studies have documented the difficulties encountered when learning about statistical inference. This literature contains studies that disclose different errors and misuses in reasoning about inference (e.g., Watson, 2002). It also focuses on the exploration of how to develop students' reasoning about statistical inference (Ben-Zvi et al., 2012; Franklin et al., 2007; Garfield & Ben-Zvi, 2008; Pfannkuch, 2005; Pfannkuch & Wild, 2000; Wild & Pfannkuch, 1999). One method of developing students' reasoning about uncertainty that has shown promise in this literature is the focus on informal statistical inference.

6.3.1 Informal Statistical Inference

Informal statistical inference is described in the literature as the process of making probabilistic generalizations from a sample to a population without running a formal statistical test. It is the act of looking beyond the data to cases outside of the sample at hand (Makar & Rubin, 2009) and the cognitive activity involved in drawing conclusions or making predictions about "some wider universe" (Garfield & Ben-Zvi, 2008). Informal statistical inference takes into account multiple dimensions such as data, distributions, measurements, representations, and statistical models.

In making informal statistical inferences, language is essential since it articulates the level of confidence or uncertainty in the prediction (Ben-Zvi et al., 2012; Makar & Rubin, 2009). One of the tools in statistical inference is the information gathered from samples. However, the results from a sample might lead the learner to think that the sample reflects the behavior of the population (Rubin, Bruce, & Tenney, 1991). This is known as over-reliance on sample *representativeness*.

In contrast, the learner might doubt the information given from the sample because of the variability intrinsic in every sample. The learner might conclude that the sample does not give relevant information about the population and attribute the results exclusively to chance. This is known as over-reliance on sample variability. These two methods of judging results from a sample reflect either a deterministic or a relativistic view of the learner that influences the language used in the informal inference. According to Rubin and colleagues (1991), the information from a sample should not tell everything or nothing, but something about the subjacent population.

A possible educational approach that might help to support the development of informal statistical inference is the statistical investigations inspired in the "investigative cycle" (Wild & Pfannkuch, 1999). In this approach, the participants are involved in the solution of a problem that takes them through the stages of the investigation process (questioning, planning, gathering data, analysis, and interpretation). The investigative cycle is also highlighted in the *Guidelines for Assessment and Instruction in Statistics Education* (Franklin et al., 2007). This document, endorsed by the American Statistical Association, emphasizes that the statistical question is a very important beginning of the investigation.

At the *Eighth International Research Forum on Statistical Reasoning, Thinking and Literacy* (SRTL-8), the results from several different research studies related to people's reasoning about uncertainty were presented. For example, there were studies looking at new ideas of uncertainty emerging in students during an introductory statistics course focused on bootstrapping and randomization methods (Arnold et al., 2013). Others were interested in students' web of actions and reasons involved in reducing uncertainty in solving a real problem (Bakker et al., 2013). While others were interested in confronting how confident students were with their inferences after working with sampling distributions (Manor et al., 2013). Different ways to approach the study of uncertainty lead to different ideas about uncertainty. Whereas the research conducted by Arnold et al. (2013) and (Bakker et al., 2013) was focused on aspects of uncertainty related to study design and hypothesis testing, the research

carried out by Manor et al. (2013) was concerned with the level of confidence expressed by the participants when making inferences.

6.4 Subjects and Methods

This study follows a qualitative research paradigm since the key interest—the ideas of uncertainty—is a phenomenon that is qualitative in nature (Sánchez-Gamboa, 1998). The research was carried out in a professional development program for teachers. Such a natural environment, according to Suter (2006), is favorable when the research is focused on discovering how study participants construct their own meaning of events or situations. A qualitative research orientation "honors the understanding of a whole phenomenon via the perspective of those who actually live it and make sense of it" (p. 344). In other words, it takes into account the subjective experiences or internal states (emotions, thoughts, reflections, etc.) of the participants: In this particular case, the expressions teachers use to refer to uncertainty.

6.4.1 Setting and Participants

To address the goal of the research, data were gathered from teachers' discussions and reflections in a professional development program carried out during the first semester of 2013. The unit of analysis was teachers' discourse, since discourse and cognition are related (Lerman, 2001). The program followed the principles of a community of practice (Wenger, 1998) where every teacher was expected to contribute with something to the team (experience, reflections, tasks, lesson plans, etc.). The program started with ten in-service statistics teachers (one elementary, nine secondary) who were part of a self-called research group[1] in teaching mathematics promoted by the Secretary of Education of Medellin, Colombia. The teachers represented a broad range of experience—from a teacher in his second year, to a teacher with fifteen years of experience. The teachers all worked for public schools in the city and held undergraduate degrees in mathematics education; some of them were pursuing Master's degrees in the sciences. All the teachers had taken a course in which descriptive statistics was taught, but none had taken a statistical methods course.

The professional development program, which met for an entire semester, met for three-hours once a week. In every meeting, the teacher educators brought some exemplary materials to discuss and promote reflection among the teachers; however, the teachers were also welcome to bring and share activities, teaching materials, difficulties, reflections, and achievements in their statistics teaching.

As a mechanism to start the program, we asked teachers to bring a lesson plan for a statistics class that they were about to teach. They planned the lesson in pairs and presented it to their colleagues to get comments to either improve the lesson or reflect

[1] Although the teachers referred to themselves as a research group, they had not been involved in a research project. The "research group" expression might reflect the positive perceptions they had of themselves as a team.

on crucial aspects from the planning. Some teachers even taught and videotaped a class that was based on their lesson plan, and brought the video to share and discuss with their colleagues. Throughout the professional development program, we paid close attention to the teachers' discussions and reflections because their comments were important inputs to orient the subsequent gatherings and to propose specific activities.

During the professional development program, the teachers were engaged in solving statistical tasks. In the process of solving the tasks, the teachers were encouraged to think about their level of certainty when they were making inferences. Most of the tasks involved familiar settings for the teachers. For example, one task was related to the probability of success for a student answering a multiple-choice test by guessing. Although there were several tasks that were worked on during the professional development program, in this chapter we focus on only two of the tasks. These tasks were selected because they involved decision-making based on quantification of uncertainty, promoted rich discussion among teachers and offered details that might help address the goal of the research. Although the goal of this research focused on the professional development program and on the development of teaching materials, the data offered multiple opportunities to study uncertainty.

6.4.2 Tasks

The tasks proposed to the teachers were designed taking into account the investigative cycle suggested by Wild and Pfannkuch (1999). While teachers worked on the tasks, we paid close attention to the probabilistic language used. In addition, we paid attention to the participant's engagement with predictions, followed by data generation to support informal inferences, and finally, reflection on the discrepancies between predictions and outcomes. Taking this approach was very important since the participants had only taken an algebra-based statistics course. We think that having the teacher go through the process of an investigative cycle, within an interesting task, opens the door for promoting informal statistical inference and encourages that teacher's reflection on his or her own practice as a teacher of statistics.

The Horse Race Task. One proposed task was a horse race game. The teachers were asked to answer the following question: "Which horse would you bet on?" The teachers received six different color chips, two dice and a handout with the diagram shown in Figure 6.1. There were six horses numbered from zero to five; the players associated each horse with a color chip. The rules of the game were explained. The game was played in pairs; each team ran a race in which the horses progressed according to the difference in tossing two dice. The teachers had to make their bets before starting the race. Each teacher could select up to two different horses. The race was over when a horse got to the finish line.

The Discrimination Task. The second task, taken from Scheaffer, Gnanadesikan, Watkins, and Witmer (1996), explored a situation of possible discrimination. The task was inspired by a study published in the *Journal of Applied Psychology* (Rosen

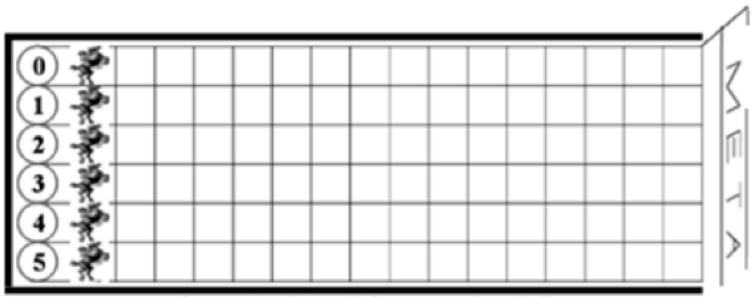


Figure 6.1. Handout for the simulation of the horse race.

& Jerdee, 1974). The statistical question of interest was whether bank supervisors were discriminating against women during a promotion process for bank manager positions. In the actual case, there were 48 applicants (24 males and 24 females). Of the 35 applicants recommended for promotion, 21 were male and only 14 were female. Table 6.1 summarizes this information.

Table 6.1
Counts for the Discrimination Task

	Recommended for Promotion	**Not Recommended for Promotion**	**Total**
Male	21	3	24
Female	14	9	24
Total	35	13	48

The Discrimination Task was an interesting scenario to engage teachers in discussions related to informal inferential reasoning. We wanted to study what information they used to decide whether there was any discrimination. We also wanted to know how certain they felt about their decision.

6.4.3 Data Analysis

Each session of the professional development program was video recorded. Videos were observed, transcribed, translated into English and annotated to help capture teachers' ideas of uncertainty when making informal statistical inferences. Three undergraduate students, three graduate students, and two faculty members from the mathematics education and science education programs took part in the data analysis. All were also participants in the research seminar. This team met once every three weeks to share progress about other ongoing research (four different projects were going on at that time) and also to discuss results and interpretations from the video analysis. When there was disagreement, the research team would continue discussion

until reaching a consensus. Entries from the teachers' journals were also examined, whenever possible, and used to validate interpretations from the video analysis. Despite the large amount of data collected, only those episodes that offered information about the probabilistic language used by the teachers when making informal inferences were selected for the analysis. Although there were ten participants engaged in the professional development program, we only report on the analysis for the teachers who voluntarily decided to share their reasoning about the two tasks.

6.5 Analysis and Results

Having described the research setting, participants, tasks, and data analysis methods, we are ready to set out to respond to our research goal of examining teachers' ideas of uncertainty in the process of their informal statistical reasoning. In this section, we describe the analysis of the data collected for the two tasks. In the first task, teachers needed to quantify uncertainty in order to decide which of the six horses to bet on. In the second task, the teachers needed to quantify the uncertainty in order to make an unbiased judgment about whether there was sex discrimination in a company's promotion process.

6.5.1 The Horse Race Task

Once the task was explained, the teachers[2] selected their horses (see Table 6.2) and started the game. As the game progressed, many of the teachers felt surprised because the horses they selected did not advance in the race as they had initially expected. To stimulate discussion we asked the teachers to explain the aspects they took into account for their selection.

Table 6.2
Teacher Pairs' Wagers and Results

Teacher Pairs	**Horse Number Wagered**	**Winning Horse Number**
Nancy and Elmer	3 and 0	2
Cristina and Daniel	2 and 5	1
Andrés and Germán	3 and 5	1
Zaida and Juan	3 and 1	1
Wilson and Francisco	4 and 5	1
Research Assistants 1 and 2	4 and 5	3

[2]The teachers' names are pseudonyms to protect the identity of the participants.

While the game was in progress, we asked the teachers whether they noticed any special pattern in the race thus far. One of the teachers (Daniel) said, "the number five has bad luck" (he had selected Horse #5). This expression, at that point, makes us think that the teacher did not use any strategy to try to quantify uncertainty. He attributed the results exclusively to luck. He looked at the data gathered up to that moment and generalized with the expression bad luck. He took into account what he saw to conclude a pattern. At that particular moment Daniel observed the information offered by the sample and he relied on its representativeness to express that statement.

After the teachers finished the game, we explored the reasons that they used to initially select a horse. One of the teachers (Germán) said, "I choose [number] five because I like it so much". Another teacher, Juan, was a frequent player of *ludo*[3] and he had some sense of the distribution of the outcomes after throwing two dice. However, this sense of distribution was not enough to support his argument about selecting the horse number one. He said,

> I discarded [the horse] number five because I saw that only when I get seven [six and one] I could advance that number. [I discarded] The [horse number] zero because one thinks that always after throwing two dice the most difficult is to get pairs. But the [horse number] one, I chose it because of the color. Simply, I like that color [red].

Juan, in his explanation, showed some indication of statistical reasoning. He knew how to advance with number five after throwing the dice and had some clues about the scarceness of pairs. However, at the end, his perceptual beliefs were stronger for selecting the red chip associated with horse number one, the one he had selected. He admitted that the color was decisive in the selection.

Teachers' reasoning to select the horses seemed to be primarily based on perceptual beliefs (such as color preference, number, or position of the horse) and not on probabilistic judgments in which the sample space and other statistical aspects would be considered. This is consistent with results reported in previous research that state: "People do not follow the principles of probability theory in judging the likelihood of uncertain events" (Kahneman & Tversky, 1972, p. 431).

After pooling the winning horses from the different races (Table 6.2, third column), the teachers noticed that Horse #1 won in most races. However, in one of the races, the winning horse was Horse #2 (the experimental results of one race and the way the horses advanced are shown in Figure 6.3). We took advantage of the situation and asked the following question: "It seems that horse number one has a tendency to win, but in one of the races the winner was horse number two; how can you explain this phenomenon?" One of the teachers, Daniel, said, "because of the randomness. Probability helps us to make a decision but that [result] is not totally sure". His justification revealed some indication of considering a probabilistic judgment. He went to the middle ground between representativeness and variability. The expression "that is not totally sure" seems to be an intent to offer a degree of uncertainty that can

[3]Ludo is a board game in which players advance counters according to the results of throwing two dice.

be attributed to the non-deterministic nature of statistics tasks. It appears that this teacher has a sense of confidence that Horse #1 could win, but something else could happen. He recognized the sample did not tell everything, nor did it tell nothing, but something (as expressed by Rubin et al., 1991). We suspect this change happened as a result of experiencing and reflecting on the task.

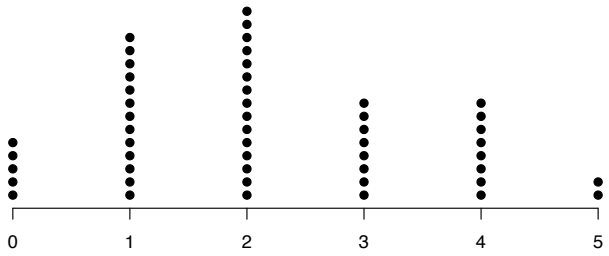


Figure 6.2. Nancy and Elmer's experimental results for the horse race task.

After noticing trends in the race results, the teachers constructed a table to show the different combinations and the theoretical probability distribution (as shown in Tables 6.3 and 6.4) and contrasted these results with their experimental results. Table 6.4 helped the teachers understand that Horse #1 and Horse #2 have the highest probabilities of winning a race.

Table 6.3
Combination of Results–Difference

Die	**1**	**2**	**3**	**4**	**5**	**6**
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

Table 6.4
Theoretical Probability Distribution

	Horse Number					
	0	**1**	**2**	**3**	**4**	**5**
Favorable events	6	10	8	6	4	2
Probability	0.16	0.27	0.22	0.16	0.11	0.05

In one of the races, the winner was Horse #3. The teachers were curious about this outcome and asked about a suitable explanation for this result. Daniel used the pooled results and the constructed table of the probability distribution to justify his reasoning,

> To get number one, we have ten favorable events out of thirty-six possible ones, but we have twenty-six [events] that are not favorable. This means that it is easier to get a number whose difference is different from one than to get one. Having the highest probability does not mean that it is always going to win.

It is interesting that Daniel was able to relate the results back to the table of the probability distribution in order to support his explanation. However, the first part of his argument does not really explain why having Horse #3 as a winner could be a suitable outcome. The second part of his explanation was stronger and offered, again, a quantification of the uncertainty involved in this situation. A high chance does not guarantee a certain event. This was another indication that Daniel used the sample to obtain some information—not nothing, not everything, but something.

Another associated result was the evolution of language used by the teachers. The probabilistic language the teachers used when making predictions or generalizations in this task moved from deterministic or relativistic extremes to the middle. At the beginning of the task, the teachers' expressions revealed that they saw the sample as a provider of all the information or they did not see the sample as a provider of any information. However, as the teachers progressed through the task, their language became more and more refined. They went from deterministic expressions such as, "the number five has bad luck", to relativistic expressions, such as, "having the highest probability does not mean that it is always going to win".

The horse race task gave the teachers the opportunity to be involved in a hands-on activity, to talk about uncertainty, and at the same time gave them inputs for their own reflection. One of the teachers, Nancy, expressed,

> It is too obvious that the horses advance in counts according to the difference on the two dice. I wonder why we did not bet on those values [those with the highest probability]. This activity was quite entertaining and if we liked it, students might enjoy it even more.

Similar to Nancy, we were also surprised that at the beginning of the game the teachers selected the horses for the race using personal criteria and not necessarily taking into account probabilistic criteria. However, the teachers' engagement in the task allowed them to contrast the performance of their selection with the theoretical

probability, which we consider a crucial point for promoting reflection about their own decisions.

Proposing tasks that involve uncertainty where the teachers go through the investigative cycle (generating simulated data, analyzing the information, and making sense of the results to explain a particular behavior) seems to constitute an appropriate tool for helping teachers to contrast their initial intuitions with simulated and theoretical data. Many of their initial intuitions were associated with perceptual beliefs, external attributions of the counters and luck. However, as they progressed through the cycle to solve tasks with a degree of uncertainty, their decisions became more based on basic principles of probability theory. Additionally, exposing teachers to the investigative cycle seems to help them refine their probabilistic language. They moved from deterministic and relativistic expressions about the results of a sample to expressions in the middle ground of this continuum.

In relation to these results, Wild and Pfannkuch stated that the investigative cycle could be useful for building "a more holistic feel for statistical investigation" (1999, p. 243), since it "is a high-level description of a systematic approach to investigation. It identifies major elements. It can be used as the foundation for something that is much more of a prescription" (p. 243). This make sense if we take into consideration that the final goal of the statistical investigation is learning from real contexts where uncertainty is present.

6.5.2 The Discrimination Task

After presenting the discrimination scenario, we asked teachers to argue, based on the provided information, whether there was enough evidence to believe the company had discriminated against the women in their promotion process. The initial discussion was intense. It took a while for teachers to understand the actual study and they gave several suggestions for designing a different study to detect discrimination. Some teachers strongly believed that there was no evidence of discrimination, some were undecided and some stated that there was evidence of discrimination.

For example, Daniel, one of the teachers who indicated no evidence of discrimination, said that he took the information given into account, but he did not consider a comparison of the proportion of people promoted between the genders. He based his argument on different reasoning, saying, "if there were 24 women and 14 were recommended for promotion, that is more than a half. I do not think that there is discrimination". He also stated, "one does not promote everyone". Daniel also argued that the counts on the table might have been possible just by chance alone. It is plausible, based on his statements, to think that the results of the sample did not offer him enough information and he simply thought that it was a reasonable result among all the possible samples—over-reliance on sample variability (as is mentioned by Rubin et al., 1991).

To study the discrimination scenario, Daniel suggested comparing the counts within the female group; he compared the number of women promoted with the total number of female applicants. He felt that making a comparison between males and females was not necessary in order to make a decision about discrimination.

Another teacher, Juan, mentioned that he did not see how statistics could be used to solve the problem of discrimination. He said, "in these type of situations, one could see statistics as a tool. However, to what extent, with this information and using statistics, could one decide whether there was discrimination or not?" In spite of his comment, Juan ultimately compared the proportions of the males and females promoted, and based on the differences stated, "I think that there was discrimination". But, he did not mention how certain he was about the stated difference. It seems that Juan did not see relevant information from the results shown in the sample, suggesting he has an over-reliance on sample variability—no information. However, he went back to the data from the sample and compared the values on the table to suspect some discrimination.

It is interesting to note that teachers' answers to the discrimination question were often at the extreme ends of the scale of probability. They either saw discrimination or did not see it, but they failed to consider a range of possibilities to quantify uncertainty. This might be explained from a philosophical point of view where determinism makes us feel comfortable and we try to stay away from situations that require us to consider variation. "The stronghold of the deterministic sentiment is the antipathy to the idea of chance" (James, 2007, p. 153). According to James, determinism professes the universe is already established and the future does not have ambiguous possibilities.

To help teachers understand the phenomenon, we simulated the random variation model by using a deck of cards. We asked the teachers to create a deck of cards with 24 black cards (males) and 24 red cards (females). They shuffled the cards and then dealt out 35 cards (the promoted applicants) and counted the number of black cards out of 35. They recorded this number, repeated the simulation four more times, and put together their results in a dot plot on the board (see Figure 6.3).

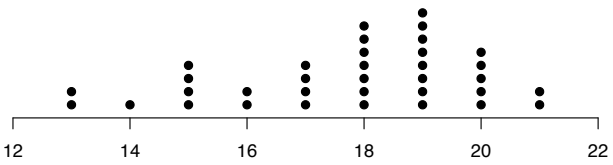


Figure 6.3. Simulation results of the number of men promoted.

In collecting and organizing the data from the simulations, the teachers who initially predicted discrimination began to confirm their previous intuitions. Those who did not anticipate discrimination started to reconsider their predictions. The generation of simulated data also allowed teachers to visualize a sampling distribution

and discover the rarity of having 21 out 24 men promoted assuming a model of no effect (no difference, no discrimination). The sampling distribution was an important artifact that gave the teachers some statistical evidence to confirm or refute their previous predictions.

The simulated distribution also allowed teachers to quantify, at least experimentally, the degree of confidence for their conclusions. A value of 21 or higher only occurred two out of 35 times; that is, it occurred less than 6% of the time! The value of 6% was a simulated p-value (the chance of observing 21 or higher, if there is no discrimination). We finished the activity at this point and decided not to carry out a more formal statistical analysis.

6.6 Discussion and Implications

The goal of the professional development program described in this chapter was to bring together in-service teachers to share experiences in their statistics teaching. The tasks presented here were inputs to study the ideas of uncertainty held by the teachers when they go through an investigative cycle. We found that teachers attributed important value to perceptual beliefs while probabilistic reasoning was not a resource commonly used to support their decisions. At the same time, we found that the teachers' informal statistical inferences relied either on sample representativeness or sample variability. However, after being involved in solving statistical tasks that resemble statistical investigations, teachers' inferences were located in the middle ground between representativeness and variability.

The tasks proposed in this research allowed us to see some degrees of quantification of the uncertainty in teachers' talks. However, the probabilistic language to quantify uncertainty was scarce. This could suggest teachers' strong tendency to see the world from a deterministic point of view but at the same time could suggest the lack of power of the tasks to promote the use of probabilistic language. The tasks that state a statistical question and require generation of data to analyze and make a decision seem to be valuable resources for the teachers. The discrimination task, for example, allowed them to confront their initial intuitions that were at the extremes of the continuum from representativeness to variability. In addition, it supported teachers' decision-making process based on a visual distribution of possible values.

In future research, it is clear that the number of tasks needs to be increased. This study is a report based exclusively on two tasks and perhaps, coincidentally, the results from them are comparable—the participant teachers moved from the extremes of over-reliance on sample representativeness and sample variability to less deterministic expressions of uncertainty. However, it is admissible to think about increasing the number of tasks in future research. Otherwise, we ourselves would be falling into the trap of over-reliance on sample representativeness, the same condition we helped teachers to overcome. Increasing the number of tasks would also help to establish other patterns that might emerge in teachers' consideration of uncertainty.

The epistemological foundations of this research are anchored in a social perspective of teachers' professional development. Following this perspective, the tasks pro-

posed to the participants were undertaken by a community of teachers. We are aware that when problem solving takes place in a group setting, many teachers' voices are often not heard—only a few have the courage to lead the discussions. It would be interesting to use smaller groups (e.g., 2–3 teachers at a time) to study whether similar patterns in teachers' articulation about uncertainty would be found.

Although the purpose of this research was not the study of teachers' content knowledge, it was evident that a large limiting factor was the teachers' lack of knowledge and exposure to statistical inference. The results from this study provide preliminary evidence that we need to give more thought to our pre-service teachers' programs of study. The statistics courses the teachers are taking in their undergraduate programs do not seem to prepare them for the content and statistical inquiry they would need to teach in a statistics course at the school-level. Teachers will likely need other professional development opportunities, not only for developing deeper content knowledge than their pre-service education provides, but also for qualifying their practice.

Designing a program that gradually helps teachers to develop sophisticated notions of uncertainty is challenging. Although we are not comfortable making gross generalizations due to the small number of participants, we do think that some of the structure and findings from this study could be considered when teaching school statistics. We have come to believe that the structure of the community, the group discussions, the reflections and the confrontation of simulations with theoretical results all interacted in promoting the development of the teachers' ideas of uncertainty. Educational approaches that encourage teachers to talk and reflect on making sense of data and statistical inferences can support teachers learning. Professional development programs for statistics teachers should be designed to expose in-service teachers to statistical investigations.

Acknowledgements

This research was supported under a grant from the University of Antioquia Research Committee–CODI–Social Sciences and Humanities, 2012.

References

Arnaus, R. (1999). La formación del profesorado: Un encuentro comprometido con la complejidad educativa. In J. F. Angulo, J. Barquín, & A. I. Pérez (Eds.), *Desarrollo profesional del docente: Política, investigación y práctica* (pp. 599–635). Madrid, Spain: Akal.

Arnold, P., Budgett, S., & Pfannkuch, M. (2013). Experiment-to-causation inference: The emergence of new considerations regarding uncertainty. In A. Zieffler & E. Fry (Eds.), *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 119–146). Two Harbors, MN: University of Minnesota.

Bakker, A., Ben-Zvi, D., Makar, K., & Kurvers, T. (2013). Reducing uncertainty in a hospital laboratory: A vocational student's web of reasons and actions involved in making a statistical inference. In A. Z. . E. Fry (Ed.), *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 34–48). Two Harbors, MN: University of Minnesota.

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM—The International Journal on Mathematics Education*, *44*, 913–925. doi: 10.1007/s11858-012-0420-3

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? young students reason and argue about some wider universe. In *Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5).* United Kingdom. Retrieved from https://sites.google.com/site/danibenzvi/allpublications

Fay, A. L., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development*, *67*, 689–716. doi: 10.1111/j.1467-8624.1996.tb01760.x

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer.

Humes, W. (2001). Conditions for professional development. *Scottish Educational Review*, *33*(1), 6–17.

James, W. (2007). The dillema of determinism. In *The will to believe and other essays in popular philosophy* (pp. 145–183). New York: Cosimo, Inc.

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

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*, 143–157. doi: 10.1016/0010-0277(82)90023-3

Kirkwood, M., & Christie, D. (2006). The role of teacher research in continuing professional development. *British Journal of Educational Studies*, *54*(4), 429–448. doi: 10.1111/j.1467-8527.2006.00355.x

Konold, C. (1991). Informal conceptions of probability. *Cognition and Instruction*, *6*, 59–98. doi: 10.1207/s1532690xci0601_3

Kuzmak, S., & Gelman, R. (1986). Young children's understanding of random phenomena. *Child Development*, *57*, 559–566.

Lerman, S. (2001). Cultural, discursive psychology: a sociocultural approach to studying the teaching and learning of mathematics. *Educational Studies in Mathematics*, *46*(1–3), 87–113. doi: 10.1007/0-306-48085-9_3

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Manor, H., Ben-Zvi, D., & Aridor, K. (2013). Students' reasoning about uncertainty while exploring sampling distributions in an "integrated approach". In A. Z. . E. Fry (Ed.), *Proceedings of the Eighth International Collaboration for Research on Statistical Reasoning, Thinking, and Literacy (SRTL-8)* (pp. 18–33). Two Harbors, MN: University of Minnesota.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction*, *16*(3), 285–365. doi: 10.1207/s1532690xci1603_3

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267–294). New York: Springer.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1–2), 27–46. doi: 10.1080/10986065.2011.538302

Pfannkuch, M., & Wild, C. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, *15*(2), 132–152.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. London: Routledge & Kegan Paul.

Rosen, B., & Jerdee, T. (1974). Influence of sex role stereotypes on personal decisions. *Applied Psychology*, *5*(9), 9–14.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics (icots-3)* (Vol. 1, pp. 314–319). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf

Sánchez-Gamboa, S. (1998). *Fundamentos para la investigación educativa: Presupuestos epistemológicos que orientan el investigador [Foundations for educational research: Epistemological assumptions that guide the researcher].* Bogotá: Cooperativa Editorial Magisterio.

Scheaffer, R., Gnanadesikan, M., Watkins, A., & Witmer, J. (1996). *Activity-based statistics*. New York: Springer.

Suter, W. N. (2006). *Introduction to educational research: A critical thinking approach*. Thousand Oaks, CA: Sage.

Wenger, E. (1998). *Communities of practice learning, meaning and identity*. Cambridge: Cambridge University Press.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. doi: 10.1111/j.1751-5823.1999.tb00442.x

Zapata-Cardona, L., & Rocha, P. (2011). *Actitudes de profesores hacia la estadística y su enseñanza.* Conferencia Interamericana de Educación Matemática-CIAEM. Recife, Brasil.

Zapata-Cardona, L., & Rocha, P. (2012). Teachers' questions in the statistics classroom. In D.Ben-Zvi & K. Makar (Eds.), *Teaching and learning statistics.*

Proceedings of Topic Study Group 12, 12th International Congress on Mathematial Education (ICME-12) (pp. 53–59). Seoul, Korea.