ESTABLISHING STATISTICAL FOUNDATIONS EARLY: DATA MODELING WITH YOUNG LEARNERS

Lyn D. English
Queensland University of Technology, Australia
l.english@qut.edu.au

This paper addresses research from a three-year longitudinal study that engaged children in data modeling experiences from the beginning school year through to third year (6-8 years). A data modeling approach to statistical development differs in several ways from what is typically done in early classroom experiences with data. In particular, data modeling immerses children in problems that evolve from their own questions and reasoning, with core statistical foundations established early. These foundations include a focus on posing and refining statistical questions within and across contexts, structuring and representing data, making informal inferences, and developing conceptual, representational, and metarepresentational competence. Examples are presented of how young learners developed and sustained informal inferential reasoning and metarepresentational competence across the study to become "sophisticated statisticians".

INTRODUCTION

The ability to reason effectively about data is integral to making meaningful, informed decisions across all spectrums of life. One cannot participate effectively in political debates about community issues such as the environment, health care, and education without this reasoning ability. Indeed, being a "literate" and informed citizen necessitates skills in investigating, interpreting, understanding, and critically evaluating statistical information especially in the abuse of data and statistics (Batanero, Burrill, & Reading, 2011). Young children are very much a part of our data intensive community, where they are exposed to a vast array of statistical information designed to captivate and convince their inquisitive minds. Statistical literacy takes a long time to develop and must begin in the earliest years of schooling (English, 2013; Franklin & Garfield, 2006; Shaughnessey, 2006). The foundations of such literacy, however, need broadening--traditional experiences involving fictitious data that require minimal interpretation are no longer sufficient.

Numerous definitions of statistical literacy exist, usually incorporating the ability to understand and use data effectively and critically to make informed decisions (e.g., Mandinach & Gummer, 2013; Watson, 2006). Watson (2006), for example, defined statistical literacy as the nexus of statistics, probability, and the everyday world where "statistical tools, general contextual knowledge, and critical literacy skills" are applied in "spontaneous decision-making" within "unrehearsed contexts" (p. 11). One approach to fostering young children's statistical literacy is through data modeling (e.g., English, 2013; Lehrer, Kim, & Jones, 2011). Data modeling engages young learners in the entire complexity of inquiry, encompassing core components from statistics and probability. Interpreting contexts, posing and refining statistical questions, and entertaining possible data are essential starting points, aspects that have been neglected in early statistics education. Likewise, the application of informal inference including variation, uncertainty, and prediction throughout the inquiry process is critical, yet currently largely ignored (English & Watson, 2013). Specifically in this paper, I give consideration to young children's capabilities in detecting variation and making predictions within data contexts, and their metarepresentational competence in displaying and analyzing data.

DATA MODELING

Research on data modeling with young children has been limited, possibly reflecting an apparent lack of awareness or appreciation of their statistical capabilities. A further likely obstacle is the challenge faced in designing activities where children are free to generate their own statistical ideas prior to formal instruction. Yet research by Lehrer and Schauble (e.g., 2007), English (2013), and Kinnear (2013) has demonstrated quite clearly the statistical competencies young children possess. Their research has shown the developmental nature of data modeling, beginning with young children posing their own questions about meaningful phenomena, identifying and

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

measuring attributes in undertaking subsequent investigations, and progressing to interpreting, organizing, structuring, and representing their data (Lehrer & Lesh, 2003). Opportunities for extending their data through further inquiry are also incorporated. Drawing informal inferences including identifying variation and making predictions, is foundational in undertaking a data modeling activity yet such inference is frequently overlooked in early statistical investigations. Figure 1 displays some of these core components of this data modeling approach.

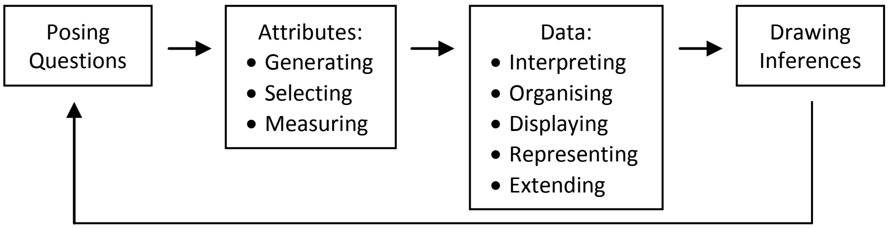

Fig. 1. Data Modeling Cycle

Informal Inference

Informal or beginning inference has received increased attention in recent years (e.g., English & Watson, 2013; Makar, Bakker, & Ben-Zvi, 2011), although much less so with respect to young children whose inferential capabilities have been overlooked in the elementary school. Beginning inference encompasses several factors including variation, prediction, hypothesizing, and uncertainty, the first two of which are addressed here.

The importance of variation has been extensively documented, with researchers highlighting how it lies at the heart of statistical reasoning and is linked to all aspects of statistical investigations (Cobb & Moore, 1997; Pfannkuch, 2005; Watson, 2006). Indeed, as Watson (2006) indicated, the reason data are collected, graphs are created, and averages are computed is to "manage variation and draw conclusions in relation to questions based on phenomena that vary" (p. 21). Little research has been conducted on young children's reasoning with variability and variation, although the work of Watson (e.g., Watson, 2006) has revealed a primitive understanding of these concepts.

Young children's abilities to make predictions based on data also remain untapped. Obviously they do not have the mathematical background to undertake formal statistical tests, but they can nevertheless draw informal inferences based on their understanding of the data presented and on aspects of the problem scenario and context. As Watson (2006) indicated, one of the aims of statistics education is to help students make predictions that have a high probability of being correct. Yet in the real world, decisions are required where there is uncertainty and where several alternatives might be reasonable. Hence, young children's exposure to informal inference involving uncertainty is an important learning foundation if a meaningful introduction to formal statistical tests is to take place in the secondary school.

Metarepresentational Competence

Data modeling fosters metarepresentational competence (e.g., diSessa, 2004; English, 2013) because models are typically conveyed as systems of representation, where structuring and displaying data are fundamental. The structure is constructed, not inherent (Lehrer & Schauble, 2007). When children construct and display data models, they generate their own forms of inscription. By the first grade, children already have developed a wide repertoire of inscriptions, including common drawings, letters, numerical symbols, and other referents. As they invent and use their own inscriptions, children also develop an "emerging meta-knowledge about inscriptions," which provides a basis for their mathematical activity (Lehrer & Lesh, 2003). The term, *metarepresentational competence*, coined by diSessa and his collaborators (diSessa, 2004; diSessa, Hammer, Sherin, & Kolpakowski, 1991), includes students' abilities to invent or design a variety of new representations, explain their creations, understand the role they play, and critique and compare the adequacy of representations. Unlike the standard representational techniques students might have learned from specific instruction, metarepresentational competence encompasses students' "native capacities" (diSessa, 2004, p. 294) to create and re-create their own representational forms.

Limited research has been conducted on developing young children's metarepresentational competence, which is perhaps not surprising given their typical exposure to adult-created data displays. Even more uncertain is whether and how children's conceptual understanding develops in conjunction with this metarepresentational competence, and further, whether the latter might advance or hinder the former.

STUDY OVERVIEW

To illustrate young learners' capabilities in drawing informal inferences and representing data, examples are taken from a three-year longitudinal study from first through to third grade (6-8 years). Three classes of children and their teachers from a middle socio-economic Queensland state school engaged in multiple, life-based experiences incorporating other disciplines such as health and nutrition, and environmental studies. Literature was used as a basis in designing the classroom activities, with story picture books serving to both introduce the activities and provide the data context for the children's statistical reasoning. The multi-component activities of each year involved the children in each phase of data modeling as displayed in Fig. 1.

A design-based research approach was adopted, specifically a design experiment, involving the learning of students, teachers, and researchers (Kelly, Lesh, & Baek, 2008). Comparative case studies (small focus groups of students) were also included. The activities were designed in collaboration with the teachers, who implemented them in their classrooms while we (author and senior research assistant) observed and assisted the teacher where necessary. The adults acted as facilitators not directors of the children's learning, that is, the children were free to respond in ways they chose. Data collection included videotaping and audiotaping of the focus groups as well as all whole-class discussions. Using iterative refinement cycles for analyses of children's learning (Lesh & Lehrer, 2000), the transcripts of all focus group work and classroom discussions were reviewed many times in conjunction with all of the children's artifacts. The data were coded and examined for patterns and trends using constant comparative strategies (Creswell, 2012).

CHILDREN'S INFORMAL INFERENTIAL REASONING

Examples of the children's capabilities in detecting variation and making predictions are drawn from two activities, namely, *Litterbug Doug* (first year) and *Baxter Brown's Picnic* (second year). The third of four activities implemented in the first year, *Litterbug Doug* was based on the picture book by Bethel (2009), relating the story of how Litterbug Doug was originally a dirty creature who lived in a pile of rubbish in a very clean town. A "green-caped crusader" then swooped to the Earth to reform Litterbug Doug, who subsequently became the Litter Police for the town and enthusiastically monitored their environment. Following the previous activities in which the children posed questions, identified and generated attributes, organized and displayed their data in multiple ways, *Litterbug Doug* engaged the children in interpreting tables of data, identifying variation in the data, posing questions, and making predictions. On reading the book, the teacher explained that, "Now that Litterbug Doug has become the Litter Police, the townsfolk are interested to see what he collects in Central Park during his first 3 days. They also want to know if Litterbug Doug is doing a good job of collecting litter in Central Park." The children were then shown a simple table of the numbers of rubbish items he collected on the Monday, with the teacher explaining, "As a start, the town's mayor asked Litterbug Doug to show him what he collected on his first day, Monday. Litterbug Doug showed the mayor what he saw and what he collected in the park." Following discussion on this first data collection, the table was then extended to display values for the items collected Tuesday and Wednesday (<10), with the Thursday column left blank. The direction was then given, "Litterbug Doug has now collected litter in Central Park for 3 days and the townsfolk are keen to see how much he has collected." In their groups, the children were to explore the data, first noting the numbers of items collected on the second and third days, then how the data varied across the first 3 days and why this might be the case. Finally, the children were to consider the blank Thursday column and predict how many different items Litterbug Doug might have collected on that day.

Baxter Brown's Picnic adopted a similar format to *Litterbug Doug*, with the central character, Baxter Brown, inviting his five canine friends to each bring six different food items to a picnic. The numbers of item types each dog brought were displayed in a table, with the last column ("Oinkers") left blank. After discussing what they observed about the data, the children were to predict the number of Oinkers each dog might have taken to the picnic. The next component of the activity, not addressed here, involved the children in planning a class picnic, posing and refining questions for investigation, and collecting and representing their data in their own way.

Selection of Findings

Data "lenses," from the unpublished work of Konold, Higgins, Russell, and Khalil's (2004), provided a valuable framework for analyzing the children's responses to the tables of data in the *Litterbug Doug* and *Baxter Brown's Picnic* activities. Specifically, the *case values*, the *classifiers*, and the *aggregate* lenses, with the latter modified to a *pre-aggregate* lens for this study (English, 2012) provided insights into the children's developments in identifying variation in the data and predicting the missing column values. For the most common lens, that of *case values*, the unit of analysis is an individual case with the analysis focusing on a consideration of the values of particular cases. Examples include identifying items with the same values (e.g., "There are three 3s," referring to 3 cans and 3 cheeses on Tuesday, and 3 newspapers on Wednesday); comparing item values across rows (e.g., "And that's a pattern, 1, 2, 1, 2;" "It's little, big, little, except for this one); and totaling across rows or totaling both rows and columns.

Use of the more sophisticated lens, that of *classifiers*, considers the frequency of cases with particular values, without attention to the data collection as a whole. Examples include, "He collected more apple cores on Tuesday and he collected less on Monday;" "On Monday he collected more drink cans than Wednesday;" and "Daisy wants four Dentastix and the rest want lower numbers than her." The final lens, the most difficult, is the *aggregate lens*, where the entire distribution of values is the perceptual unit. The use of a modified, *pre-aggregate lens* was suggested in responses where all of the data in the table are considered and frequencies compared and/or trends noted, for example, "Monday and Wednesday are both the same but the rows are not the same; not the same in numbers" (Comparing column totals and values across the rows); and "Well, first he didn't find that much cause it was his first day. And then he knew more so he found more and then he found so much that he couldn't find that much so it went down again" (referring to column totals and applying contextual knowledge in doing so).

Children's analysis of the data and their predictions for the blank column often involved switching between lenses, suggesting an informal awareness of the range and variation in the existing data, especially in the *Litterbug Doug* activity. Here, 12 out of 13 groups recorded predictions of values ranging from 0 to 10, with all but one of these 12 groups explicitly recognizing that wild outliers (e.g., 56, 45) would be unlikely (English, 2012). In *Baxter Brown's Picnic*, however, it appeared that the task context was an inhibiting factor in the children's ability to abstract the data from the problem context (Konold & Higgins, 2003). For example, a prediction of a large outlier was considered inappropriate because too much food would make the dogs sick. On the other hand, some responses did indicate an awareness of variation in the existing values such as predicting that Baxter Brown would be unlikely to bring zero Oinkers, "because he's pretty much of a greedy guts so I think he would have more" (a decision based on the existing item values for Baxter Brown).

CHILDREN'S METAREPRESENTATIONAL COMPETENCE

To illustrate the metarepresentational competence children displayed in the third year, examples are taken from the activity, *Investigating and Planning Playgrounds*. Following a story book reading, each group was to discover more about their classmates' views on their new school playground by creating four survey questions and four answer options for each question. On answering their own questions, each group chose one key question to which the other groups were to respond. The children were to initially predict how their key question might be answered by the remaining groups. Each group then analyzed all the data for their key question and were to display their findings in more than one way using their choice of representation. A range of recording material was provided including blank chart paper, chart paper with a drawn circle, and grid paper. The children could choose whatever material they liked; no direction was given here.

Selection of Findings

A range of multiple representations was created, with bar graphs and circle graphs the most popular. Unusual representations that drew on the children's real-world observations were also created by two groups, such as a “heart monitor” representation. As one of the groups explained, “Okay, how are we going to show it our own way?! Let’s think, just think. I know, like at the hospital, so like... up here (indicates a line graph similar to a heart rate monitor display). So first we’re going to maybe do about 3cm” (child placed his ruler horizontally across the middle of the paper). Another group member further added, “We’re doing like a doctor’s sort of thing, how they go like that (indicating a rise and fall)...like a doctor does...yes, how they have those lines, so we’re sort of doing it with maths though.”

Given that the children had not been taught how to construct circle graphs, it was interesting to observe how they were developing both conceptual and metarepresentational competence in creating these representations. For example, each of the focus groups made use of a ruler and/or estimation in their efforts to represent their data. One focus group argued over how to estimate a sector for each response option, with one child insisting that “You have to find the middle first. That’s the first thing you actually do.” He then placed his ruler through the centre of the circle and drew a small sector to represent the two “for exercise” responses to the focus question, “Why do you like the equipment you chose?” His explanation was, “Two will only be like this (drawing a small sector)...cause it’s a very small amount.” He then recorded “2” in the sector. When asked how many “pieces of the pie” was needed, the group quickly replied “four, cause there’s four of them (response options).” The group also noted that the four sectors would not be the same size “because if there’s two people, this would have to be a smaller piece to fit two people and a bigger piece to fit nine people in.”

A particularly interesting finding was the children's independent, emerging understanding of percent, which they had experienced in the second year when one class shared one of the activities with a grade 7 class. In reporting on their representations, the grade 7 students explained how they used percentages in constructing their circle graphs. The grade 2 class was fascinated by this approach and transferred their learning to the third year. Furthermore, those classes who had not shared the grade 2/7 experience learned informally about percent from their peers in the third year. Although the children had developed only a primitive understanding of percent, which was well beyond their curriculum level, they nevertheless displayed an awareness of its nature and role. As one example, five groups of children received zero preferences for one of their response options; this generated substantial debate on how to display 0% and also, 1%. One group was experimenting with how to represent their data on their circle graph for the response options to the question, “Why do you like the playground?” The option, “fits more people than the oval” received zero votes, while the option, “good views” received one preference. In creating their circle graph, one group member suggested, “Maybe you could rule a little bit off it…that could be zero percent.” She further noted that the response option of good views “only has one percent…” and “has to be really small, like that small.” This group further struggled with their display of 0%, claiming that there was insufficient space to label the option of "fits more people than the oval." When asked by their teacher, “How can you show 0%?” the child responded that, “You should just rub that out…cause that got nothing.”

CONCLUDING POINTS

This paper has provided some examples of the statistical foundations young children are capable of developing and sustaining through data modeling. Although the reported findings are limited to a small sample, they nevertheless illustrate young children's capabilities in drawing informal inferences and developing metarepresentational and conceptual competence earlier than expected or indeed, considered feasible. The findings support diSessa et al.'s (1991) arguments that young children's skills are more “broadly applicable, more flexible and fluid” (p. 118), and are not just confined to a narrow set of instructed representations. Providing children with opportunities for structuring and displaying data in ways that they choose and understand is essential, even though such representations might initially appear meaningless to us. Likewise, experiences in drawing informal inferences from meaningful data are needed, including abstracting beyond the problem context. Establishing these important foundations of variation, prediction, and uncertainty from the earliest grade levels is critical for more advanced statistical reasoning.

REFERENCES

Batanero, C., Burrill, G., & Reading, C. (Eds.) (2011). *Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Dordrecht: Springer.

Bethel, E. (2009). *Litterbug Doug*. Mascot, Australia: Koala Books.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly, 104*, 801-823.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks: Sage Publications.

diSessa, A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction, 22*(3), 293-331.

diSessa, A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Metarepresentational expertise in children. *Journal of Mathematical Behavior, 10*, 117-160.

English, L. D. (2013). Reconceptualising statistical learning in the early years. In L. D. English & J. Mulligan (Eds.), *Reconceptualising early mathematics learning* (pp. 67-82). NY: Springer.

English, L. D. (2012). Data modeling with first-grade students. *Educational Studies in Mathematics, 81*(1), 15-30.

English, L. D., & Watson, J. M. (2013). Beginning inference in fourth grade: Exploring variation in measurement. In V. Steinle, L. Ball & C. Bardini (Eds.), *Mathematics education: Yesterday, today, and tomorrow* (pp. 274-281). Mathematics Education Research Group of Australasia.

Franklin, C. A., & Garfield, J. (2006). The GAISE project: Developing statistics education guidelines for grades pre-K-12 and college courses. In G. Burrill & P. Elliott (Eds.), *Thinking and reasoning with data and chance* (68th Yearbook, pp. 345 –376). Reston, VA: NCTM.

Kelly, A. E., Lesh, R. A., & Baek, J. Y. (Eds.) (2008). *Handbook of design research methods in education*. New York: Routledge.

Kinnear, V. (2013). *Young children's statistical reasoning: A tale of two contexts*. Unpublished PhD dissertation. Queensland University of Technology.

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In Kilpatrick, J., Martin, W. G., & Schifter, D. (Eds.), *A research companion to Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Lehrer, R., Kim, M., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM Mathematics Education, 43*, 723-736.

Lehrer, R., & Lesh, R. (2003) Mathematical Learning. In W. Reynolds & G. Miller (Eds.), *Comprehensive handbook of psychology* (Vol. 7, pp. 357 –390). New York: John Wiley.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149-176). NY: Taylor & Francis.

Lesh, R., & Lehrer, R. (2000). Iterative refinement cycles for videotape analyses of conceptual change. In R. Lesh & A. Kelly (Eds.), *Research design in mathematics and science education* (pp. 665–708). Hillsdale, NJ: Lawrence Erlbaum Associates.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The Reasoning Behind Informal Statistical Inference. *Mathematical Thinking and Learning, 13*(1/2), 152-173.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher, 42*(1), 30-37.

Pfannkuch, M. (2005). Thinking tools and variation. *Statistics Education Research Journal, 4*(1), 83-91.

Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts in statistics. In G. Burrill & P. Elliott (Eds.), *Thinking and reasoning with data and chance* (pp. 77-98). Reston, VA: National Council of Teachers of Mathematics.

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