# Insights into the approaches of young children when making informal inferences about data

Aisling M. Leavy

Mary Immaculate College, University of Limerick, Ireland; Aisling.Leavy@mic.ul.ie

Finbarr Sloane

National Science Foundation, Arlington, Virginia, USA; fsloane@nsf.gov

*There is growing awareness of the statistical reasoning abilities of young children. In this study the informal inferential reasoning skills of a class of 5-6 year old children are examined as they reason about data in the context of a week-long data investigation unit. The strategies young children use to make predictions about data are identified. A discussion ensues around what these strategies communicate about early understandings of statistical inference. The findings suggest that making inferences from data can be challenging for younger students primarily due to the powerful influence of their developing understandings of number. However, there is evidence that children possess some of the building blocks of informal inference most notably in the approaches that point to a pre-aggregate view of data. Situating data investigations within interesting and relevant contexts, alongside good teacher questioning and opportunities to listen to the reasoning of their peers, contributes to the creation of statistical environments that support and develop early understandings of inference.*

*Keywords: Informal inferential reasoning (IIR), early childhood, data modelling.*

## Theoretical Perspective

Statistical inference involves drawing conclusions that extend beyond a given set of data. This may involve inferring properties of a population based on a random sample selected from that population, or using inferential statistics to ascertain whether differences between groups are due to some systematic influence other than chance. In recent years, the importance of introducing younger students to the fundamental notions of statistical inference has been advocated by curriculum bodies and researchers alike (Ben-Zvi, 2006; English 2010, 2012; Leavy 2008). In this study, efforts are made to characterize the nature of statistical understandings *prior* to formal work with statistical inference, i.e., the foundational ideas, more recently referred to as 'informal inference' (Rubin, Hammerman & Konold, 2006, p. 1).

Informal ideas relating to inference are those understandings that are foundational to the development of inferential reasoning. While many different definitions of informal inference have been posited, a useful definition of informal inference is "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (Zieffler, Garfield, delMas & Reading, 2008, p. 44). Zieffler et al. (2008) identify three components of an IIR framework as: making judgments or predictions, using or integrating prior knowledge, and articulating evidence-based arguments. Arising from research with primary students, Makar & Rubin (2009, p. 85) propose three principles that are considered essential for informal statistical inference as '(1) *generalization*, including predictions, parameter estimates, and conclusions, that extend *beyond describing the given data*; (2) the use of data *as evidence* for those generalizations; and (3) employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn.'

One statistical perspective identified as a necessary building block to form a basis for IIR is the ability to *view data as an aggregate* (Rubin, Hammerman & Konold, 2006). Statistical properties of aggregates such as their centers, variability and shapes emerge from attending to features of distributions rather than features of individuals. Thinking about aggregates, while possible, has been shown to be challenging for children (Cobb 1999; Hancock, Kaput & Goldsmith, 1982). Recent work by Konold, Higgins, Russel & Khalil (2015) has resulted in the identification of four perspectives that students use when working with data. The use of these perspectives as a way to analyse an individual's particular interpretation of data may provide valuable insights into their statistical reasoning and in turn the extent to which they possess the necessary building blocks for informal inference. These perspectives include *data as pointer* (focus on the event rather than the data), *data as case value* (focus on individual data values or cases), *data as classifier* (identifying subsets of data values that may be the same or similar) and *data as aggregate* (view all the data values in aggregate as an "object" or a distribution).

A study of first-grade student's data modelling approaches carried out by English (2012) categorised children's predictions and approaches when working with data using the lenses identified by Konold et al. (2015). Using this framework to guide categorization, 6-year olds in English's study viewed data in a variety of different ways. Many children focused on the values of particular cases (*case value lens*) and others demonstrated the ability to consider the frequency of cases with a particular value (*classifier lens)*. There was also evidence of what English (2012) terms a *pre-aggregate lens* which included approaches that considered all the data, compared the frequencies and had some attention to overall trends. While not as sophisticated as an *aggregate lens*, which involves consideration of the entire distribution as an entity in itself, the presence of this pre-aggregate lens is a strong indicator of the nascent potential of young children to engage in informal inferential reasoning.

A number of studies have explore the reasoning abilities of young children when engaged in data modeling activities in environments supported by the use of picture books (English 2010, 2012; Kinnear 2013, 2016) and data-visualization tools and technologies (Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008). This study continues this line of research by exploring the informal inferences young children make about the data presented in a data modelling environment and examining what these inference tell us about children's perspective on data.

## Method

A multitiered teaching experiment (Lesh & Kelly, 2000) was carried out with twenty-five 5-6 year olds as they engaged in a weeklong data modeling activity. Statistical activity was motivated by a driving question (Hourigan & Leavy, 2016) and the context was the 'design of a zoo' as it was familiar to children and incorporated opportunities to work with data, encourage exploration of variation and make predictions about the data. The Ertle, Chokshi, & Fernandez (2001) lesson note format, developed for use as part of Japanese Lesson Study, guided lesson design considerations. This framework promoted a focus on expected student reactions, concomitant teacher responses and evaluation strategies. These foci supported examination of students' ability to engage in IIR. This study examines the final lesson which focused on making informal inferences about data.

The inquiry was stimulated by playing a video excerpt that we produced:

> Hi, I am James the zoo keeper. The elephant's home in the zoo is getting a little bit crowded. I think we need to make it a bit bigger. But, I don't know how many elephants will be in the zoo next year which makes it difficult to plan ahead. I was hoping you could look at the numbers of elephants in the zoo for the last four years, and predict how many will be there next year?

Children were then shown live video feed of the elephant enclosure in Dublin zoo and presented with a table of data illustrating the number of elephants born in the first year (3 elephants), second year (4 elephants), the third year (7 elephants) and fourth year (5 elephants) (see figure 1). Children worked in groups of 5 to reason and make predictions about the number of animals born the following year (year 5). Following the predictions, other data relating to the birth rates of wolves [5, 6, 2, 3], giraffes [8, 8, 5, 5] and monkeys [3, 5, 0, 2] across four years were presented. Children worked in groups and predicted the number of respective animals born in the fifth year. The design of these tasks was informed by the Zieffler et al. (2008) framework to support IIR by challenging students to make predictions and judgments about data and by incorporating opportunities to capture students' informal inferential reasoning.

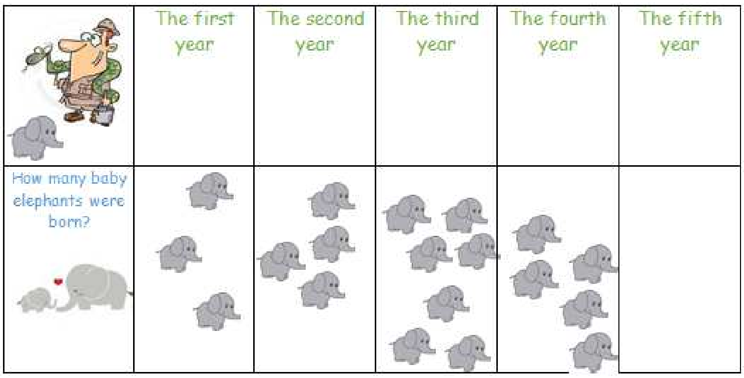


**Figure 1: The 'elephant birth task'**

Conversations in 4 of the groups were audio recorded and one group was video recorded. Our primary focus when analyzing the data was on identifying the ways in which young children make informal inferences in a context rich data modelling environment.

Recordings of group conversations were transcribed. Transcripts were coded according to whether they embodied Makar & Rubin's (2009) principles of IIR. Thus each transcript was coded at least three times in an effort to identify children's ability to generalize beyond the data, to use data as evidence and to use probabilistic language. All predictions were further categorized as representing data as pointer, case value, classifier or aggregate perspectives on data (Konold et al., 2015).

## Findings

Children understood the task and were enthusiastic when making predictions about animal births in Year 5. However, making data informed predictions was challenging for some. Initially there was some evidence of idiosyncratic reasoning that was distanced from the context and from the data presented, however, this soon disappeared once the data and context were discussed further. Many children based their predictions on their knowledge of the context and modified their prediction based on discussion with peers. The findings are structured using the three principles of IIR that are considered essential for informal statistical inference (Makar & Rubin, 2009).

### Principle 1: Generalizations beyond the data

While all children made predictions regarding the number of births, not all of the predicted values indicated an ability to generalize beyond the data. Rather, they reflected the influence and power of counting in the mathematical development of the young child. For these children, there was an awareness of frequencies and this was demonstrated in the tendency to list the numbers, order them and then compare the outcome to the counting numbers. This focus on the frequencies resulted in two approaches to predicting births. The first approach was to *fill in the gaps.* Children compared the frequencies to the sequence of counting numbers usually leading to the identification of a 'gap' in the list of numbers. Children were eager to fill this gap i.e. identify a count/frequency that hadn't occurred in the presented data and avoid presenting a value that had already occurred. Thus, they believed that this missing number would likely be the number of animals born the next year (see discussion between Sheena and Ayesha below around wolf birth rate [5, 6, 2, 3]). The second approach was to *extend the number sequence.* In these situations children were not overly perturbed by an identified gap in the counting sequence and chose instead to extend the numbers beyond the range of the presented data (see Kate below). Generally, the next highest counting number above the upper value of the range was their prediction for the number of births in year 5. Both these strategies indicate a focus on pattern in the sense of ordinal counting numbers and thus the 'power of counting'. However, from a statistical sense the reasoning was located and justified within the world of counting numbers thus indicating a lack of focus on pattern and trends in the data.

Teacher: How many wolves will be born this year (pointing to year 5)?

Sheena: We say maybe 4 cause 5, 6, 2, 3. And there's no 4.

Ayesha: We've got our reason. 2, 3, 4, 5, 6. It's 4 cause 2, 3, 4, 5, 6.

Teacher: What would happen if there was a year 6? How many animals might be born then?

Kate: 1. Cause it would start 1, 2, 3, 4, 5, 6.

**Principle II: Using data as evidence**

Analysis of the transcripts revealed an abundance of situations when children *used data as evidence* to support their predictions and conclusions about data. The explanations provided by children were categorized as falling within one of the four perspectives on data posited by Konold et al. (2015).

**Observation 1: The prevalence of a *case value lens***

The focus on individual data values indicated the presence of a *case value lens.* In particular, children were attuned to the appearance of zero births for year 3 in the monkey data [3, 5, 0, 2] and commented 'there were none that year' and 'there are zero there'. While this case value lens indicates a lack of focus on the aggregate, the individual data values were considered within the greater data context. For example, Eva drew on her knowledge of the context in her efforts to explain why no monkeys were born in year 3 when she stated 'because if they had too many babies there [pointing to the 5 born in year 2], the mommy babies would have to rest all day'.

Young children's approaches involving summing data values and calculating totals have been used as indicators of a case value lens (English, 2012). Similarly, in this study, several predictions of the births in year 5 also indicated a case value lens as they were based on summing all or some of the values and presenting this total as the prediction for year 5. Matthew predicted 'I decided there will be 10 monkeys altogether born' based on summing the births in years 1-4, and Kornelia predicted that 16 wolves would be born in year 5 'because I counted all of them'. This difficulty in attending to the variation in the data was also evident in another child's response that 10 giraffes [8, 8, 5, 5] would be born because '5+5 makes 10'.

However for many of these same children, while there was a focus on counting and the application of a *case value lens,* there was an awareness of pattern in magnitude of numbers. When large numbers were presented as predictions, children rejected these numbers as too big and drew on contextual information to justify their reasoning. In the following segment, children are predicting the number of giraffes [pattern: 8, 8, 5, 5] that will be born in year 5.

Thomas: I think 85. Because there is an 8 there and a 5 there.

Polina: They wouldn't fit into the box. They are definitely not going to fit into the zoo also.

Teacher: Really? Why do you think that?

Polina: Because they [giraffes] are very big.

**Observation 2: Awareness of trends in the data and evidence of a *pre-aggregate lens***

The responses of 20% of children suggest an awareness of overall patterns and trends in the data. This was termed a *pre-aggregate lens* by English (2012) and may point to some emerging sense of distribution. The awareness of pattern was evident in Mia's response to the wolf births which she described as 'going up and going down'. Similarly the recognition and subsequent extension of a repeating pattern in the giraffe data set [8, 8, 5, 5] was evident when Melios predicted '8. It's 8, 8, 5, 5, 8. Cause it's a pattern: 8, 8, 5, 5, 8'. During a whole class discussion about the number of monkeys that would be born in year 5, awareness of patterns was evident in the comments from Otille and Kate below:

Teacher: How many monkeys did you think were born in year 5? [3, 5, 0, 2]

Otille: I think 1 because it goes down, up, down, up, down.

Kate: 5. Cause 5 here [points to 5] and then low [points to 0 and 2] so it would go back to high.

As can be seen from the transcripts above, children's justifications did not explicitly refer to the context of the data (in this case birth rates) and hence there is the possibility that this awareness of trends stemmed more from an algebraic rather than statistical perspective. However, the greatest indication of the presence of a *pre aggregate lens* was in the reasoning of those children who married an awareness of trends in the data with their understandings of the context in constructing their predictions. In the discussion of the trends in elephant births [3, 4, 7, 6] Polina imagined that animals born in year 1 would have grown up by year 5 and be giving birth to elephants in year 5.

Polina: We put 8 elephants (born in year 5)

Teacher: Why did you put 8 in?

Polina: Because I think these are going to grow up [pointing to the 3 elephants born in year 1] and these ones will be in their tummies [pointing to her prediction for year 5]. It is always going to get bigger.

Teacher: So do you think it will always get bigger?

Polina: Yes, I think so, I think there will be babies born from these ones. These ones are going to be all grown up, they will be adults.

Thus her understanding of the variation in the data influenced her predictions and ensured she always predicted beyond the upper range of the presented data. Another child, Anna, demonstrated her ability to view the trends across the years and used this trend to inform her initial prediction. However, she subsequently used her knowledge of the context and adjusted her prediction downwards. Here, Anna is discussing her prediction for the number of giraffes born [8, 8, 5, 5] and her initial prediction of '3' may indicate some developing notion of center. However her attention to the context makes her mindful of how her prediction (if it were correct and acted upon) would affect the other animals in the zoo and she adjusts her predict to 'protect' other animals from the negative outcomes arising from her prediction.

Anna: It was different on different years sometimes 5 [pointing to the values for year 3 and 4] but here and here it was 8 [pointing to the year 1 and 2]. So I think 3.

Teacher: Why 3?

Anna: Cause it is like the others. Not too many (baby giraffes) but not none (baby giraffes).

No. No. I think 2. Because if there are too many, all of the branches and the leaves would be gone and there would be no place for a monkey.

**Principle III: Use of probabilistic language**

Makar & Rubin (2009) emphasize the importance of expressing uncertainty when making inferences – this can be identified in efforts to avoid deterministic claims and in the use of probabilistic language. Analysis of the transcripts reveal that children drew conclusions based on the data presented to them (birth rates over time) and used this data to make predictions beyond the data. All the while they were articulating uncertainty as demonstrated in their use of terms such as 'I think' (see Thomas, Otille, Polina and Anna above), 'probably' 'maybe' (see Sheena above) and 'I'm not too sure'. It is particularly interesting to note that children were comfortable with uncertainty and with the different predictions of others. This openness was evident when Eva pointed out 'we don't know' in relation to how many elephants the mother elephant would have. Her partner Paul continued the reasoning and stated that 'maybe there would be six elephants born because there are 6 elephants there and they could have 6 babies'.

## Conclusions

Young children in this study demonstrated the seeds of informal inference in their ability to 'look beyond the data' and engage in data-based argumentation to support their predictions. However, making data-based predictions was a challenging task for some children. Case value perspectives were most prevalent. The lack of repeating data values in the presented data may account for the low incidence of classifier perspectives as compared to the study by English (2012). Similar to English's study there was evidence of the presence of a pre-aggregate lens in the approaches taken by children. A large proportion of children scanned the data for patterns, sought 'missing numbers' and many made predictions based on patterns in the ordered lists of data rather than thinking from a statistical perspective. This reliance on number and algebraic reasoning is not surprising given the curricular emphases in early years mathematics curricula. It is interesting to note the influence of zero on children's deliberations about data was also a factor in the work of Kinnear (2013) and Kinnear & Clarke (2016) when engaging young children in data modeling activities.

The success that some children experienced in making informal inferences was due to a number of factors. The role played by the data and task context is particularly evident. The use of an interesting and relevant context provided a 'crutch' for the children when making predictions. Their personal experiences and high task interest ensured that rather than reasoning about decontextualized data, children were reasoning about and making sense of the situation at hand – this supported their inferences. Secondly, the development of skills in making data-informed predictions was due in large part to the use of good questioning on the part of teachers and due to their efforts in drawing children's attention to aspects of the data and clarifying misunderstandings as they occurred. Similarly, the work of Paparistodemou & Meletiou-Mavrotheris (2008) highlighted the important role that prompting by the researcher played in supporting children in speculating about larger data sets. The third factor was the importance of peer interactions. Children built on the ideas of others as they reasoned and made prediction within their groups thus providing evidence for the power of co-constructing meaning in small groups and demonstrated 'building on the ideas of others' (Whitin & Whitin, 2008, p.93). This importance of peer interaction in promoting inference and deriving conclusions from data was also evident in the work of third grade students when engaging in inference (Paparistodemou & Meletiou-Mavrotheris, 2008).

## References


Ben-Zvi, D. (2006). *Scaffolding students' informal inference and argumentation.* In A. Rossman & B. Chance (Eds.), Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf]

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5–43.

English, L.D. (2010). Young children's early modelling with data. *Mathematics Education Research Journal, 22*(2), 24–47

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

Ertle, B., S. Chokshi, and C. Fernandez. 2001. *Lesson Planning Tool.* Retrieved from www.tc.columbia.edu/centers/**lessons**tudy/doc/**Lesson_Planning_Tool**.pdf *on November* 28, 2016.

Hancock, C., Kaput, J. J., & Goldsmith, L.T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337–364.

Hourigan, M. and Leavy, A.M. (2016). Practical Problems: Introducing Statistics to Kindergarteners. *Teaching Children Mathematics, 22*(5), 283-291.

Konold, C., Higgins, T., Russell, S.J & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*, 305-325.

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

Kinnear, V. & Clarke, J. (2016). *Young Children's Abductive Reasoning about Data.* Proceedings of the13th International Congress on Mathematical Education.  Hamburg, July 24-31.

Leavy, A.M. (2008). An examination of the role of statistical investigation in supporting the development of young children's statistical reasoning. In O. Saracho & B. Spodek (Eds.) *Contemporary Perspectives on Mathematics Education in Early Childhood* (p. 215-232). Information Age Publishing.

Lesh, R., Kelly, A., (2000) Multitiered Teaching Experiments. In A. Kelly, R. Lesh (Eds.), *Research Design in Mathematics and Science Education.* (pp. 197-230). Lawrence Erlbaum Associates, Mahwah, New Jersey

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf]

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Enhancing Reasoning about Statistical Inference In 8 Year-Old Students. *Statistics Education Research Journal, 7*(2), 83-106.

Rubin, A., Hammerman, J. K. L, & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*. Salvador, Brazil. [CDROM]. Voorburg, The Netherlands

Whitin P., & Whitin D.J. (2003). Developing mathematical understanding along the yellow brick road. *Young Children, 58*(1), 36–40.

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. [Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf]