What sense do children make of "data" by Year 3?

Jane Watson
University of Tasmania
<Jane.Watson@utas.edu.au>

Noleine Fitzallen
La Trobe University
<N.Fitzallen@latrobe.edu.au>

Statistical terms are used in everyday language and, at times, used in non-statistical ways. It is often assumed students understand statistical terms because of their common use; however, research into their understanding of specific statistical terms is scant. This report focuses on 58 Year 3 students' responses to the basic question, "What does the term 'data' mean?", and associated examples of data and data representations. The results indicate students are making progress in establishing meaning about data and their representations. Recommendations include more use of varying contexts within which students can explore data to enrich and enhance their learning about the practice of statistics.

Statistics in school curricula in Australia dates to the *National Statement for Mathematics in Australian Schools* (Australian Education Council, 1991) following the US National Council of Teachers of Mathematics' (NCTM) publication of its *Curriculum and Evaluation Standards* (1989). Neither of these documents, nor the later *Principles and Standards for School Mathematics* (NCTM, 2000) or the *GAISE Report* (Franklin et al., 2007), defines the term "data". Indeed, the focus in early childhood professional learning for teachers has been on representing data and not specifically on defining the term data (e.g., Schwartz & Whitin, 2006). Reflecting this background, in the early years children are often introduced to activities that involve collecting and representing data (e.g., Taylor, 1997), apparently with the assumption that by giving them many examples of data, they will eventually "understand" what data stand for and what the term means. Russell (2006) claims that "[t]o understand what data are and how to use them, students must themselves be engaged in developing questions about their world and creating data to shed light on those questions" (p. 17) but does not go so far as to define the word. She stresses the importance of *creating data* by noting the connections data allow and the reason for their existence: "Data are not the same as events in the real world, but they can help us understand phenomena in the real world" (p. 17). In the adult world, Moore and McCabe (1989) define statistics in relation to defining data: "Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data*" (p. xvii). Cobb and Moore (1997) go further in claiming that "Statistics requires a different *kind* of thinking, because *data* are not just numbers, they are numbers with a context" (p. 801). This statement complements well Russell's (2006) linking data to events in the real world.

New Zealand was likely the first country to define data, in its *Mathematics in the New Zealand Curriculum* (Ministry of Education, 1992): "*Data* A set of facts, numbers, or information" (p. 211). In Australia, the development of the most recent *Australian Curriculum: Mathematics* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2019a) began in 2010 and currently includes a definition of data as, "*Data* is a general term for information (observations and/or measurements) collected during any type of systematic investigation." The inclusion of "systematic investigation" in this definition adds a third element, "context", to "information" and "collection," for making meaning of data. From Year 1 of the curriculum, the word "data" appears with the creation of representations with objects or drawings, along with descriptions of displays (Year 1, ACMSP263). By Year 2, students are gathering, checking, classifying, creating displays, and interpreting categorical data for a question (ACMSP048, ACMSP049, ACMSP50). In

2021. In Y. H. Leong, B. Kaur, B. H. Choy, J. B. W. Yeo, & S. L Chin (Eds.), *Excellence in Mathematics Education: Foundations and Pathways (Proceedings of the 43rd annual conference of the Mathematics Education Research Group of Australasia)*, pp. 409-416. Singapore: MERGA.

Year 3, the representations may be lists, tables, picture graphs or simple column graphs (ACMSP069, ACMSP070). Although the word "context" is not used in the content descriptors, the act of interpreting data implies that students will link data with the representations created and contexts within which the data were collected.

The other source of curriculum input on data comes from the Numeracy component of the *General Capabilities* section of the *Australian Curriculum* (ACARA, 2019a). As part of Numeracy in that document, "Interpreting statistical information" is one of six interrelated capabilities in the learning continuum. "This element involves students gaining familiarity with the way statistical information is represented. Students solve problems in authentic contexts that involve collecting, recording, displaying, comparing, and evaluating the effectiveness of data displays of various types" (ACARA, 2019a). In addition, although the Achievement Standard of the *Australian Curriculum* for Year 3 includes "conducting simple data investigations for categorical variables," the Numeracy Capability reduces the element of "Interpreting statistical information" to "interpret data displays" for all year levels. It is not until the end of Year 6 that students are expected to evaluate or analyse data representations. Prior to that, students are only expected to be able to collect, record, and display data.

Research into children's early understanding of data and their representation has focused on using data in contexts meaningful for children (e.g., Russell, 1990) without asking for a description or definition of the word itself. Similarly, Fitzallen (2012) analysed children's early appreciation of data in the context of graphing and analysis, without asking specifically about the word itself. An extensive search of the research literature found no instances where children were asked the meaning of "data." Given the importance of the term and its definition in the Glossary of the *Australian Curriculum* (ACARA, 2019a), it seems appropriate to ask this question of students.

The results reported in this paper are drawn from the beginning of a four-year teaching intervention related to studying the impact of making data the focus of learning activities, with the goal of enhancing the emerging STEM curriculum (Fitzallen & Watson, 2020). The student learning activities in the study were grounded in the concepts imbedded in the Practice of Statistics (Watson et al., 2018), which encapsulates all aspects of working with data: formulating questions, collecting data, analysing data, and interpreting results. At the beginning of the longitudinal project, Year 3 students were asked to respond to the item in Figure 1 as part of a pre-test of students' initial understanding related to the goals of the project. In retrospect, however, it also provided the opportunity to monitor the implementation of the *Australian Curriculum* definition of "data" and expectations for creating representations from data over the previous 3+ years of schooling (Foundation to Year 2 and half of Year 3). The research question hence becomes: *How well do Year 3 students understand data in relation to the Australian Curriculum's definition of "data", and its expectations related to data displays?*

Survey Questions about data
(a) What do you think "data" means?
(b) Give an example of some data you have seen or collected.
(c) Sketch a graph of the data.

Figure 1. Survey item for Year 3.

Method

For the research question asked in this report, a survey method using open-ended questions is appropriate to obtain the required data. Ballou (2008) suggests open-ended questions provide the opportunity to gain insights into how terms are understood, and ideas are developed. The three tasks in Figure 1 solicit qualitative data related to students' understanding of the topic of interest: a basic appreciation of the meaning of data and how they might be represented. The item was included in an eight-item survey. The other items focused on visual representations, sampling, and questioning.

Participants

Fifty-eight students from two Year 3 classes in a parochial K-10 school in an inner regional centre with a socio-economic status index ICSEA value of 1026 (mySchool.com.au; mean of 1000 and standard deviation of 100) were surveyed: 33 boys and 25 girls, 8-9 years of age. At the time students completed the survey, it had been two months since the NAPLAN testing for Year 3 had taken place. At the time of the survey, the researchers had no background on the teachers or the students. In terms of the results on NAPLAN testing nationally, the Year 3 cohort in this study was in the Average range for Reading, Writing, Grammar, and Numeracy, and in the Above Average range for Spelling (ACARA, 2019b). These results, and the fact that the teachers and students had no content interaction with the researchers prior to the survey, suggest that the sample can reasonably be assumed to be only marginally above average for Australian Year 3 children at this time in their education. The project had approval of the Tasmania Social Sciences Human Research Ethics Committee (H0015039).

Data Analysis

Due to the cognitive nature of mathematics learning, the method of analysing the data involved characterising similar responses with relation to a learning theory. The hierarchical model chosen was the *Structure of Observed Learning Outcomes* (SOLO) model (Biggs & Collis, 1982). The SOLO model has been used across the field of mathematics education for many years to analyse what respondents say or write (e.g., Watson, 2001) and continues to be useful in statistics education (e.g., Groth et al., 2019). For the survey questions in Figure 1, responses are expected to occur within the Concrete Symbolic (CS) mode, typical of students in the primary and middle years (ages 7 to 12 years). The levels are: Unistructural (Uni), where a single element or idea is presented; Multistructural (Multi), where responses include two or more elements presented in a serial fashion; and Relational (Rel), where responses describe links or relationships among the elements presented. Responses judged not to use elements involved in the task, including no response, are often labelled Pre-structural, but here they are examined from Groth et al.'s perspective, which considered in more detail the Ikonic mode (IK) for evidence of response compatibility (*c*) or incompatibility (*ic*) with the context of the task. Incompatible responses include superstitious, subjective, or deterministic beliefs, whereas compatible responses include personal experiences, imagery, or intuition related to the context of the task. It is hence of interest to observe responses considered to be in the IK mode for their compatibility as a step toward the CS mode. Using this structured analysis of students' responses, it is possible to suggest the degree to which a sample of children have had access to and taken on the goals of the curriculum in relation to "data", introduced by the middle of Year 3.

The coding scheme based on the SOLO model was designed to reflect the three components in the definition of "data" (information, collection, and context), the complexity of the example described, the representation created and its completeness, including the link between the representation and the context in the example. The elements that were considered appropriate for a definition of data (Figure 1, Part a) included a word interpreted as an appropriate synonym for "information" at the Year 3 level, a word related to the process of "collecting" information, and a word or phrase suggesting a meaningful context (systematic investigation) for information to be collected. Providing single elements was classified as Uni; putting two together as Multi; and combining all three in a meaningful sentence as Rel. For the example of data given (Figure 1, Part b), a single suggestion of a "variable" was considered Uni, whereas if it were connected with a second variable, it was considered Multi. This question did not lead to the expectation of a Rel response. With respect to Part (c), the representations were categorised according to the three representations noted in the content descriptor for Year 3 of the *Australian Curriculum* (ACMSP069): pictographs, tables, and column graphs. Within each type of graph there were increasing levels of combining the elements required to construct the representation. For pictographs, a picture without labels or categories was considered IK. Supplying categories but no variation in represented data was Uni, whereas displaying variation across categories was Multi. For both Tables and Column Graphs, an incomplete representation or no labels added was considered IK, whereas Uni or Multi representations included either one or two, respectively, of the essential components of the entity. For Tables the components were tallies and totals and for Column Graphs they were one or both axes meaningfully labelled including the column bars. Given the way that the questions were linked, if a complete pictograph, table, or column graph was labelled to reflect the context of the example suggested in Part (b), the response was considered Rel. Given the expectations of the content descriptors and the definition of "data" in the *Australian Curriculum* (ACARA, 2019a), Table 1 outlines the SOLO response levels for the three questions asked of the students. The coding was initially completed by the first author and repeated separately by an experienced research assistant. Agreement was 83% with discrepancies decided by negotiation.

Table 1
SOLO Levels of Response to the Three Parts of the Survey Item on Data

Level	Part (a)	Part (b)	Part (c) Sketch a graph		
	Defining data	Example data	Pictograph	Table	Column Graph
IK (*c*, *ic*)	Idiosyncratic; self-reference; specific case	Idiosyncratic; source; tallies with no categories	Icons without labels or categories	Table without information	Incomplete/no labels; unconventional
CS Uni	Single element mentioned	Single aspect of process or context	Icons in categories not showing variation	Labelled categories with either tallies or totals	Column graph with one dimension meaningfully labelled
CS Multi	Linking information either to the process or context	Clear summary of data or process with context	Icons in categories displaying variation	Labelled categories with both tallies and totals	Column graph with both dimensions meaningfully labelled
CS Rel	Linking information to both process and context		Complete pictograph related to context in Part (b)	Complete table related to context in Part (b)	Complete column graph related to context in Part (b)

Results

The results are considered with respect to the three parts of the item. Table 2 contains the total number and percentage of representations coded for the four SOLO levels for the question, "What do you think 'data' means?" (Figure 1, Part a). Also included are indicative examples of student responses for each level. Two IK responses were considered incompatible (*ic*) with the context, and three were compatible (*c*). The word "information" (or an abbreviation) was used 23 times across the levels but sometimes, the meaning was conveyed in general terms. The ability to construct a sentence that related the ideas of collecting, information, and context, which was needed to be coded at the Rel level, was not demonstrated by many students.

Table 2
SOLO Levels for Part (a) of the Survey Item on Data

Level	What do you think "data" means?	%
IK (c, ic)	Like Friday 7th August 2015. (c) (ID115) It means a graph? (c) (ID151) Something you do. (ic) (ID165)	17%* (n=10)
CS Uni	Tells you stuff. (ID105) Information. (ID108) Like a survey. (ID132)	28% (n=16)
CS Multi	Collecting information. (ID103) Calculating graphs and collecting information. (ID145) Data means what you know and you put it into a graph. (ID148)	45% (n=26)
CS Rel	It means collecting information about people or a person. (ID104) Information collected on a question like = what is your favourite colour? (ID122) Data means that you collect knowledge about something and put it in a graph. (ID142)	10% (n=6)

*This value includes five students who did not reply to the question.

Table 3 contains responses to the request for examples of data (Figure 1, Part b). The difference between Uni and Multi responses depended on the implied action of collecting information or asking questions related to the example provided. Two IK responses were *ic*.

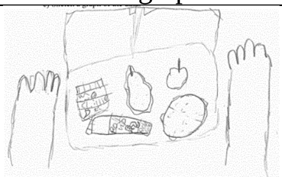
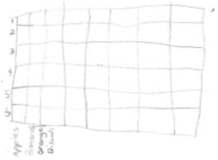
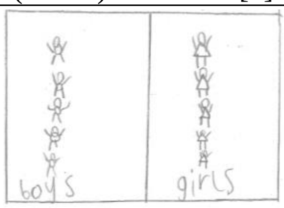
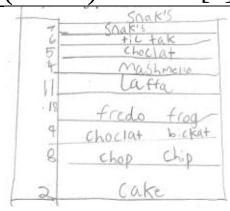
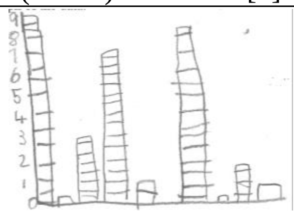
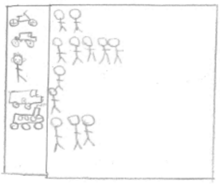
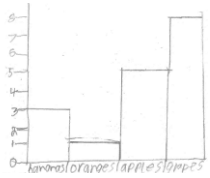
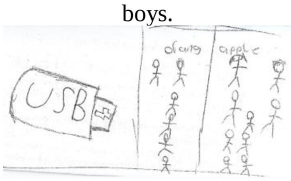
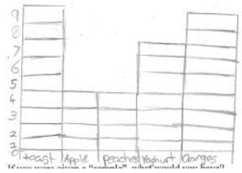
Table 3
SOLO Levels for Part (b) of the Survey Item on Data

Level	Give an example of some data you have seen or collected.	%
IK (c, ic)	I have seen data with pictures in data. (c) (ID116) We/have done graph work. (c) (ID129) Making lunch? (ic) (ID164)	19%* (n=11)
CS Uni	How many boys or girls. (ID102) A food graph. (ID104) I have some data of the earth. (ID140)	26% (n=15)
CS Multi	How many people had cake for recess. (ID103) What is your favourite colour. (ID108) Who ate what fruit and veg. (ID119)	55% (n=32)

*This value includes six students who did not reply to the question.

With respect to Part (c) (Figure 1), the students produced three types of representation: Pictographs (n=5), Tables (n=13), and Column graphs representing frequency (n=37), as expected by Year 3 (ACARA, 2019a). Table 4 contains examples of each level of representation that was assessed for the three graph types. The numbers in square brackets in each cell indicate the number of representations in that category, whereas the percentages in the right column represent the percentages across the three categories combined. For Pictographs, one variable is represented at the Uni level and two variables for Multi. At the Uni level for Tables, the list is supplemented by totals (could be tallies), whereas both are present at the Multi level. Similarly, for the Column graphs, the bars are accompanied by labels on either one (Uni) or two axes (Multi). At the Rel level, the response provided in Part (b) is included to demonstrate the connection made between the two questions by the student. The seven IK responses were considered compatible with the context.

Table 4
SOLO Levels for Part (c) of the Survey Item on Data

	Sketch a graph of the data			
Level	Pictograph	Table	Column graph	%
IK (*c*)	 (ID164) [1]	 (ID108) [2]	 (ID155) [4]	17%* (n=10)
CS Uni	 (ID103) [1]	 (ID120) [4]	 (ID131) [11]	28% (n=16)
CS Multi	 (ID102) [1]	 (ID143) [1]	 (ID121) [5]	12% (n=7)
CS Rel	7 boys liked oranges and 8 girls liked apples, so did two boys.  (ID111) [2]	Who ate fruit and veg.  (ID119) [6]	I collected data for a healthy breakfast.  (ID145) [17]	43% (n=25)

*This value includes three students who did not reply to the question.

Although blank responses are a concern, the presence of 17 IK responses across the three questions, with only four considered incompatible with the contexts, suggests that expectation of CS responses is reasonable in Year 3. Of particular interest is the association between the responses to Parts (a) and (c). Whereas 43% of students could produce a Relational level representation linked to the data in their examples, only 10% could provide a complete Relational definition of data. Fifteen students performed better on Part (a) than Part (c), whereas 26 did better on Part (c), with 14 consistent across the parts. An indicative Pearson's correlation coefficient (r=0.302, p<0.05) suggests significance but only about 9%

of shared variance. This is a sign that there is not a strong relationship between these two aspects of early learning about data.

Discussion and Conclusion

Interest in the question in the title of this paper arose at the beginning of a longitudinal project that was underpinned by the practice of statistics, fundamental to which are the data collected to answer a statistical question. Aware of the *Australian Curriculum*'s (ACARA, 2019a) definition of "data" but finding no published report of students' responses to the question prompted including the question as a survey item.

The official definition of data in the *Australian Curriculum* (ACARA, 2019a) elaborates on the word "information" in parenthesis with "observations and/or measurements", as well as with the reference to collecting data for "any type of systematic investigation." In the definitions provided by the students in this study, 40% mentioned a version of the word "information" but only one student mentioned "measuring"; none mentioned observations or observing. It may be that teachers are not making the distinction that information in the context of statistical investigations can be numerical or categorical, and measurable or observable in nature. It is possible closer attention to the definition and meaning of data will make the use of data more meaningful for students when answering statistical questions (Russell, 2006) and conducting systematic investigations (Watson et al., 2018).

It does, however, appear that young students are given the background to represent data in many ways. The students in this study utilised tallies, tables, pictographs, and column graphs, all of which are expectations of the curriculum at Year 3. This reflects appreciation of the quantifiable nature of data and the notion that data are plural in nature and collected from multiple sources as seen at all SOLO CS levels. That many responses to "What do you think data means?" described data in very general, non-quantifiable ways suggests a disconnect between how data are described and how they are represented (e.g., the correlation reported). Making explicit the connections between these two aspects of a statistical investigation in Year 3 may help students in posing questions that generate meaningful data that can be represented and analysed, part of the practice of statistics with which they have been shown to have difficulty (e.g., English et al., 2017; Wright et al., 2020).

Making meaning from data and creating data are emphasised in both the curriculum and the extant literature on student learning of statistical concepts. In terms of the contexts suggested in Parts (b) and/or (c) of the survey item, 44 students (76%) based their contexts around food, including food at recess, food for breakfast, and fruit choices. Although investigations about the contents of young students' lunch boxes provide convenient and legitimate data collection opportunities, they potentially limit exposure to contexts in which students can conduct a systematic investigation, learn about different data types, and explore how data explain and are influenced by the context of the investigation (Fitzallen & Watson, 2011; Russell, 2006). There are many resources available that provide engaging contexts for investigations that require observations and measurements to collect information (e.g., Fitzallen & Watson, 2020). It is recommended teachers embrace the learning opportunities made available when students' experiences with statistical concepts are positioned within investigations across the curriculum that explore issues related to a range of contexts.

Acknowledgements

This research project was funded by the Australian Research Council [DP150100120]. We acknowledge the contributions of Suzie Wright in assisting with this research. She was a valued colleague who will be remembered fondly.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2019a). *Australian Curriculum*. Author. https://www.australiancurriculum.edu.au/

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2019b). *NAPLAN: National Assessment Program Literacy and Numeracy*. Author. http://www.nap.edu.au/naplan/naplan.html

Australian Education Council. (1991). *A national statement on mathematics for Australian schools.* Author.

Ballou, J. (2008). Open-ended question. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 548-550). SAGE Publications.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

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

English, L., Watson, J., & Fitzallen, N. (2017). Fourth-graders' meta-questioning in statistical investigations. In A. Downton, S. Livy, & J. Hall (Eds.), *40 Years On: We Are Still Learning. Proceedings of the 40th annual conference of the Mathematics Education Research Group of Australasia*, Melbourne, July 2-6 (pp. 229-236). MERGA.

Fitzallen, N. (2012). *Reasoning about covariation with TinkerPlots*. [Doctoral dissertation]. University of Tasmania, Hobart.

Fitzallen, N., & Watson, J. (2011). Graph creation and interpretation: Putting skills and context together. In J. Clark, B. Kissane, J. Mousley, T. Spencer, & S. Thornton (Eds.), *Mathematics: Traditions and [new] Practices. Proceedings of the joint AAMT/MERGA conferences*, Alice Springs, July 3-7 (pp. 253-260).

Fitzallen, N., & Watson, J. (2020). Using the practice of statistics to design students' experiences in STEM. In B. Shelley, K. te Riele, N. Brown, & T. Crellin (Eds.), *Harnessing the transformative power of education* (pp. 74-99). Koninklijke Brill.

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.* American Statistical Association. http://www.amstat.org/education/gaise/

Groth, R. E., Austin, J. W., Naumann, M., & Rickards, M. (2019). Toward a theoretical structure to characterize early probabilistic thinking. *Mathematics Education Research Journal*, *33*, 241-261. https://doi.org/10.1007/s13394-019-00287-w

Ministry of Education. (1992). *Mathematics in the New Zealand curriculum*. Author.

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. Freeman.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Author.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Author.

Russell, S. J. (1990). Counting noses and scary things: Children construct their ideas about data. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on the Teaching of Statistics*, Dunedin, August (pp. 158-164). http://iase-web.org/documents/papers/icots3/BOOK1/A2-6.pdf

Russell, S. J. (2006). What does it mean that "5 has a lot"? From the world to data and back. In G. Burrill (Ed.), *Thinking and reasoning with data and chance* (pp. 17-29). NCTM.

Schwartz, S. L., & Whitin, D. J. (2006). Graphing with four-year-olds: Exploring the possibilities through staff development. In G. Burrill (Ed.), *Thinking and reasoning with data and chance* (pp. 5-16). NCTM.

Taylor, J. V. (1997). Young children deal with data. *Teaching Children Mathematics*, *4*(3), 146-49.

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, *47*, 337-372.

Watson, J., Fitzallen, N., Fielding-Wells, J., & Madden, S. (2018). The practice of statistics. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 105-137). Springer.

Wright, S., Fitzallen, N., Shelley, B., & Lang, M. (2020). Inspiring the next generation of scientists: Children as researchers and storytellers. *Teaching Science, 65*(4), 11-25.