Math Ed Res J (2014) 26:477–502
DOI 10.1007/s13394-013-0113-4

ORIGINAL ARTICLE

Average: the juxtaposition of procedure and context

Jane Watson · Helen Chick · Rosemary Callingham

Received: 10 April 2013 / Revised: 23 August 2013 / Accepted: 9 December 2013 /
Published online: 12 January 2014

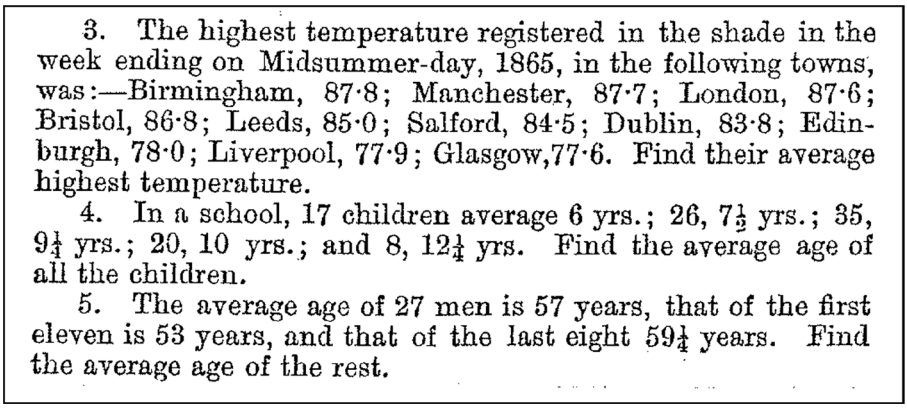
Abstract This paper presents recent data on the performance of 247 middle school students on questions concerning average in three contexts. Analysis includes considering levels of understanding linking definition and context, performance across contexts, the relative difficulty of tasks, and difference in performance for male and female students. The outcomes lead to a discussion of the expectations of the curriculum and its implementation, as well as assessment, in relation to students' skills in carrying out procedures and their understanding about the meaning of average in context.

Keywords Average · Context · Central tendency · Mean · Median · Middle school students

The word "average" has a long history, dating back to the late fifteenth century when its meaning in French was "financial loss incurred through damage to goods in transit" from the word meaning "damage to ship" (Kirkpatrick 1983; Dictionary Central n.d.). The meaning evolved by the seventeenth century to mean "equal sharing of such loss by those with a financial interest in the goods". This was extended to mathematics in the middle of the eighteenth century. Today, average can be a noun, verb, or adjective, depending on grammatical usage and context. Colloquially, it can mean "normal" or even "not very good". The degree of technicality associated with more formal uses also varies. On one hand, it may be used in a descriptive sense, such as "the average shopper", implying measurement but not explicitly providing an explanation of what algorithm is being used. On the other hand, it may be used in science, sport, accounting, or other fields with precise expectations of the definition being used. The many interpretations of average and its common usage put pressure on curricula, teachers, and students to be explicit in their application and interpretation of the term.

Over the history of mathematics education, the word average has been equated with the arithmetic mean. At least as early as 1866 (Smith 1866, p. 145) students were asked to calculate the mean, find missing values to produce a given mean, and work out a weighted mean problem. As shown in Fig. 1, these were not trivial problems, and

J. Watson (✉) · H. Chick · R. Callingham
University of Tasmania, Private Bag 66, Hobart, Tasmania, Australia 7001
e-mail: Jane.Watson@utas.edu.au



3. The highest temperature registered in the shade in the week ending on Midsummer-day, 1865, in the following towns, was:—Birmingham, 87·8; Manchester, 87·7; London, 87·6; Bristol, 86·8; Leeds, 85·0; Salford, 84·5; Dublin, 83·8; Edinburgh, 78·0; Liverpool, 77·9; Glasgow,77·6. Find their average highest temperature.

4. In a school, 17 children average 6 yrs.; 26, 7½ yrs.; 35, 9¼ yrs.; 20, 10 yrs.; and 8, 12¼ yrs. Find the average age of all the children.

5. The average age of 27 men is 57 years, that of the first eleven is 53 years, and that of the last eight 59¼ years. Find the average age of the rest.

Fig. 1 Problems from *A Shilling Book of Arithmetic* (Smith 1866, p. 145)

although a context was provided for each, they appear more like problems of practice in addition, division, algebra to work the algorithm backward, and working with fractions. This raises questions such as how has mathematics education in relation to average evolved since 1866? What do we know about the development of students' understanding of average in general, and what is important for them to understand in the twenty-first century? In the Smith text, average was the only statistical term introduced, and it meant arithmetic mean. It was the practical algorithm that was used in the society of the time and was also included by Pendlebury (1896), with examples similar to Smith's, because it was part of "so much of the science of Arithmetic as is needed for school use and for the Civil Service and other examinations". It appears that perhaps working with the mean was a criterion skill used as a benchmark for employment; for example, one might imagine a clerk sitting at a desk solving one of Pendlebury's problems: "If the cost of a prisoner to a county be £26. 19*s*. 10¾*d*. in a common year, find his average cost per day" (p. 263).

Following the requirements of classical statistics, the equating of average and arithmetic mean persisted in school textbooks until the middle of the twentieth century (e.g. Denbow and Goedicke 1959; Pendlebury and Robinson 1928). Texts for users of statistics (e.g. Boddington 1936) noted the three measures of central tendency—mean, median, and mode—but discussion of the three only reached school education at the primary and middle school level in the last decade of the century when "Data and Chance" entered the curriculum in various countries (e.g. Australian Education Council (AEC) 1991; Department of Education 1995; Ministry of Education 1992; National Council of Teachers of Mathematics (NCTM) 1989).

The issue of context as a justification and purpose for needing to know about average was at least superficially acknowledged from the beginning, as seen in Fig. 1, but the focus of problems was using the correct algorithm rather than making sense of the results in context. It was usual, for example, for exercises to begin with number-only problems to "find the average of the following numbers" (Pendlebury 1896, p. 262; Smith 1866, p. 145). The emphasis on the algorithm has persisted to this day, and one suspects that, like many word problems in arithmetic, students extract the numbers from the context and try to solve the puzzle of finding the missing piece. As the material related to statistics has gained greater attention in curriculum documents, it

has been acknowledged that there is no statistics without context (Rao 1975). Average must be seen, however, not only in a superficial real-world context but also as being representative of the data set that derives from that context (Mokros and Russell 1995). The plea of Gal (1995) that students should see a genuine need to *use* average in the tasks set was reflected in the *Used Numbers* series (Friel, Mokros, and Russell 1992). In recent years, however, the need to develop national assessment instruments of a short-answer nature, for example, National Assessment Program: Literacy and Numeracy (NAPLAN) tests in Australia, has often refocused attention on procedures and basic concepts rather than contextual understanding. Figure 2 shows two examples of this (Australian Curriculum, Assessment and Reporting Authority [ACARA] 2009, 2010).

Since chance and data became a substantive part of school education about 25 years ago, curriculum documents have dealt with the definition of average and the importance of context in various ways. The NCTM *Standards* document (1989) directly addressed the potentially ambiguous meaning of average in Standard 10 for grades 5–8, by discussing different ways of determining the characteristics of an "average" student using numerical and non-numerical data (pp. 105–107), including mean, median, mode, and the overall concept of centre. The Australian *National Statement* of 1991 avoided use of the word average and instead talked of "measures of location" (AEC, p. 178), except when referring to the impact of statistics on daily life and suggesting the investigation of "the use of the word 'average' in a range of social contexts" (p. 178). By 2000, the NCTM *Principles and Standards* deleted reference to average, opting for "measures of centre", emphasizing median in grades 3 to 5 and mean in grades 6 to 8. Later still, the NCTM's *Curriculum Focal Points* (2006) also did not refer to average. The *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* (Franklin et al. 2007) mentioned mean, median, and mode but also acknowledged the language issue associated with average in terms of "typical". In contrast to this usage, it defined the "mean absolute difference" (for measuring variation) as an "average", calculated using a mean algorithm. It also implied equivalence of the terms average and mean in reference to an example from a media article. Similarly, the US *Common Core State Standards for Mathematics* (Common Core State Standards Initiative [CCSSI] 2010) did not mention average in its sections on Statistics and Probability but did employ the word elsewhere in using an example of "batting averages", in finding a midpoint in the complex plane and in defining "expected value" in the glossary. These

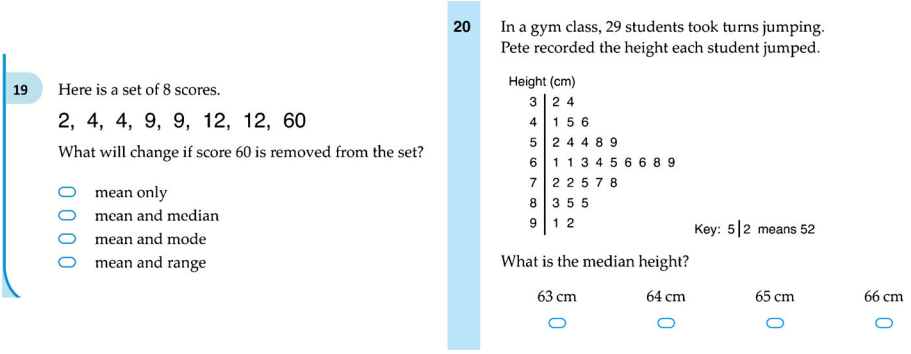


Fig. 2 Two NAPLAN questions focusing on measures of centre from year 9: Q19 (ACARA 2009, p. 7) and Q20 (ACARA 2010, p. 8)

last two examples would suggest that the colloquial use of average, when the mean is intended, continues still in highly regarded mathematics education resources. The latest version of the *Australian Curriculum: Mathematics* (ACARA 2013b) does not include the word average in the content section and introduces mean, median, mode, and their uses from year 7. Although not defined in the Glossary section, the word average is used within the definition of several other terms, such as when contrasting variable types and when linked to arithmetic mean within the definition of mean itself.

The use of average to describe mean with the implicit expectation to be able to remember and use the algorithm adds to the potential confusion for students, because intuitively they may also have other colloquial understandings of average such as "most" or "middle". The issue of which average is intended is exacerbated further when it is realized that outside of the mathematics curriculum there are other subject areas that employ the concept of average, usually without a precise definition. The *Australian Curriculum: Science* (ACARA 2013c), for example, defines "data" in its glossary and indicates that a datum "does not necessarily mean a single measurement: it may be the result of averaging several repeated measurements, and these could be quantitative or qualitative" (p. 107). No indication is given as to how averaging would differ for quantitative or qualitative measurements. Because of the requirement for statistics to make sense in many contexts, mathematics teachers/educators must be able to make clear to others what they mean by the term average. The *General Capabilities in the Australian Curriculum* (ACARA 2013a) document makes explicit the expectation for numeracy across the curriculum in its requirement for general capabilities in all discipline areas. As average may encompass several tools used in other subject areas, it is one of the elements of numeracy that students need to consolidate during the middle years, in its various interpretations.

Previous research

Probably because of its history within the school curriculum, average was the earliest area of statistics investigated by mathematics education researchers. The focus was initially on the weighted mean and difficulties of tertiary students in dealing with it. In the early 1980s, Pollatsek, Lima, and Well (1981), Mevarech (1983), Reed (1984), and Hardiman, Well, and Pollatsek (1984) considered different aspects of student difficulty, including the mathematical axioms involved (e.g. closure and associativity) and algebra word problems in contexts such as average speed, work time, and mixtures. Specific instruction on the balancing properties of the mean (Hardiman et al.) or corrective feedback (Mevarech) led to some improvement in student performance.

Following work with the more complex algorithm for weighted mean, researchers investigated properties of the basic arithmetic mean itself and became interested in students at the high school level. Representativeness, location, and expectation were all found by Goodchild (1988) to cause difficulty for these students in building a concept of mean as average. Strauss and Bichler (1988) avoided the use of the word average, instead using tasks involving "equal sharing", with their 8- to 14-year-old students, although they were investigating student understanding of properties of the mean. Various properties—both mathematical (e.g. the sum of deviations from the mean is zero) and statistical (e.g. representativeness)—caused problems for students. The results

were partially replicated by Leon and Zawojewski (1991), and both studies found different difficulties at different ages but did not suggest a developmental sequence.

Significant progress came at the school level with the work of Mokros and Russell (1995) who, rather than using surveys, began with a more open-ended approach to average in in-depth interviews with 21 students in grades 4 to 8. Without expecting a specified interpretation of the term average, they were able to classify responses to a series of concrete data-based problems into non-representative and representative approaches. The mode and the mean algorithm were considered non-representative; "reasonable", midpoint, and point of balance were seen as representative. Cai (1995, 1998, 2000) continued to explore students' understanding of the properties of the arithmetic mean, finding that although most students could identify the correct algorithm in a multiple-choice setting, they were not as successful at working the algorithm backward (e.g. starting with the mean and finding a missing value), with very few employing levelling or balancing to achieve success.

Throughout the 1990s, Watson and Moritz (1999, 2000) considered the longitudinal development of student understanding of the various aspects of average through surveys and interviews with students across years 3 to 11. Their analysis was based in the neo-Piagetian SOLO model of Biggs and Collis (1982, 1991) making possible a hierarchical classification of responses based on the complexity and relevance of information provided: prestructural (outside the concept), unistructural (a single relevant point), multistructural (more than one relevant point in sequence), and relational (integrated points to address the task). Among the outcomes that provide a foundation for the expectations of student responses in the current study are the responses provided to the two questions shown in Table 1, categorised by intended meaning; that is, "middle" was interpreted as median and "most" as mode.

In a detailed analysis of interviews with students in year 3 to first year university, asking what average meant in the statement "students watch an average of 3 h of TV per day", Watson and Moritz found that students in years 3 and 5 mentioned ideas associated with median or mode but not the mean. All other years mentioned ideas associated with the three concepts of mean, median, and mode, with many responses including two or all three concepts (2000, Table 3, p. 30). In response to questions dependent on the mean, working the algorithm backward for finding family size or calculating a weighted mean for hours of TV watching, Watson and Moritz found successful performance at around 25 % from year 8 until year 12 and first year university, where it improved for the small sample of students. These results indicate that most students appear to have sound and varied intuitions about average but lack the procedural understanding (or memory of the algorithm) to apply the mean in familiar contexts.

Table 1 Interpretations of average (summarised from Watson and Moritz 1999)

Question	Percent of responses[a] related to …		
	Mean	Median	Mode
What would it mean if someone said you were average?	1 %	86 %	18 %
What does "average wage earner buying an average home" mean?	9 %	60 %	36 %

[a] Percents add to more than 100 % due to multiple descriptions

Since the beginning of the twenty-first century, the attention of those interested in average has moved in two directions. One of these has been from the students to their teachers or to pre-service teachers. Jacobbe and Carvalho (2011) summarise the research in this area, concluding "that teachers' understanding of average seems to be similar to that of students" (p. 207). Although there has been less work done with teachers than students, it began by focussing on the mean before moving on to the median and mode. Callingham (1997) found teachers performed well on basic questions and questions involving graphical displays but struggled with a weighted mean problem. Leavy and O'Loughlin (2006) found similar results to Callingham. Although having a tendency to describe average as "in the middle", Begg and Edwards (1999) found teachers had a better understanding of mean than median or mode. Leavy and O'Loughlin also found teachers confusing the mean with the median and mode and observed that over half of their sample saw mean as synonymous with average, without considering alternatives.

The other expansion of interest in the twenty-first century has related to a focus on variation. Responding to calls from Green (1993) and Shaughnessy (1997), research began to consider issues of spread and how students come to an appreciation of variability in data. Shaughnessy suggested that perhaps the previous lack of interest in the concept of variation was related to the curriculum focus on measures of centre. This focus, particularly on the mean, rather than on measures of spread, especially the standard deviation, was likely because of the relative simplicity of the algorithm for the mean compared to the one for the standard deviation. Considering variation, however, does not eliminate the need to consider average, as discussed so pointedly by Konold and Pollatsek (2002) using the analogy of signal and noise. They focused on average, usually as mean, in order to simplify their exposition and because the mean is so commonly used, particularly to compare two groups, which is a very basic goal of much of data analysis. They suggested four interpretations of average, from calculating a specific number with an algorithm, to fair sharing, through finding a typical value, to finding the best estimate of the signal within the noise of the data. Considering several contexts for data collection (e.g. repeated measurements, measuring individuals, and comparing rates), they explored the relationship of the variation in the noise in data to the information in the signal and how it is used to compare groups. It is interesting that with the emphasis on the mean in the school curriculum, students often do not think to use it in a situation where they are asked to compare two groups (Gal, Rothschild, and Wagner 1990; Watson and Moritz 1999).

In previous research on average, context has been used in two ways. On one hand, it has been employed to set a scene such as "time spent watching TV" (Watson and Moritz 2000) where students are asked open-ended questions about how an average might have been obtained. On the other hand, it has been used to provide information (usually numbers in some form) with an expectation to use a particular algorithm (usually mean or median) to solve a specific problem requiring an explicit answer (such as a mean, total, or median). The interviews of Watson and Kelly (2005) focussing on variation in the context of the weather, provided a window on students' linking of average as a measure of centre and variation more generally. The extended interview setting and prompts, however, gave students an optimal opportunity to display their understanding. Comparing outcomes for different uses of context, however, does not appear to have been on the research agenda.

With this background in mind, the current study used survey items to explore three different contexts (family size, weather temperatures, and house prices) involving three types of task expectations (specific numerical answer, description acknowledging

variation, and definitional understanding). The expectation to engage with procedures and with context provides the opportunity to compare and contrast student performance across the tasks. The focus is on middle school students because these are the years when measures of central tendency are introduced in the school curriculum (ACARA 2013b). Specific research questions are the following:

1. What levels of understanding do middle school students display in relation to linking their definitional knowledge of central tendency to the context within which it is placed? Do these change over the middle school years?
2. How consistent is performance across contexts and tasks, and which tasks/contexts are the most difficult for students?
3. Are there differences for male and female students?

Methodology

Sample

The 247 students who completed a survey including the six questions related to average considered here were part of the 3-year StatSmart project (Callingham and Watson 2008; Callingham 2010) and had completed at least one similar survey previously during the project that included three of the questions reported here. By the final year of the project, these students had completed the required longitudinal surveys for the project and were given a survey including some previously unused items, three of which are included here. The numbers of students at each year level are given in Table 2. For the purpose of analysis, the year groups were combined in pairs, 6/7, 8/9, and 10/11. In one state, year 7 was a primary grade, and there were relatively few students in year 6. Years 8 and 9 are, according to the curriculum, appropriate for consolidating the concept of average, and there were very few year 11 students. The students came from three Australian states (Tasmania, 136; Victoria, 64; South Australia, 47). Overall, there were 47 % females and 53 % males.

Instruments

The questions employed in this study were adapted from earlier research and are presented in Fig. 3. They were part of a larger survey completed by the students and are presented in the figure in the order in which they appeared throughout the survey, interspersed among questions on other chance and data topics. The first question about the average family having 2.3 children was originally used as part of an interview protocol by Watson and Moritz (2000). In the original study, this question was asked

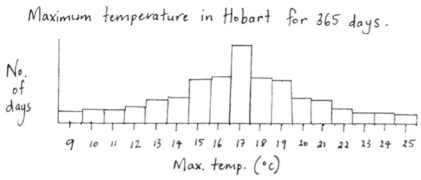
Table 2 Students in each year

Year	6	7	8	9	10	11
Male	5	34	19	36	33	4
Female	9	22	17	23	45	0
Total	14	56	36	59	78	4

Q5, 2.3Chn: The average number of children in 10 families in the neighbourhood is 2.3. One family with 5 children leaves the neighbourhood. What is the average number of children per family now? Show your work here.

Q9, Temp17: Some students watched the news every night for a year, and recorded the daily maximum temperature in Hobart. They found that the average maximum temperature in Hobart was 17°C. Explain what this tells us about the average maximum temperature in Hobart throughout the year.

Q10, Temp17Grph: Ben drew the following graph to show the maximum temperatures in Hobart throughout the year. Explain what the graph shows.



This article appeared in a newspaper.

Hobart defies homes trend

AGAINST a national trend, Hobart's median house price rose to $88,200 in the March quarter - but, Australia-wide, the average wage-earner finally can afford to buy the average home after almost two years of mortgage pain.

Q21, House1Avg: What does "average" mean in this article?

Q22, House2Med: What does "median" mean in this article?

Q23, House3Why: Why would the median have been used?

Fig. 3 Questions used in the survey

following a question about what it means for a family to have 2.3 children, which was helpful in setting the context for the question as used there. The second and third questions were part of a protocol used by Watson and Kelly (2005) exploring students' understanding of variation in the context of a city's daily maximum temperature over a year. As noted earlier, spread, outliers, and range provide links within the mathematics curriculum for considering the mean and its contribution to an overall description of a distribution. The last three questions were first used by Watson and Moritz (1999) in a survey format with 1,654 students in years 6, 8, 9, and 11. The context of median house prices is one of the most common applications of measures of central tendency in social settings. Some data are missing for the last three questions because in some situations there was not enough time for students to finish the survey.

The other questions in the survey included tasks related to chance and luck, graph interpretation, sampling, and interpretation of two-way tables. There were no other questions related specifically to average.

Analysis

The rubrics used to score responses to the six questions are given in Table 3. They reflect the influence of the SOLO Model (Biggs and Collis 1982, 1991) in recognising not only the greater structural complexity at higher levels but also the required correctness or appropriateness of responses. The rubrics for Q5, Q9, and Q10 are based

Table 3 Rubrics for six questions

Code	Criteria
Q5, 2.3Chn	
0	No response or incorrect answer with no explanation or unintelligible reasoning
1	Partial attempt—can recognise some aspect of the problem
2	Correct answer with no explanation
3	Correct answer with appropriate explanation
Q9, Temp17	
0	No response or imaginative or idiosyncratic comments
1	Single initial comment about an aspect of temperature or a single statement describing 17 on a continuum
2	A multi-step comparison of temperature with other places
3	A comment about the temperature with an acknowledgment of variation focusing around 17 or an explicit reference to variation away from 17
Q10, Temp17Grph	
0	No response or idiosyncratic responses or misinterpretation of the graph
1	A comment about the *shape of graph*; however, no indication of an understanding of the purpose of the graph
2	Statement about *frequency* and understanding of *purpose* of the graph, but no acknowledgement of the importance of 17 or focus on 17 only as more/most frequent temperature
3	Statement about frequency and purpose of the graph, and acknowledgement of the *importance of 17*
Q21, House1Avg	
0	No response or no idea of central tendency, often tautological
1	Single idea not related to context
2	Describes the central tendency for a data set or the method of obtaining the average from a data set (sometimes related to context)—(needs to state both the central tendency and the data set it comes from)
Q22, House2Med	
0	No response or no idea of central tendency, often tautological
1	Single idea not related to context
2	Describes the central tendency for a data set or the method of obtaining the median from a data set (sometimes related to context)
Q23, House3Why	
0	No response or response that does not refer to question (e.g. language/price)
1	Usefulness or fairness (without explicit mention of outliers)
2	Mention of outliers or extreme values

on those used in interview settings by Watson and Moritz (2000) and Watson and Kelly (2005). Those for Q21 to Q23 are found in Watson and Callingham (2003) where the responses split into fewer discernible groups. The range of codes reflects the range of levels of response for the different questions. What is of relevance for this study are the differences within the levels of response because this is what will help teachers take into account the relationship between procedures and context in devising pedagogies for remediation or sharing of student approaches to solutions.

To compare the levels of understanding across the years, ANOVAs were carried out for each question. Two-way tables were created to consider the association between performances on all pairs of questions. Indicative correlation coefficients were calculated (using the discrete data) to give an indication of the strength of the relationships. The difficulty of tasks was judged by comparing the mean score for each question with the total score as a percentage and ranking these percentages across the six questions. For gender differences, indicative *t* tests were carried out at each of the three paired year levels for each question and the total score. Given that 6 ANOVAs and 21 *t* tests were carried out, including for the total scores for the three paired years, the Bonferroni correction was applied. This means that the α-level for the tests reduces from 0.05 to $0.05 \div 6 = 0.0083$ for the ANOVAs and to $0.05 \div 21 = 0.0024$ for the *t* tests.

Results

Results for research question 1—levels of understanding of average

A summary of the number and percent of responses at each level of each question is given in Table 4. For Q21, Q22, and Q23, the percentages are calculated from the number of students who reached these questions. NA records the number of missing students. The trend in performance across the codes for the questions are similar, but with the mean performance of year 8/9 students better than year 6/7 students on all questions and marginally better than the year 10/11 students except for Q9. The results for the ANOVAs considering the average levels of performance across the three year groupings are presented in Appendix 1. With the Bonferroni correction applied, the differences for Q5, 2.3Chn, Q21, House1Avg, and Q22, House2Med are considered significant across the three groups. The plot of mean scores for each group and question is shown in Fig. 4. For all questions except Temp17, the year 8/9 group had the highest mean.

Because of the difference in pattern of performance across the years for Q9 and Q10 and the closer relationship of Q5 and Q21 to Q23 to the definitions of mean and median in the *Australian Curriculum* (ACARA 2012a), these four questions are presented first, followed by Q9 and Q10.

> Q5, 2.3Chn: The average number of children in 10 families in the neighbourhood is 2.3. One family with 5 children leaves the neighbourhood. What is the average number of children per family now? Show your work here.

Of the half of responses that were coded 0, most (89) put question marks, said they did not know, or left the space blank. Of the 37 responses coded 0 that wrote something, 13 put down a number different from 2 with no explanation (e.g. 1.3, 1.6, 1.7, 1.5, 1.8, 2.1, 5, 1.15). Two responses were drawings of boxes with tallies (likely to represent families with children) with no conclusion. Two responses questioned the existence of 0.3 of a child. Ten responses provided calculations that could not be interpreted, e.g.

- $10 \div 4 = 2.3$, $5 \div 4 = 1.25$
- $5 \times 9 = 45$, $45 \div 9 = 5$
- $10 \div 2 = 5$, $3 \times 5 = 15$, $10 - 1 = 9$, $15 - 5 = 15$
- $10 - 5 =$ half, $2.3 -$ half $= 1.1/5$
- 13 because $10 - 23 = 13$.

Table 4 Responses at each code level and mean for questions and Year

Question and code levels	Year 6/7 n=70	Mean	Std dev	Year 8/9 n=95	Mean	Std dev	Year 10/11 n=82	Mean	Std dev
Q5, 2.3Chn (0, 1, 2, 3)[a]	(73, 17, 4, 6)	0.40	0.83	(34, 17, 6, 43)	1.57	1.34	(45, 20, 4, 32)	1.20	1.31
Q9, Temp17 (0, 1, 2, 3)	(59, 24, 4, 13)	0.71	1.05	(43, 39, 2, 16)	0.91	1.05	(56, 18, 2, 23)	0.93	1.24
Q10, Temp17Grph (0, 1, 2, 3)	(30, 11, 44, 14)	1.43	1.07	(15, 13, 49, 23)	1.81	0.96	(22, 12, 40, 26)	1.70	1.99
Q21, House1Avg (0, 1, 2, [NA[b]])	(55, 34, 11, [25])	0.57	0.70	(16, 61, 23, [21])	1.07	0.63	(31, 51, 18, [21])	0.87	0.69
Q22, House2Med (0, 1, 2, [NA[b]])	(66, 16, 18, [26])	0.52	0.79	(31, 34, 35, [21])	1.04	0.82	(28, 47, 25, [25])	0.96	0.73
Q23, House3Why (0, 1, 2, [NA[b]])	(86, 14, 0, [26])	0.17	0.35	(62, 24, 14, [21])	0.51	0.73	(73, 18, 9, [26])	0.36	0.64

[a] (x_0, x_1, x_2, …, x_n) indicates the percent attempting the question who performed at levels (0, 1, 2, …, n)

[b] Number who did not reach the question

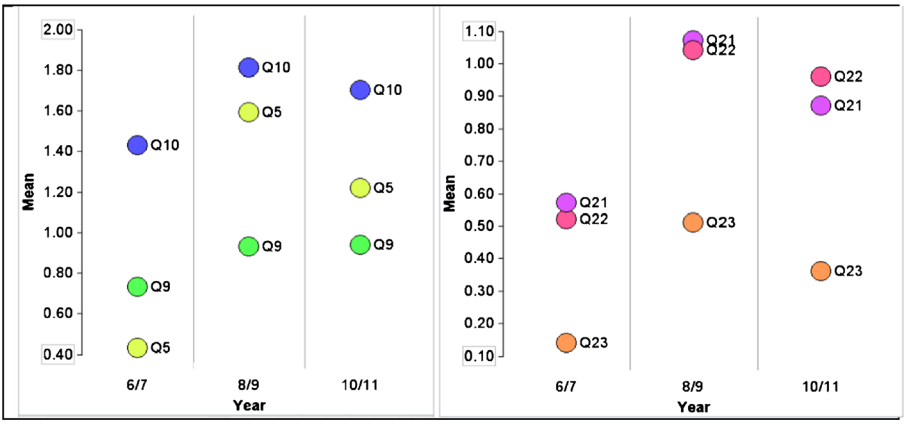


Fig. 4 Mean values for each question and group (*Left*: means for Q5, Q9, and Q10, for which the maximum code score was 3. *Right*: means for Q21, Q22, and Q23, for which the maximum code score was 2)

Three responses provided a written answer indicating that the average did not change because "the maximum number of children would be 5 to start with most likely", because "the family had more than the average amount of children", or "2.3 cause it doesn't say that the family was the only one that had five kids".

Of the 47 responses coded 1, 24 worked with the number 23 (10×2.3), subtracted 5 but divided by 10 rather than 9, obtaining an answer of 1.8 (or 1.9). Other answers provided errors in calculations using numbers that indicated the students knew or picked out from the statement of the problem numbers relevant to the solution, perhaps not completing the process. Six of these answers were larger than 2.3 indicating that students did not check the answer against the implication of the question, which resulted in reduction in the mean due to a value larger than the mean being deleted. Four responses expressed in some way or other that there was not enough information provided to work out the answer.

- It depends if there are more families with heaps of children, so it's really impossible to tell without looking at all the families that got surveyed.
- It would depend on how many children the other nine houses had.
- I don't have numbers to work out the average.
- I need more information.

Of the responses coded 2, three responses showed the appropriate intuition but did not calculate the new average.

- 1–2, I guessed that because it wouldn't change much because it's only one family.
- Less than 2.3.
- Decreases, because a high number was taken out which was keeping the average high.

Nine other students gave the correct answer of 2.0 but with no explanation or nothing that could be rationalised.

Of the 71 responses coded 3 that showed appropriate calculations to reach the answer of 2.0, there was variation in the amount of detail provided. Basically, the total of 23 children was calculated for the 10 families, five children were removed, and the

remaining 18 children were divided by nine for the remaining nine families. Eight responses created 10 families with appropriate numbers of children to equal 23 with one of five (e.g. 5, 2, 2, 3, 4, 1, 1, 1, 2, 2); that family was removed and the average shown for the nine remaining. It would appear that the context was supportive to these students in solving the problem, as the students could model numbers representing family size.

> Q21, House1Avg: What does "average" mean in the article "Hobart defies homes trend"? Q22, House2Med: What does "median" mean in the article "Hobart defies homes trend"? Q23, House3Why: Why would the median have been used in the article "Hobart defies homes trend"?

As is seen in Table 4, these questions were more difficult for students than the others. For the three questions, of the responses coded 0, an increasing number in each case wrote that they did not know or could not remember the concept involved or why the median was relevant (20 for Q21, 48 for Q22, and 66 for Q23). A further 12 responses to Q21 were tautological, as was one response to Q22. For Q21, about the meaning of average, other answers that were coded 0 focussed on irrelevant aspects.

- Equal
- The amount of money earned in a month
- The amount of mortgage in Australia
- The total
- How much you need to buy to get a house

For Q22, about the meaning of median, examples coded 0 included the following:

- That you don't have quite enough.
- House prices rose to $88,200 in the March quarter.
- The price that came up the most.
- Maybe the main price.

The modal idea occurred on several occasions. For Q23, many responses coded 0 had reasons why the median was used that did not focus on dealing with outliers.

- To produce more variability
- To simplify
- So it sounded like more
- Because it was a good choice
- To add in statistical stuff and make it interesting

For Q21 on the meaning of average, responses were not required to focus on a particular meaning of average. A code of 1 reflected a single appropriate aspect of any type of average.

- Usual
- The everyday worker type of Australian
- Typical

A code of 2 was given for responses that added to this the context and/or data set from which the average is obtained.

- Someone who doesn't earn a lot or not much. Somewhere in the middle.
- It's all the data added and divided by the number of data.

- In the middle, not too rich, not too poor.
- The group that most people are likely to be in.

Similarly for Q22 on the meaning of median, code 1 responses reflected a single aspect of the concept.

- The normal house price I think.
- Average houses, simple houses.
- The usual price range.

Code 2 responses were more specific about the properties of median, although perhaps somewhat more colloquial than a textbook definition.

- The median means the middle amount. So 88,200 is the middle amount.
- Middle between wealthy houses and non-wealthy houses.
- The middle price. For example, I think it's like 1, 1, 2, 3, 4, 7, 10. Three in this case is the median.

For Q23, on why use the median, code 1 responses reflected the usefulness of the median without referring to outliers.

- Because it's in the middle.
- Because they wanted to tell people what the 'normal' house price in Hobart is.
- The median would have been used because it provides a near-accurate middle number without too much mathematical calculations.
- Because it's a regular house.

Code 2 responses did acknowledge the need to avoid the influence of outliers.

- Because there may have been one very expensive house and thrown the mean out of whack
- Because there would have been extremes so the average wouldn't have been used
- Because it is more accurate against outliers
- So extreme cost homes wouldn't effect [sic] the average by a lot, e.g., a mansion or a run down house

Q9, Temp17: Some students watched the news every night for a year, and recorded the daily maximum temperature in Hobart. They found that the average maximum temperature in Hobart was 17 °C. Explain what this tells us about the average maximum temperature in Hobart throughout the year. Q10, Temp17Grph: Ben drew the following graph to show the maximum temperatures in Hobart throughout the year. Explain what the graph shows.

Q9 and Q10 were more open ended than the other questions, inviting students to reflect on the meaning of average in the wider context of weather rather than just on its mathematical definition. Fewer students left these questions blank or said they did not know than occurred for the other questions (30 for Q9 and 18 for Q10), suggesting the context and question encouraged engagement. For Q9, which asked about average temperature, there were 13 tautological responses. Other responses coded 0 for Q9 presented views about the weather not related directly to the average maximum temperature or a misinterpretation of the average.

- Hobart has pretty nice weather.
- It's low.
- This tell me that Hobart is not a very dry land based on the low temperature's [sic] 17 °C.
- That it is never very hot in Hobart.

Code 1 responses to Q9 contained a single comment about an aspect of temperature or about 17° on a continuum.

- The sum of every temperature divided by the number of nights=average temperature.
- The maximum in Hobart was in mid 10s to early 20s.
- This tells us that Hobart had a relatively cool year and there weren't many extremely hot days.
- It tells us that the max. temp. in Hobart is usually cold.

Code 2 recognised responses that compared Hobart temperatures with other places. As seen in Table 4, there were few of these.

- Hobart's weather for maximum temperature is around 17 °C, not going up much like the mainland.
- That it's more likely to be 17 °C. And that it depends if it's raining … Tasmania is a colder state.

The highest level of response (code 3) acknowledged variation focussing around 17 °C or an explicit reference to variation away from 17 °C.

- That Hobart is quite cold but it could also vary from 0 °C–25 °C but apparently 17 °C was the average maximum temperature in Hobart throughout the year.
- It tells us that the average for that year was 17 °C. It was recorded in summer and winter and autumn and spring. It may never have been 17 °C at all.
- This tells us that the average temperature in Hobart is quite cold and most temperatures would have been between 11° and 30°.
- It's often colder than that, or warmer. It would have been [good] to do a summer/winter average.
- Even though Hobart's maximum temperature may have reached very low or very high, the average remained at 17 °C throughout the year.

Q10, which referred to a graph of the year's daily maximum temperature, was the most open-ended question of the six analysed here. The purpose was to assess the link between the knowledge that the average maximum for the year was 17 °C and a graph depicting the frequency of temperatures throughout the year. The graph presented in the item was created for the item as a "student response" not based on actual data. Responses coded 0 were idiosyncratic or misinterpreted the graph.

- The rise and fall of temperature over the year.
- The middle of the year is a lot hotter.
- The graph shows that the temperature rises each day until the 17th day. Then it reduces and keeps decreasing.

Responses coded 1 made a statement about the shape of the graph but no indication of its purpose.

- The temp. increases, stays steady, skyrockets up, steadies again and slowly decreases to about the first recorded temp.
- The graph is symmetrical. Shows the rise and drop in the temp. The left side doesn't give numbers.
- It goes in a pattern, starts low, goes up, then low again.
- The graph shows the spikes and falls of temperature in Hobart.

At the code 2 level, responses were of two types. They either made statements about frequency and the purpose of the graph, ignoring the importance of 17 °C, or they focussed only on 17 °C as the more frequent temperature.

- The graph shows that the high and low temperatures are quite low compared to the middle ones.
- The least it got to was 9 °C.
- The graph shows that the range of temp. is between 9 °C and 25 °C. This is unusual as the temp. would be expected to reach higher than 25 °C.
- They chose the most common temperature rather than adding them all up and dividing them by 365 days.
- That 17 came up the most.
- 17 is the average, median, and mode.
- 17 was high so it [was] seventeen degrees a lot.
- It only shows that the temperatures around 17 are fairly common. The no. of days is not shown.

Code 3 responses combined the elements from code 2 responses and made statements about frequency and the purpose of the graph, including acknowledgment of the importance of 17 °C.

- Range 9–25°. Mean 17.
- This graph is fairly symmetrical—so the weather was never overly cold (below 9 °C) or overly hot (above 25 °C). A 17 °C max temp was observed for many more days than other max. temperatures.
- This graph shows that the highest number of days had a maximum of 17. The same number of days were higher than 17 than the days lower the 17.
- It shows that 17 °C was the most frequent of all the temperatures but there is a fair amount around that temp. ranging 14 °C–20 °C while the rest had a very little amount of days that reached that temperature.

Results for research question 2—consistent performance across tasks and most difficult items

The association of outcomes among the questions can be judged to some extent because of the hierarchical coding (despite its discrete nature) by considering correlation coefficients. The results, along with sample size and percent of variance explained, are presented in Table 5. Despite the significance of the coefficients, the small amount

Table 5 Correlation between pairs of questions and percentage of variance explained (*n* is sample size)

	Q5 2.3Chn	Q9 Temp17	Q10 Temp17Grph	Q21 House1Avg	Q22 House2Med
Q9, Temp17	0.147* *n*=247 [2.2 %]				
Q10, Temp17Grph	0.247*** *n*=247 [6.1 %]	0.307*** *n*=247 [9.4 %]			
Q21, House1Avg	0.180* *n*=179 [3.2 %]	0.073 *n*=179 [0.5 %]	0.141* *n*=179 [2.0 %]		
Q22, House2Med	0.268** *n*=175 [7.2 %]	0.114 *n*=175 [1.3 %]	0.205* *n*=175 [4.2 %]	0.293*** *n*=175 [8.6 %]	
Q23, House3Why	0.214** *n*=174 [4.5 %]	0.097 *n*=174 [0.9 %]	0.159* *n*=174 [2.5 %]	0.223** *n*=174 [5.0 %]	0.446*** *n*=174 [19.9 %]

*Significant at the 0.05 level; **significant at the 0.01 level; ***significant at the 0.001 level

of variance explained indicates that there is little relationship between student performance on one question and another.

Exploring further the association of performance between pairs of questions based on prior expectation or strength or weakness of association, four pairs are presented in detail in Fig. 5. Because they were based in the same context (weather), the codes for the two questions about the weather, Q9 and Q10, are shown in the two-way table in Fig. 5(a). The 52 % of responses scoring 0 on Q9 were spread across all codes with 10 % (13/128) receiving the highest code for Q10. Overall, only 5 % (13/247) achieved the highest code for both questions. From the point of view of definitions in the school curriculum, it might be expected that understanding and applying the mean (Q5) and knowing the meaning of median (Q22) would be associated, but as seen in Fig. 5(b), there is considerable variability across all codes for Q5.

a

n = 247				
Q10, Temp 17Grph	Q9, Temp17			
	0	1	2	3
0	37	15	0	1
1	21	4	2	3
2	57	26	2	26
3	13	24	3	13

b

n = 175				
Q22, House 2Med	Q5, 2.3Chn			
	0	1	2	3
0	45	7	5	11
1	22	12	6	19
2	14	13	1	20

c

n = 179				
Q21, House 1Avg	Q9, Temp17			
	0	1	2	3
0	36	8	2	9
1	40	34	1	16
2	14	13	1	5

d

n = 174			
Q23, House 3Why	Q22, House2Med		
	0	1	2
0	66	36	23
1	2	18	14
2	0	5	10

Fig. 5 Association of levels of performance among participants for four pairs of questions (codes defined in Table 3)

The lowest correlation occurred for Q9 and Q21, both of which asked for an explanation of average, either in the context of the weather or of wages and house prices. This association is shown in Fig. 5(c), where 60 % (54/90) of those scoring 0 on Q9 could provide more reasonable responses on Q21. Only 5 (3 %) of all responses reached the highest code on both questions. The highest correlation was between Q22 and Q23. As might be expected, there were somewhat higher correlations among the house price questions, but as can be seen in Fig. 5(d), of the 72 % (125/174) of responses receiving 0 for Q23, 53 % (66/125) scored 0 on Q22 as well, but the rest could do better on Q22. The higher correlation reflects the fact that very few responses (7) were better on Q23 than Q22, and there was a high percentage (37 %) scoring 0 on both.

One way to consider the relative difficulty of the questions is to rank them by the average score achieved as a percentage of the total possible score. Using this criterion for each of the year groups in Table 4, the percentages are presented in Table 6. They indicate that Q23 was the most difficult question for all three groups, and Q10 was the easiest for all three. The second most difficult for year 6/7 was Q5, whereas the second most difficult for years 8/9 and 10/11 was Q9. For the other three questions not yet mentioned for each year level, the percentages (mean in relation to total possible score) were approximately the same within the year groups: about 26 % for year 6/7 (Q9, Q21, Q22), about 53 % for year 8/9 (Q5, Q21, Q22), and about 44 % for year 10/11 (Q5, Q21, Q22).

Results for research question 3—gender differences

Comparing gender differences across each question for the three paired years resulted in 18 t tests across six questions. These are presented in Appendix 2 and give an indication of difference, acknowledging that assumptions of normality and equal distance between numerical codes (see Table 3) are unlikely to be completely appropriate. The tests revealed very few differences. The girls performed slightly better than the boys in years 6/7 on Q10 (t=2.50, p<0.02). In years 8/9, the boys performed slightly better than the girls on Q5 (t=2.42, p<0.02) and Q23 (t=2.50, p<0.02), as well as overall on the total score (t=2.52, p<0.01). Using the Bonferroni correction, however, for the 21 tests, none of these differences is statistically significant. The effect sizes for those questions showing significant differences range from 0.50 to 0.59, in the range that Cohen (1969)

Table 6 Mean scores from Table 4 as a percentage of the total possible score by year

Question	Year 6/7	Year 8/9	Year 10/11
Q5, 2.3Chn	13.3 %	52.3 %	40.0 %
Q9, Temp17	23.7 %	30.3 %	31.0 %
Q10, Temp17Grph	47.7 %	60.3 %	56.7 %
Q21, House1Avg	28.5 %	53.5 %	43.5 %
Q22, House2Med	26.0 %	52.0 %	48.0 %
Q23, House3Why	8.5 %	25.5 %	18.0 %

would call "medium". There were no gender differences for years 10/11 on any individual question or the total score. One might speculate that girls would find the context of Q10 and the weather more approachable than boys, but as the difference for this question only occurred in years 6/7, it is a difficult argument to sustain. The same is true for the two more procedural or definition-based questions, Q5 and Q23, for years 8/9; these two questions contribute to the overall difference for males at this year level, but in fact, males performed marginally better on all questions than females.

Limitations

As in all studies involving surveys, there is the danger that questions may not be interpreted in the way intended by the researchers. There also may be variations based on previous understanding of the meaning of average, of the decimal 2.3, of the representation in the plot, or of the context of house prices. These understandings provide a range of background experiences and knowledge that teachers need to be aware of in the classroom. As well, it has been acknowledged that statistical tests have been applied to discrete (ordered categorical) rather than continuous, scaled data.

Discussion and implications

In relation to the developmental levels displayed by students in answering these questions on average, it may not be surprising that the year 6/7 students struggled, particularly with Q5 based on the mean, but given the presence of mean, median, and mode in the Australian curriculum from year 7 (ACARA 2012a), the results for year 8/9 are disappointing. The percentage of year 8/9 students achieving the highest level of response ranged from 47.3 % on Q22 and 45.3 % on Q5, down to 18.1 % on Q23 and 16.8 % on Q9. The percentages at the highest level were lower for year 10/11 on Q5, Q21, Q22, and Q23 but higher on Q9 and Q10. The fact that the best performance of the year 8/9 students was related to using the mean algorithm and defining the median is somewhat encouraging given the curriculum. The lack of mention of outliers (Q23), however, suggests perhaps this aspect of the curriculum is not being implemented consistently in classrooms. It is likely that a lack of continued reinforcement of the concepts is reflected in the drop in performance at the highest level in year 10/11.

One suspects that the poor performance of year 8/9 on Q9 (16 %) and Q10 (23 %) at the highest level reflects the lack of engagement with tasks that are based in contexts to develop understanding of average as a representative measure embedded in a data set displaying variation. It appears difficult for students to move away from the explicit procedural knowledge to a wider appreciation of how the concepts link with others in statistics, especially variation and its representation in graphical form. The fact that year 10/11 students performed better at this level on Q9 (23 %) and Q10 (26 %) may indicate that although their procedural/definitional knowledge has not been reinforced, their intuitions have improved, perhaps through interactions with the average concept in other subject areas where context is significant in understanding what an average represents.

Overall the linking of procedure and context was not strong in Q5. Students who remembered the algorithm used it without reference to the families. As noted, the closest link to context for successful responses was shown by the eight responses that created 10 imaginary families to satisfy the conditions of the task. Some of the unsuccessful attempts drew diagrams with tallies to try and represent the families. The link to context was somewhat stronger in Q21, Q22, and Q23 because of the newspaper article and the specific questions, but often students focused only on a dry definition, and the poor performance on Q23 suggests that the understanding of outliers, potentially present in the context, was not appreciated. Perhaps forced by the question and context of the task, Q9 and Q10 encouraged an engagement with the weather context. For both questions, the highest level responses incorporated the appreciation of variation, acknowledged within the context of the task.

Given the sample sizes for males and females within the three year pairings, it is difficult to make any judgments or hypotheses on the reasons for the four small differences found. None appeared for more than one year-level pair. In a study of middle school students of the same age as this study, Carmichael and Hay (2009) reported gender differences for males and females on questions relating to the interests they had related to statistical literacy contexts. They found significantly more positive responses for males to three questions that potentially could contribute to better performance on Q5 based on the mean and Q23 on outliers: "I'm interested in working on problems involving data and statistics"; "I'm interested in looking up unusual statistics"; "I'm interested in using averages to compare sports." The only significant differences reflecting a more positive interest in statistical literacy for females were related to collecting or completing surveys in various contexts. On a question related to the weather, "I am interested in the average rainfall for my area", there was no difference between male and female interest.

The lack of consistency across questions has been noted in the description of the levels of development, but the correlations and percentage of variation explained between pairs of questions reinforce the supposition of an inconsistency in presentation in the middle school classroom. The difficulty of the questions shown in the comparison of mean scores to total possible scores point to it being easier for students to engage with a pictorial/graphical image and link it to the discussion of average than to perform the procedurally based tasks. This is a very tentative supposition and would require further research with a variety of tasks.

It is likely that the statement of the tasks in this study encouraged either procedural (Q5, Q21–Q23) or context-based (Q9, Q10) responses. On one hand, from anecdotal evidence and previous research (e.g. Cai 1995, 2000), one suspects that the emphasis in the classroom has remained to a large extent on the procedural aspects of average. If this is the case, the outcomes of this research are disappointing. Historically, looking in detail at the *Australian Curriculum: Mathematics* (ACARA 2012) at year 7, the Descriptor asking students to "Calculate mean, median, mode and range for sets of data. Interpret these statistics in the context of data" (ACMSP171, p. 37) had an Elaboration, "Calculating mean areas set aside for parkland, manufacturing, retail and residential dwellings to compare land use in the local municipality". In the later version of the curriculum (ACARA 2013b, p. 68), this Elaboration has been removed. The following Descriptor about interpreting

mean and median in data displays, however, still includes the Elaboration, "Locating mean, median and range on graphs and connecting them to real life" (ACARA 2013b, ACMSP172, p. 68). Although examples of "real life" contexts are noted elsewhere in the Statistics and Probability section of the curriculum, it is disappointing they are not linked to measures of centre. The somewhat better performance of students on Q9 and Q10 suggests hope that this emphasis on context, if it is not now, will become a reality. It would also be interesting to conduct a longitudinal study to follow student development over the next few years as curricula such as the *Common Core State Standards for Mathematics* (CCSSI 2010), GAISE (Franklin et al. 2007), and the *Australian Curriculum: Mathematics* (ACARA 2013b) are implemented. There is a further challenge for writers of NAPLAN questions in Australia to devise items that reach across meaningful contexts and allow students the opportunity to display their understanding beyond basic procedures and definitions.

One has the feeling that if an approach to statistics is taken that reflects "big picture" investigations as starting points—with average seen as one of the tools—as suggested in the inferentialism approach suggested by Bakker and Derry (2011), then the links hinted at in this study would be reinforced both to the rest of the statistics curriculum and to the contexts associated with the rest of the school curriculum. This approach is illustrated by the experience of a year 10 class that began by exploring a media claim that brown-eyed people had faster reaction times than other people (Watson 2008). The ensuing investigation reached across the school curriculum as well as linking many parts of the statistics curriculum. In this case, the average as mean was one of the tools assisting the class to draw an inference from the data they collected. At no time, however, did the class employ the algorithm personally to complete the calculation of the mean. The question might also then be, should there continue to be such a strong emphasis on procedures and algorithms for average? Maybe the outcomes in terms of students being able to apply their understanding in social settings when they leave school would be more positive with an intentionally contextual approach. Technology can provide the straightforward algorithms but not the nous for how to plan an investigation using average or interpret an answer in context for decision making.

The outcomes of this study suggest that adding realistic contexts to problems about average does not always result in the same impact. Some contexts are more easily taken up by students, and the actual wording and expectation for calculation, description, or interpretation of graphical information appear to produce different tensions for students, depending on both their procedural and contextual knowledge. More research is needed to delve further into the characteristics of the learners and the problem settings to refine researchers' and teachers' understanding of how to deal with the difficulties students experience. Given the lack of research related to students' understanding of average in recent years, following this avenue should produce a new flow of outcomes that are not only very interesting but also valuable in terms of achieving learning outcomes that are useful when students leave school.

Acknowledgments This project was funded by Australian Research Council Grant No. LP0669106. An earlier version of some of these results was presented at the Mathematics Education Research Group of Australasia conference in Singapore, 2012 (Watson and Chick 2012).

Appendix 1

Table 7 ANOVA results for three year groupings

	Sum of squares	*df*	Mean square	*F*	Sig.
Q5, 2.3Children					
Between groups	55.049	2	27.524	18.901	0.000
Within groups	355.325	244	1.456		
Total	410.373	246			
Q9, Temp17					
Between groups	2.105	2	1.052	0.844	0.431
Within groups	304.253	244	1.247		
Total	306.358	246			
Q10, Temp17Grph					
Between groups	5.963	2	2.982	2.778	0.064
Within groups	261.865	244	1.073		
Total	267.828	246			
Q21, House1Avg					
Between groups	6.906	2	3.453	7.731	0.001
Within groups	78.610	176	0.447		
Total	85.515	178			
Q22, House2Med					
Between groups	7.960	2	3.980	6.472	0.002
Within groups	105.764	172	0.615		
Total	113.724	174			
Q23, House3Why					
Between groups	3.783	2	1.892	4.849	0.009
Within groups	66.697	171	0.390		
Total	70.480	173			

Appendix 2

Table 8 Gender differences for each problem by paired years

	Males			Females					
Group	Mean	Std dev	*N*	Mean	Std dev	*N*	Diff	*t*	*p*
Q5: 2.3Chn									
Year 6/7	0.49	0.94	39	0.29	0.59	31	0.20	1.02	0.31
Year 8/9	1.84	1.32	55	1.18	1.32	40	0.66	2.42	0.02[a]
Year 10/11	1.05	1.42	37	1.07	1.25	45	0.28	0.97	0.34
Q9, Temp17									
Year 6/7	0.56	0.91	39	0.90	1.16	31	−0.34	1.37	0.18
Year 8/9	0.93	1.05	55	0.88	1.04	40	0.05	0.24	0.81
Year 10/11	0.97	1.30	37	0.89	1.19	45	0.08	0.31	0.76
Q10, Temp17Grph									
Year 6/7	1.15	1.06	39	1.77	0.99	31	−0.62	2.50	0.02[b]
Year 8/9	1.82	0.96	55	1.80	0.97	40	0.02	0.09	0.93
Year 10/11	1.78	0.92	37	1.62	1.21	45	0.16	0.67	0.51
Q21, House1Avg									
Year 6/7	0.54	0.66	24	0.60	0.75	20	−0.06	0.27	0.79
Year 8/9	1.11	0.65	44	1.00	0.59	30	0.11	0.76	0.45
Year 10/11	0.93	0.68	27	0.82	0.72	34	0.10	0.57	0.57
Q22, House2Med									
Year 6/7	0.50	0.78	24	0.55	0.83	20	−0.05	0.21	0.84
Year 8/9	1.16	0.86	44	0.87	0.73	30	0.29	1.52	0.13
Year 10/11	1.00	0.66	24	0.94	0.79	33	0.06	0.31	0.76
Q23, House3Why									
Year 6/7	0.21	0.41	24	0.05	0.22	20	0.16	1.53	0.13
Year 8/9	0.68	0.77	44	0.27	0.58	30	0.42	2.50	0.01[c]
Year 10/11	0.39	0.66	23	0.33	0.65	33	0.06	0.33	0.74
Total Score									
Year 6/7	3.67	2.51	24	4.40	1.70	20	−0.73	1.11	0.27
Year 8/9	7.61	2.94	44	5.87	2.91	30	1.75	2.52	0.01[d]
Year 10/11	6.26	3.58	23	5.39	3.33	33	0.87	0.93	0.36

[a] Effect size=0.50

[b] Effect size=−0.60

[c] Effect size=0.59

[d] Effect size=0.59

References

Australian Curriculum, Assessment and Reporting Authority. (2009). *National Assessment Program Literacy and Numeracy. Numeracy non-calculator. Year 9 2009*. Sydney: Australian Curriculum, Assessment and Reporting Authority.

Australian Curriculum, Assessment and Reporting Authority. (2010). *National Assessment Program Literacy and Numeracy. Numeracy non-calculator. Year 9*. Sydney: Australian Curriculum, Assessment and Reporting Authority.

Australian Curriculum, Assessment and Reporting Authority. (2012). *The Australian curriculum: mathematics, version 3.0, 23*. Sydney: Australian Curriculum, Assessment and Reporting Authority.

Australian Curriculum, Assessment and Reporting Authority. (2013a). *General capabilities in the Australian curriculum, January, 2013*. Sydney: ACARA.

Australian Curriculum, Assessment and Reporting Authority. (2013b). *The Australian curriculum: mathematics, version 5.0, 20*. Sydney: Australian Curriculum, Assessment and Reporting Authority.

Australian Curriculum, Assessment and Reporting Authority. (2013c). *The Australian curriculum: science, version 5.0, 20*. Sydney: Australian Curriculum, Assessment and Reporting Authority.

Australian Education Council. (1991). *A national statement on mathematics for Australian schools*. Carlton, VIC: Australian Education Council.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*, 5–26.

Begg, A., & Edwards, R. (1999). Teachers' ideas about teaching statistics. *Proceedings of the 1999 Combined Conference of the Australian Association for Research in Education and the New Zealand Association for Research in Education*. Melbourne: Australian Association for Research in Education. Retrieved on November 1, 2012 from http://www.aare.edu.au/data/publications/1999/beg99082.pdf

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

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. A. H. Rowe (Ed.), *Intelligence: reconceptualization and measurement* (pp. 57–76). Hillsdale: Lawrence Erlbaum.

Boddington, A. L. (1936). *Statistics and their application to commerce* (7th ed.). London: Sir Isaac Pitman.

Cai, J. (1995). Beyond the computational algorithm: students' understanding of the arithmetic average concept. In L. Meira & D. Carraher (Eds.), *Proceedings of the 19th Psychology of Mathematics Education Conference* (Vol. 3, pp. 144–151). São Paulo: PME Program Committee.

Cai, J. (1998). Exploring students' conceptual understanding of the averaging algorithm. *School Science and Mathematics, 98*, 93–98.

Cai, J. (2000). Understanding and representing the arithmetic averaging algorithm: an analysis and comparison of US and Chinese students' responses. *International Journal of Mathematical Education in Science and Technology, 31*, 839–855.

Callingham, R. A. (1997). Teachers' multimodal functioning in relation to the concept of average. *Mathematics Education Research Journal, 9*, 205–224.

Callingham, R. (2010). Trajectories of learning in middle years' students' statistical development. Refereed paper in C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July). [CD-ROM] Voorburg, The Netherlands: International Statistical Institute.

Callingham, R., & Watson J. M. (2008). Overcoming research design issues using Rasch measurement: The StatSmart project. In P. Jeffery (Ed.), Proceedings of the AARE annual conference, Fremantle, December, 2007. Retrieved on September 12, 2009 from: http://www.aare.edu.au/07pap/cal07042.pdf

Carmichael, C., & Hay, I. (2009). Gender differences in middle school students' interests in a statistical literacy context. In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides: proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 97–104). Palmerston North: MERGA.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic.

Common Core State Standards Initiative (CCSSI). (2010). Common core state standards for mathematics. Washington, DC: National Governors Association for Best Practices and the Council of Chief State School Officers. Retrieved on 29 April, 2012 from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Denbow, C. H., & Goedicke, V. (1959). *Foundations of mathematics*. New York: Harper & Row.

Department for Education (England and Wales). (1995). *Mathematics in the national curriculum*. London: Author.

Dictionary Central. (n.d.). http://www.dictionarycentral.com/definition/average.html

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 preK-12 curriculum framework. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/gaise/. Retrieved July 3, 2009

Friel, S. N., Mokros, J. R., & Russell, S. J. (1992). *Statistics: middles, means, and in-betweens: a unit of study for grades 5–6 [Used numbers: real data in the classroom]*. Palo Alto: Dale Seymour.

Gal, I. (1995). Statistical tools and statistical literacy: the case of the average. *Teaching Statistics, 17*, 97–99.

Gal, I., Rothschild, K., & Wagner, D. A. (1990, April). *Statistical concepts and statistical reasoning in school children: convergence or divergence*? Paper presented at the meeting of the American Educational Research Association, Boston

Goodchild, S. (1988). School pupils' understanding of average. *Teaching Statistics, 10*, 77–81.

Green, D. (1993). Data analysis: what research do we need? In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: who should teach it?* (pp. 219–239). Voorburg: International Statistical Institute.

Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). Usefulness of a balance model in understanding the mean. *Journal of Educational Psychology, 76*(5), 792–801.

Jacobbe, T., & Fernandes de Carvalho, C. (2011). Teachers' understanding of average. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Teaching statistics in school mathematics—challenges for teaching and teacher education* (pp. 199–209). New York: Springer.

Kirkpatrick, E. M. (Ed.). (1983). *Chambers 20th century dictionary*. Edinburgh: Chambers.

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

Leavy, A., & O'Loughlin, N. (2006). Preservice teachers' understanding of the mean: moving beyond the arithmetic average. *Journal of Mathematics Teacher Education, 9*, 53–90.

Leon, M. R., & Zawojewski, J. S. (1991). Use of the arithmetic mean: an investigation of four properties, issues and preliminary results. In D. Vere-Jones (Ed.), *Proceedings of the third International Conference on Teaching Statistics* (School and general issues, Vol. 1, pp. 302–306). Voorburg: International Statistical Institute.

Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*, 415–429.

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

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*, 20–39.

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

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

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: a quest for coherence*. Reston: National Council of Teachers of Mathematics.

Pendlebury, C. (1896). *Arithmetic* (9th ed.). London: George Bell and Sons.

Pendlebury, C., & Robinson, F. E. (1928). *New school arithmetic*. London: G. Bell and Sons.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191–204.

Rao, C. R. (1975). Teaching of statistics at the secondary level: an interdisciplinary approach. *International Journal of Mathematical Education in Science and Technology, 6*, 151–162.

Reed, S. K. (1984). Estimating answers to algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 778–790.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), People in mathematics education (Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia, vol. 1, pp. 6–22), Waikato, NZ: MERGA

Smith, B. (1866). *A shilling book of arithmetic for national and elementary schools*. Cambridge: Macmillan and Co.

Strauss, S., & Bichler, E. (1988). The development of children's concept of the arithmetic average. *Journal for Research in Mathematics Education, 19*, 64–80.

Watson, J. (2008). Eye colour and reaction time: an opportunity for critical statistical reasoning. *Australian Mathematics Teacher, 64*(3), 30–40.

Watson, J. M., & Callingham, R. A. (2003). Statistical literacy: a complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3–46.

Watson, J., & Chick, H. (2012). Average revisited in context. 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, eBook* (pp. 753–760). Singapore: MERGA, Inc.

Watson, J. M., & Kelly, B. A. (2005). The winds are variable: student intuitions about variation. *School Science and Mathematics, 105*, 252–269.

Watson, J. M., & Moritz, J. B. (1999). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*(4), 15–39.

Watson, J. M., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1&2), 11–50.