

Chapter 8

Posing Comparative Statistical Investigative Questions



Pip Arnold and Maxine Pfannkuch

Abstract A “good” statistical investigative question is one that allows rich exploration of the data in hand, discovery, and thinking statistically. Two outcomes from four research cycles over a period of five years were: The development of criteria for what makes a good statistical investigative question and a detailed two-way hierarchical classification framework for comparative statistical investigative questions that are posed. With a focus on the last research cycle, responses from pre- and post-tests are explored, and the level of comparative statistical investigative questions that students posed is discussed.

Keywords Comparisons · SOLO taxonomy · Statistical enquiry cycle
Statistical investigative questions

8.1 Introduction

Arnold (2008) highlighted posing statistical questions as a problematic situation because of its role in assessment for qualifications in New Zealand and because teachers lacked knowledge in this area. The problem arose in the first of four research cycles where students in a test situation posed a statistical question, which was checked as satisfactory by the teacher. The students subsequently were unable to finish the test because their statistical question was not suitable for the given data. This raised the question “What makes a good statistical question?”, as the teacher and researcher together had marked the student posed questions correct. In an attempt to answer the question “What makes a good statistical question?” the literature was reviewed extensively and the conclusion drawn was that generally the literature gave

P. Arnold (✉)
Karekare Education, Auckland, New Zealand
e-mail: pip@karekareeducation.co.nz

M. Pfannkuch
Department of Statistics, The University of Auckland, Auckland, New Zealand
e-mail: m.pfannkuch@auckland.ac.nz

© Springer Nature Switzerland AG 2019
G. Burrill and D. Ben-Zvi (eds.), *Topics and Trends in Current Statistics Education Research*, ICME-13 Monographs, https://doi.org/10.1007/978-3-030-03472-6_8

mixed messages about what makes a good statistical question and the purpose of a statistical question. Indeed, Arnold (2013) concluded that the identified problem was actually about “What makes a good statistical *investigative* question?” Over four research cycles what makes a good statistical investigative question was explored, and the resultant criteria for what makes a good statistical investigative question were formed. These criteria informed the teaching experiment for research cycle four with a particular focus on comparative [statistical] investigative questions. Investigating comparative situations is a major focus in the New Zealand statistics curriculum at year 10 (ages 14–15) where this research took place. Hence, it is important for New Zealand teachers to know what makes a good statistical question at the school level, the components and concepts underpinning a good statistical question, and the learning in which students should be immersed to support the posing of good statistical questions. This chapter focuses on this fourth research cycle and explores the research question—*What level of comparative investigative questions are year 10 (ages 14–15) students posing?*

8.2 Literature Review

8.2.1 Statistical Investigative Cycle

The first dimension of the four-dimensional framework for statistical thinking in empirical enquiry (Wild and Pfannkuch 1999) is concerned with what one thinks about and the way in which one acts during a statistical investigation. Wild and Pfannkuch (1999) worked with the PPDAC (problem, plan, data, analysis, and conclusion) model (MacKay and Oldford 1994) of the statistical investigative cycle, and this is the model that underpins the work in this research:

- The **problem** stage deals with grasping a particular system’s dynamics and understanding and defining the problem.
- The **planning** stage involves deciding what to measure and how, how the sample will be taken, the design of the study, and how the data will be managed, including the recording and collecting of data. It also includes piloting the investigation and planning the analysis.
- The **data** stage is concerned with collecting, managing and cleaning the data.
- The **analysis** involves sorting the data, constructing tables and graphs as appropriate, exploring the data, looking for patterns, planned and unplanned analysis, and generating hypotheses.
- The final stage of the cycle involves interpreting, generating **conclusions**, new ideas and communicating findings.

In the statistical investigative cycle, questions and questioning arise in all areas. Questions are formally posed in both the problem and planning stages, in particular. Definitions and clarification of the purposes of these questions are now discussed.

8.2.2 Questions Within the Statistical Investigative Cycle

The initial motivating question for this research was: What makes a good statistical question? A number of studies were found where forming statistical questions were part of the researched process (e.g. Burgess 2007; Hancock et al. 1992; Lehrer and Romberg 1996; Pfannkuch and Horring 2005; Russell 2006) and a number of papers or books were located that reported an overview of the current status of statistics education, including forming statistical questions (e.g. Graham 2006; Konold and Higgins 2002; Whittin 2006). After reviewing existing literature and considering the statistical investigative cycle, the picture of what makes a good statistical question was still unclear. There were mixed messages about the purpose of statistical questions and whether they were used for an investigation or to collect data from people. From the literature (e.g. Burgess 2007; Russell 2006; Pfannkuch and Horring 2005) and from experience, it was concluded that within statistical investigations we can consider two types of questions: those that are formally posed and those that are spontaneously asked throughout the investigative process.

The theory proposed by Arnold (2013), therefore, is that there is *question posing* and *question asking*. Question posing results in a question being formally structured, whereas question asking is a continual spontaneous interrogative process. Question posing arises as a result of having a problem that needs to be addressed using a statistical investigation. Posed questions may be asked for investigative or survey purposes: *investigative* questions are those to be answered using data (the problem), while *survey* questions are those asked to get the data (the plan). Question asking also has two purposes, both of which involve an interrogation element: *interrogative* questions are those asked as checks within the PPDAC cycle, while *analysis* questions are those asked about the statistics, graphs and tables in order to develop a description of and an inference about what is noticed (the analysis).

As this research is focused on situations where students are working with secondary data, i.e. data that has already been collected and is given to them, Fig. 8.1 shows where questions fit within the statistical investigative cycle when students are given data. The purpose of Fig. 8.1 is to show how many different types of “statistical” questions are used within the PPDAC cycle, reinforcing how it could be very confusing for students if the questions are not defined and named according to their different purposes.

8.2.3 Posing Statistical Investigative Questions

In the big picture of statistical enquiry the investigative question is the statistical question or problem that needs answering or solving. In most instances the investigative question starts from an “inkling” and is developed into a precise question. The process of developing or creating the investigative question is iterative and requires considerable work to get it right (e.g., delMas 2004; Franklin et al. 2005;

The PPDAC cycle in situations where the data are given to the students (secondary data).
<p>Original Investigators DATA/PLAN</p> <p>Data set is given.</p> <p><i>Interrogate</i> the background of the data.</p> <p>Understand the original <i>investigative</i> question(s) and data collection plan including how the data were collected and who the data were collected from. Find out the <i>survey</i> questions used. Identify the variables of interest and what they measure, and identify the population of interest.</p> <p><i>Interrogate</i> the data.</p>
<p>PROBLEM</p> <p>Motivating situation/question/idea of what “I” will investigate using this provided data.</p> <p>POSE a new <i>investigative</i> question.</p> <p><i>Interrogate</i> investigative question.</p>
<p>ANALYSIS</p> <p>Calculate statistics, draw graphs, build tables as appropriate.</p> <p>ASK <i>analysis</i> questions about the statistics, graphs and tables and describe what is noticed and what is inferred.</p> <p><i>Interrogate</i> analysis.</p> <p>Revisit the investigative question, updating if necessary (maybe pose new investigative questions that have become apparent after analysing the data).</p>
<p>CONCLUSION</p> <p>Write the “conclusion” answering the new investigative question(s).</p> <ul style="list-style-type: none"> • Use supporting evidence from the analysis. • Make inferences about the population. <p><i>Interrogate</i> conclusion.</p> <p>POSE further investigative questions as a result of “conclusion”.</p>

Fig. 8.1 Questions within the statistical investigative cycle: secondary data (Arnold 2013, p. 22)

Hancock et al. 1992; Russell 2006; Wild and Pfankuch 1999). There is also a need when developing the investigative question to have “an understanding of the difference between a question that anticipates a deterministic answer and a question that anticipates an answer based on data that vary” (Franklin and Garfield 2006, p. 350).

Posing investigative questions has been identified as a problem area for students, for example, the idea of asking questions of the data. Pfankuch and Horring (2005) noted that students lacked understanding of what a question is and the idea that one can pose a problem by asking questions of data: “Maybe students haven’t yet formed that understanding of what a question is—how you can ask a question in a set of data” (p. 208). Lehrer and Romberg (1996) found that students initially had problems with asking questions of data: “students believed that questions cannot be asked of data, only of people” (p. 80). Burgess (2007) noted that students found posing investigative questions a problem but did not specify the particular issue that arose. Other issues related to investigative questions include the need for teachers to model posing investigative questions, initially as seed or starter ideas (e.g. Lehrer and

Romberg 1996), but also to push students thinking about, for example, “typicalness” and data as an aggregate rather than individual cases (e.g. Konold and Higgins 2003).

In order to get precise investigative questions that can be interpreted and that yield useful information, an *interrogative* process, which involves asking questions of the investigative question, is necessary (e.g., Burgess 2007; Graham 2006; Konold and Higgins 2003). For example, Burgess (2007) acknowledged that some of the specialised content knowledge teachers needed for teaching statistics related to their ability to be able to decide if a question posed by their students was suitable, unsuitable, or whether changes could be made to make the question suitable. Graham (2006) provided five useful considerations for forming a good investigative question, which were different aspects of interrogating the investigative question. The considerations were whether the question was: “(1) actually a question, rather than simply an area for investigation...; (2) personally interesting to you...; (3) likely to draw on data that will be available within the time frame of the investigation...; (4) specific, so that it is answerable from data...; (5) measurable....” (p. 88). With this perspective, investigative questions are formulated through an interrogative process with regard to the considerations.

8.3 Methodology

Design based research (DBR) or design experiments were used. DBR has its foundations in design science (Brown 1992) and typically involves a planned intervention that develops ideas based on theoretically grounded innovations to inform practice while simultaneously conducting research on the intervention (Brown 1992; Cobb 2000). In particular, DBR focuses on the types of learning that differ from common or current practice and explores new and novel practices with the intent to change systems by being innovative (Bakker 2004; Bakker and van Eerde 2015; Schwartz et al. 2008).

A design experiment is a form of interventionist research that creates and evaluates novel conditions for learning. The desired outcomes include new possibilities for educational practice and new insights on the process of learning. Design experiments differ from most educational research because they do not study what exists; they study what could be. (Schwartz et al. 2008, p. 47)

The research, using DBR, started with an initial preparation and design phase, followed by a teaching experiment, then a retrospective analysis phase, which fed into another preparation and design phase, with the cycle repeated four times (e.g. Bakker 2004; Bakker and van Eerde 2015). A hypothetical learning trajectory (HLT) (Simon 1995) was used in the design of instructional materials. In the teaching experiment phase the teacher and researcher (as observer) together experienced the students’ learning and reasoning in the classroom. Each lesson was reflected on and informed the next lesson. During the teaching experiment phase, evidence was collected in the form of video-recordings of lessons, field notes, pre-and post-tests and interviews

of some students for the retrospective analysis, which occurred at two levels. An ongoing retrospective analysis informed subsequent planning and was motivated by what seemed best for the students (Cobb 2000). The retrospective analysis at the end of a teaching experiment was orientated by the HLT and conjectures both of which provided a basis for developing the instruction theory (Bakker 2004; Cobb 2000). The research process was iterative—design, test and redesign.

For the pre- and post-tests the retrospective analysis involved writing hierarchical descriptors based on the student data and criteria derived from the literature followed by the subsequent classification of student responses into categories. The categories evolved over four cycles and were based on the SOLO taxonomy (Biggs and Collis 1982). The SOLO taxonomy then provided the basis for quantification of the responses, which were then analysed quantitatively. Transcriptions of the video recordings were used to identify salient moments within the class lessons in order to provide evidence and illustrations of how students were scaffolded to interrogate and pose investigative questions.

Four research cycles were undertaken in 2007, 2008, 2009 and 2011. This chapter reports on the findings and outcomes from 2011, the last cycle. At the end of the first teaching cycle the problematic situation, what makes a good statistical question, was identified.

8.3.1 *Participants*

The first two teaching experiments were undertaken in a state, mid-socio economic, multicultural, suburban co-educational school with Teacher A, who in 2007 was in her fifth year of teaching. Her year 10 students (ages 14–15) in 2007 were average to below average in ability, while in 2008 the students were above average in ability. The last two teaching experiments were undertaken in a state, mid-socio economic, multicultural, inner-city girls' school with Teacher B, who in 2009 was in her ninth year of teaching. Her year 10 students in 2009 were average in ability. For the 2011 class focused on in this chapter, there were 29 students of above average ability involved in the research. The class had a mix of ethnicities including New Zealand European, Māori, Pasifika and Chinese.

8.4 Teaching Experiments

To situate the research question, the relevant elements of the four teaching experiments are given. These are the elements that: (1) contributed to the criteria; and (2) were relevant for comparative situations, the focus of the research question.

In every instance the teaching experiment is within the context of the statistics topic in a year 10 (ages 14–15) mathematics class. The main focus at this year level in New Zealand is on comparative situations. Generally the students would have about

4–5 weeks of the statistics topic across this one year of schooling. It is important for the reader to note that at the time of this research, the teaching of posing investigative questions in New Zealand was limited to teachers putting an investigative question on the board and then expecting the students to pose their own investigative questions with little or no formal teaching about how to pose investigative questions. For many students this would have been the first time they would have met comparative situations and especially the expectation to pose comparative investigative questions.

8.4.1 Teaching Experiment One

Posing investigative questions was identified at the end of the first teaching experiment as a problematic situation that was in need of further exploration (Arnold 2008). The hypothetical learning trajectory for posing investigative questions evolved over the teaching experiments. In the first teaching experiment, as questioning was not identified as a problematic area specifically, the teaching and learning sequence was created based on previous best practice, while focusing on using the statistical investigative cycle as envisioned in the new curriculum (Ministry of Education 2007). An initial linear hierarchical categorisation system was proposed for judging investigative questions (Arnold 2008) based on initial evidence in the students' post-tests [for a full account of student pre- and post-tests see Arnold (2008)].

8.4.2 Teaching Experiment Two

In the second teaching experiment the problematic situation, what makes a good investigative question, was initially addressed. During the teaching experiment the teacher focused on ensuring that the variable and the target population were clear in the question and that the question was asking about “some type of relationship or comparison” (Teacher A, 2008, lesson 2). In summarising questions within the statistical investigation cycle, three points were noted: (1) posing investigative questions requires students and teachers to have a clear idea of what the variable(s) are in which they are interested; (2) what they want to do (summarise, compare or relate); and (3) what the population of interest is. The planning involved deliberately teaching these criteria to the students and providing sufficient examples to allow them to practise with a number of different variables and populations. Teacher A deliberately discussed and highlighted the three criteria.

In addition to the initial lesson on posing investigative questions, the teacher decided to spend an additional lesson sorting, critiquing and improving investigative questions that had been posed by others. This involved the students first sorting the questions into the different types (summary, comparison and relationship) and then improving the investigative questions by making sure the investigative questions met the three criteria given by the teacher. In this lesson a number of points

Criteria
<ol style="list-style-type: none"> 1. The variable(s) of interest is/are clear and available 2. The population of interest is clear 3. The intent is clear 4. The question can be answered with the data 5. The question is one that is worth investigating, that it is interesting, that there is a purpose 6. The question allows for analysis to be made of the whole group

Fig. 8.2 Criteria for posing investigative questions

were mentioned by the teacher that have subsequently been linked to posing a good investigative question or understanding the question posed.

- The teacher mentioned several times during the lesson the need to consider whether the question was worth investigating. This links to Graham's (2006) second consideration (see Sect. 8.2.3).
- The actual variable that could be investigated was clarified; for example, they were not investigating foot size; they were investigating right foot length.
- The use of comparative words when posing comparison questions was explored to clarify the type of question; for example, using longer, taller or faster. Linked to this was the use of the appropriate comparing word (precise language) ; for example, use longer for right foot length, but not for right foot width (in this case they would use wider).

Between the second and third teaching experiments there was extensive dialogue between the researcher and colleagues at the university based largely on the retrospective analysis of student responses in the post-test from teaching experiment two. This dialogue addressed language and the preciseness of wording, in particular, the use of the article words *a* and *the* in investigative questions and the implications of these as to which group the question was about. Through this dialogue and through analysis of student responses, particularly poorly posed investigative questions, other ideas of suitable criteria for "What makes a good investigative question?" were generated. At this point six criteria were established (see Fig. 8.2) for what makes a good investigative question. These combine the three features the teacher used in the second teaching experiment, moderating questions from the first teaching experiment (Arnold 2008), and detailed analysis of the investigative questions that students posed in their pre- and post-tests.

The researcher then trialled some teaching ideas with a year nine (ages 13–14) class at another school to test how the criteria might be introduced to students. This was not recorded as it was not part of the research for which permission to video was granted, but it did provide an opportunity to trial some of the material before using it in the third teaching experiment.

8.4.3 *Teaching Experiment Three*

In the third teaching experiment criteria for what makes a good investigative question were used and the teaching focused on the underlying conceptual knowledge needed to understand the investigative question. The teaching and learning activities around posing investigative questions in the third teaching experiment built on the work from the second teaching experiment. The teacher (Teacher B, 2009) had been exposed to much deeper thinking about posing investigative questions prior to teaching the statistics unit. This had included workshops for all the teachers in the school on the material, and the teacher was a member of a linked Teaching and Learning Research Initiative project team (Pfannkuch et al. 2011).

Students posed investigative questions in class and then a selection of these was used in the following lesson. During the following lesson the teacher asked the students to sort questions that they had previously posed. The students identified which questions they thought were investigative questions and which ones were not. The students came to the conclusion that they did not like most of their questions. Through a teacher-led discussion the students generated ideas that aligned with the criteria for what makes a good investigative question. Students felt that the questions they had been given were not suitable as investigative questions: the question was not able to be answered because the variable was not one of the variables available in the given data set; there was not enough data to answer the question; and some questions were about an individual and not the whole group, which the students felt was unacceptable. Generating the criteria from student discussion and their findings was a deliberate strategy rather than the teacher just giving the criteria. An additional activity was used later in the topic where students critiqued questions that had been posed by others and improved on them based on the developed criteria.

As the teacher moved into new concepts, such as sampling, she started always with an investigative question, which was posed collectively as a class and checked against the criteria that had been established. In addition in later lessons on using samples to answer investigative questions about populations, care was taken to reinforce the actual population about which the students were posing and answering investigative questions. A fictitious school was invented and data cards for each “student” were created to help to develop the concept of population and sample. The “population”, Karekare College students, was constantly referred to, and this population was also physically shown as the data cards in a bag (see Fig. 8.3). This material representation of the population, coupled with the actual drawing of samples from the bag, was designed to reinforce the connection between sample and population and the investigative question.

In a wrap-up session the students again came back to the criteria about what makes a good investigative question, and, as well as posing investigative questions themselves, they had to critique questions posed by others. During this activity, an interesting observation was made by one of the students to another student in the group that was being observed: “Have you noticed that all the good ones are really long?” (2009 student, final lesson).



Fig. 8.3 Karekare College population bag with data cards (Arnold 2013, p. 152)

8.4.4 Teaching Experiment Four

In the final teaching experiment the teacher's (Teacher B, 2011) approach to posing investigative questions was different than in the previous experiments. She gave the students questions that had been posed by others before she required them to pose their own. The activity, where the students had to sort a number of investigative questions into groups, provided a catalyst to talk about what questions were good questions and what questions were not. From this discussion some of the criteria that had previously been established by the research were re-established by the students. That is, the students and teacher developed the criteria based on the class discussion about the questions they were sorting. Criteria that the students came up with included that the question needs to be about the overall distribution of the data, it must be interesting, and the variable and group need to be stated. Student reflection at the end of the lesson elicited a further criterion that had not been mentioned in class: that the type of question needed to be clear. At this point the teacher resisted the urge to "finish" the criteria (the students had identified five of the six criteria) and left the sixth criterion for when it naturally arose in the teaching and learning sequence.

Defining the context, i.e. the variable and the population, became a focus, and throughout the unit the teacher constantly asked the students to define the variable and the population for each situation. This was also linked to moving from questions about "these" students (the sample) to questions about the population. An example of the teacher helping the students define the variable is given in the excerpt below. It occurred in a lesson where students were exploring a situation where survey participants had ranked themselves as to how good they thought they were at a particular subject; for example, maths, reading, sport and the arts. The discussion was around exactly what the variable is, i.e. is it boys rating themselves higher than girls rate

themselves, or is it boys rating themselves as better when they compare themselves to girls?

Teacher: The question they were asked was how good do you think you are at maths. That was the question that they were asked. That was the survey question. ... How good do you think you are at maths? So remember we're comparing the boys and the girls. So when we're posing an investigative question we're looking at the first one, so those were the survey questions. The investigative question can someone give it to me, the first one?

Student: I wonder whether boys tend to think that they are better at maths than girls. Year 4–13 boys and girls.

Teacher: Year 4 to 13 New Zealand boys tend to think ...

Student: They are better at maths than girls.

Teacher: They are better at maths in this case than year 4 to 13 New Zealand girls.

Student: No not think they are. Because the boys wonder if they're better than the girls.

Teacher: Remember the question wasn't "Are you better than girls?", it's just how good you think you are so it's not rating against the other. But in the overall rating.

...

Teacher: What did we say up there? Boys rate themselves better at maths than girls. The boys aren't rating themselves compared to girls, it's just when they rate themselves, boys' ratings tend to be higher than girls' ratings. So the question could have been: "I wonder whether ratings for maths ability by year 4 to 13 New Zealand boys tend to be higher than ratings for maths ability by year 4 to 13 New Zealand girls."

In addition to the discussion regarding how to frame or describe the variable, the teacher was clearly differentiating between the two types of questions that are posed, i.e. survey questions and investigative questions. It is also worth noting her use of the phrase "tend to" for comparison questions. This phrasing had become part of the teacher's natural language she used in relation to comparison questions, a key element for a "good" comparison investigative question (see Fig. 8.4, and also links to criteria 6, Fig. 8.2).

The teacher persisted throughout the unit of work reinforcing the criteria for what makes a good investigative question, for example, getting the context sorted out by getting the students to correctly define the variable(s) and the population (criteria 1 and 2, Fig. 8.2), and making the questions about the population not the sample (criteria 2, Fig. 8.2). In addition she required them to make predictions of what they expected, particularly in the comparison situation, asking students all the time which group they thought would be bigger, taller or faster (links to criteria 3, 4 and 6, Fig. 8.2). The implication of these predictions was about how the comparison question was framed—for example, did they have boys taller than girls or girls taller than boys?—with the expectation being that the question was framed so that it aligned with what the students expected to be true. So if they thought boys were taller than

Question category	Student question example	Commentary
A. Nonsense, not related or not a comparison question.	I wonder Auckland region and Wellington region have the same student in year 10? (2009 student, post-test)	This question is irrelevant and does not meet criteria 1, 3, 4 and 6 (Fig. 8.2).
	I wonder if the popliteal length relates to armspan. (2009 student, post-test)	This is a relationship question.
B. A question that is partially related to the data, but not answerable by the given data (either due to sample size issues or variable not in the data set).	I wonder if all the ambidextrous students are capable of kicking a ball with both left and right foot. (2009 student, pre-test)	Handedness was in the data set provided; however, there wasn't a question about ambidextrousness for "footedness". Therefore criteria 1 and 4 (Fig. 8.2) are not met.
	If Asian girls have a longer armspan than Indian boys. (2007 student, post-test)	In the 2007 post-test there were only two Indian boys and two Asian girls. Does not meet criteria 4 (Fig. 8.2).
C. A question that hints at comparison.	I wonder if more year 10 boys are physically fit than year 10 girls. (2011 student, post-test)	This question suggests comparison, though as it reads it is probably only comparing a couple of categories, therefore not meeting criteria 6 (Fig. 8.2).
	I wonder if ambidextrous hand writers can speak different languages. (2009 student, pre-test)	This question hints at comparing the number of languages spoken across handedness, and therefore not meeting criteria 3 and 4 (Fig. 8.2).
D. A question that has all of one group bigger/smaller than all of another group or compares an individual.	I wonder if all girls have longer hair than all boys. (2008 student, pre-test)	A good example of the type of thinking, and therefore the type of question, where students are thinking something is bigger and think all of one is bigger than all of the other. They have not yet grasped the idea of tendency or tending to be bigger/longer.
	I wonder if the average resting rate for a boy is lower than a girl? (2011 student, pre-test)	Comparing a boy with a girl, comparing individuals.
E. A question that compares categorical data.	I wonder if secondary students that live in southland region are fitter than secondary students from Auckland region. (2009 student, post-test)	In the data set given, the variables that might be used to answer this question were both categorical, region they live in, and fitness levels (unfit, a little fit, ...).
F. A question that compares a summary statistic.	I wonder if the typical right foot length for year 11 boys is greater than the typical right foot length for year 11 girls from the 2007 NZ CensusAtSchool database. (2008 student, post-test)	This question is comparing "the typical", which is interpreted as a summary statistic; for example, the median or the mode.
	I wonder if the average hair length of 16 year old girls is greater than the average hair length of 16 year old boys. (2008 student, pre-test)	This question is comparing the average, which could be median, mean or mode.
G. A question that assumes the idea of tendency. This includes questions that ask how much bigger or if there is a difference.	I wonder if secondary girl students have bigger wrist circumference than secondary boy students. (2009 student, post-test)	This question uses the phrase "have bigger" but, unlike the example in category D, they haven't indicated that they are thinking all girls bigger than all boys, so this style of question has been categorised as assuming tendency.
	I wonder if boys have longer popliteal lengths than girls. (2009 student, post-test)	A second example showing a different variable; commentary above relevant for this question.
H. A question that includes the idea of tendency; for example, question includes words or phrases such as on average, generally or tends.	I wonder if boys in year 10 tend to be taller than girls in year 10. (2009 student, post-test)	This question structure has one population tending to be taller/heavier OR have a longer/shorter [variable] than the other population.
	I wonder if on average right handers have longer hair than left handers. (2008 student, pre-test)	This is a similar structure to the first, but instead of using "tend", they have used "on average".
	I wonder if Yr 9-13 NZ boys have typically higher pulse rates compared to Yr 9-13 NZ girls. (2011 student, post-test)	This is a similar structure to the first also, but this time they have used "typically" to express the idea of tendency.
	I wonder if the popliteal length of Yr 9-13 NZ girls tend to be longer than Yr 9-13 NZ boys popliteal length (2011 student, post-test)	This question structure has the variable (of one of the groups) tending to be bigger/smaller than the variable (of the other group), a different structure to the previous three.

Fig. 8.4 Comparison question examples (Arnold 2013, pp. 119–120)

girls, then the question was framed that way. A key realisation from this research was that students were conjecturing based on their general knowledge about which group would tend to have bigger values. The students were not explicitly aware they were making such a conjecture, but their posed questions strongly suggested that they were. The teacher was drawing on a new insight from the second and third teaching experiments.

8.5 Retrospective Analyses

Two findings came out of the retrospective analyses of student-posed investigative questions, these are: (1) criteria for what makes a good investigative question (Fig. 8.2, not the focus of the research question for this chapter); and (2) a detailed two-way classification matrix for comparative investigative questions that are posed (Sect. 8.5.4). The teaching experiments described above provide evidence of changes made to the teaching experiments as part of the ongoing retrospective analysis between cycles. This section describes in detail the retrospective analysis in relationship to the research question: What level of comparative investigative questions are year 10 (ages 14–15) students posing? The main sources of data were student pre- and post-test responses. In the pre- and post-tests the students were given a multivariate data set with 13 variables. Examples of discrete variables were: gender, year level at school and fitness level. Examples of continuous variables were: arm span, popliteal length (length from behind the knee to the floor, when the leg is bent at a right angle) and resting pulse rate. Students were asked to pose three comparison investigative questions.

8.5.1 *Classification of Comparison Investigative Questions*

A possible framework for comparison investigative questions was developed based on findings in the first teaching experiment (Arnold 2008). This initial framework considered questions that were not answerable with the data given and questions that were answerable, and the inclusion of the population signalled a higher level of question. This initial framework proved to be too simplistic, as it was found that the population descriptor required its own set of categories.

Comparison question categories were updated from the initial framework (Arnold 2008) following the second and third teaching experiments where student responses generated new categories. The categories were updated further following the fourth teaching experiment, as student responses signalled a need for further new categories. Figure 8.4 gives the final comparison question categories that were proposed for all year levels up to and including year 11. From year 12 onwards students have developed additional statistical knowledge which allows for more sophisticated investigative questions. This is not discussed in this chapter. Included in Fig. 8.4 are examples

for each of the different categories and commentary about the example to aid the reader. As signalled previously, the population descriptors (Criteria 2, Fig. 8.2) are not included. A good comparison investigative question needs to meet criteria 1, 3, 4, 5 and 6 from Fig. 8.2. Categories A–C are not comparison investigative questions, categories D and F are moving towards good comparison investigative questions, category E captures categorical data, which was considered inappropriate for the particular level and curriculum focus. Categories G and H are considered good comparison investigative questions with H being better than G.

8.5.2 Reflection on Final Framing of Comparison Investigative Questions

Two reflections on the final framing of the comparison investigative questions need to be mentioned. Firstly, the use of “tend to” to describe the idea of comparison, where one group “tends to be higher” than the other for a given variable, was signalled right from the start of the work on posing investigative questions (Pfannkuch et al. 2010). Researching students’ thinking about comparison situations (Pfannkuch 2006; Pfannkuch and Horring, 2005) had already identified “tend to” as being an important consideration in teaching thinking about the question framing for comparison situations.

Secondly, from the second teaching experiment to the third teaching experiment the framing of the question used in the pre- and post-tests moved from “I wonder if Year 11 NZ boys tend to have shorter **hair** than Year 11 NZ girls?” to “Does the **hair length** of Year 11 NZ boys tend to be shorter than the **hair length** of Year 11 NZ girls?” Both are acceptable as suitable comparison questions at this curriculum level, but the second question puts the variable (of the populations) clearly as the item that is being compared.

8.5.3 Population Descriptors

In the initial classification for comparison questions, the top category, H, was identified as being a “good” question and has the population included in the question. As student pre- and post-test responses were analysed from the second (2008) and third (2009) teaching experiments, it became clear almost immediately that the “super” category of population was not going to work. Students who had similar types of questions had a wide range of populations. For example, in the 2008 post-test 22 of the 24 students posed an investigative question about one group being taller than another group. Aside from the variation in the question format, 14 different population or group descriptors were used. The descriptors fell into three main categories:

(1) boys and girls (four variations); (2) various combinations of age groups (five variations); and (3) year 11 boys and girls (five variations).

Within the three broader categories there are multiple ways that students could phrase a descriptor based around whether they acknowledged that the broader population is New Zealand students and that the sample was taken from a particular CensusAtSchool database. It could be possible to make a fine graded scale for population descriptors, but pragmatism and what would be useful to teachers and students meant that fewer categories were better than more. Initially there seemed to be three clear categories: (1) Broad student population; for example, boys, girls, students (very general, could mean all boys and girls in the world); (2) Broad New Zealand student population; for example, New Zealand boys, New Zealand students (better than 1, doesn't consider the target subgroup of New Zealand boys and girls); (3) Actual New Zealand student population; for example, New Zealand year 10 students, New Zealand year 11 students, New Zealand secondary school girls.

However, as can be reasonably expected, student responses did not fall nicely into the three categories. Where, for example, did year 11 boys and year 11 girls fit? Clearly it is more specific than New Zealand students, but it doesn't specify New Zealand. An additional category was needed between broad New Zealand student population and actual New Zealand student population. Two other types of questions occurred that did not fit within these four categories. In the first type, students went broader than boys and girls but didn't use a specific population descriptor; for example, they asked about typical heights of males and females or of people. The second type of population descriptor that didn't fit into the four categories was when students specifically or inadvertently posed their investigative question about the sample. Examples of the second type of question are: "What are typical heights of **these** year 11 students?" and "What are typical heights for year 11 students sampled from the 2007 NZ C@S database?" Hence six population categories were considered as part of the overall question classification. These categories were confirmed through analysing the questions posed in the fourth teaching experiment. The final six population categories are:

1. Referring to the sample.
2. Broad population, not specifying students.
3. Broad student population; for example, boys, girls, students.
4. Broad New Zealand student population; for example, New Zealand boys, New Zealand students.
5. Any relevant student population that can be generalised about from the actual New Zealand student population used; for example, year 11 students, teenagers, secondary school girls.
6. Actual New Zealand student population; for example, New Zealand year 10 students, New Zealand year 11 students, New Zealand secondary school girls.

8.5.4 Two-Way Classification Framework (Matrix)

In order to classify a posed investigative question, the two categories, (1) question category and (2) population descriptor category, need to be considered as the student is working with both aspects (categories) at the same time. The combination of the two aspects gives rise to a two-way classification framework for comparison investigative questions. The framework is an 8 by 6 matrix (Fig. 8.5) made up of the eight question categories (rows, Fig. 8.4) and six population descriptor categories (columns, listed above).

The shaded portion of the matrix in Fig. 8.5 shows where the two aspects combine to give all the combinations to describe the investigative questions posed; for example, **H6** (in Fig. 8.5) is a comparative investigative question that includes the idea of tendency and has the actual New Zealand student population. The two-way classification framework developed during the retrospective analysis allowed for data to be gathered from each student to answer the research question—*What level of comparative investigative questions are year 10 (ages 14–15) students posing?*

8.6 Data Analysis

As described previously, students were asked to pose three comparison investigative questions in the pre- and post-test. These questions were each individually graded according to the comparison question category and the population descriptor category. For example, “*I wonder if the popliteal length of Yr 9–13 NZ girls tend to be longer than Yr 9–13 NZ boys popliteal length*” (student, post-test) was graded as **H6** because as a comparison question it *includes* the idea of tendency and it also has the actual New Zealand student population correct. On the other hand, “*I wonder if boys have longer popliteal lengths than girls*” (student, post-test) was graded as **G3** because it *assumes* the idea of tendency and has only specified a broad student population (boys and girls).

		Population categories					
		1.	2.	3.	4.	5.	6.
Comparison question categories	A.	A1	A2	A3	A4	A5	A6
	B.	B1	B2	B3	B4	B5	B6
	C.	C1	C2	C3	C4	C5	C6
	D.	D1	D2	D3	D4	D5	D6
	E.	E1	E2	E3	E4	E5	E6
	F.	F1	F2	F3	F4	F5	F6
	G.	G1	G2	G3	G4	G5	G6
	H.	H1	H2	H3	H4	H5	H6

Fig. 8.5 Comparison investigative question matrix (Arnold 2013, p. 125)

The grading system gave 48 different possibilities when the question categories and the population descriptor categories were combined. In order to look at the difference from pre- to post-test, the 48 possibilities were simplified into six overall grades (see Fig. 8.6) using the SOLO taxonomy (Biggs and Collis 1982). The grades were based on the category of the question (A to H) and the population category (1 to 6). Pre-structural to extended abstract responses were scored from 1 to 5. Hence the above **H6** grade was scored as extended abstract or numerically as 5 while the **G3** grade was scored as multi-structural or numerically as 3.

A final pre-test and a final post-test score were determined for each student by finding the mean of their three SOLO grades for the three questions they posed. These final scores were analysed to look at the difference between pre- and post-test. Figure 8.7 shows the pre- and post-test questions posed by three different students chosen to give a range of responses, to demonstrate the grade given for each question and to show the subsequent SOLO score. For each question a student grade (a combination of comparison category and population) is given, for example E3, and their SOLO score for the question. This is summarised in the first column with their mean pre-test score, mean post-test score and the difference between the pre- and post-test means.

Student A moved from a combination of questions that were mostly non-comparison questions with a general student population to posing comparison investigative questions that include both the idea of tendency and the target population. Student A moved from pre-structural/uni-structural thinking to extended abstract thinking. Student B on the other hand was posing comparison questions, either comparing categorical variables or assuming the idea of tendency, but using a general student population. Student B moved to generally better population descriptors and also having more questions that included the idea of tendency, from multi-structural

SOLO taxonomy level	Grade	Description of evidence
No response or idiosyncratic	0	Questions that are not comparison questions, nonsense or not-related questions. Category A questions.
Pre-structural	1	Questions that are partially related to the data, but not answerable by the given data. Category B questions, any population.
Uni-structural	2	Questions that hint at comparison or have all of one group bigger/smaller than the other. Category C and D questions, any population.
Multi-structural	3	Questions that compare categorical data. Category E questions, any population. Relational or extended abstract categories (F, G and H) with population categories 1–4.
Relational	4	Questions that compare summary statistics or assume the idea of tendency, including the idea of difference. Population is “acceptable”. Category F and G questions with population categories 5 and 6.
Extended abstract	5	Questions that include the idea of tendency. Population is “acceptable”. Category H questions with population categories 5 and 6.

Fig. 8.6 SOLO criteria for grading comparison investigative questions (Arnold 2013, p. 130)

Student	Pre-test responses	Post-test responses
Student A	I wonder whether the gender affects your fitness level (e.g. Are boys fitter than girls)? [E3, 3]	I wonder if the yr 9-13 NZ boys tend to have larger neck circumference than the yr 9-13 NZ girls. [H6, 5]
Mean pre-test: 1.7	I wonder whether the armspan length is meant to be at a certain length whether you are a boy or girl? [C3, 1]	I wonder if the yr 9-13 NZ girls tend to have longer armspans than the Yr 9-13 NZ boys [H6, 5]
Mean post-test: 5	I wonder whether the ring finger of the students are meant to be smaller than the index finger or not? [A3, 0]	I wonder if the yr 9-13 NZ boys tend to have larger popliteal lengths than Yr 9-13 NZ girls [H6, 5]
Difference: 3.3		
Student B	I wonder if the boy's wrist will be larger than the girls [G1, 3]	I wonder if yr 9-13 boys ringfinger at census [at] school tend to be longer than the yr 9-13 girls ring finger at census [at school] [H6, 5]
Mean pre-test: 2.7	I wonder if more girls are less fit than boys [C3, 2]	I wonder if yr 9-13 girls tend to speak more languages than yr 9-13 boys at census [H6, 5]
Mean post-test: 4.7	I wonder if girls are able to speak more languages compared to boys [G3, 3]	I wonder if yr 9-13 girls resting pulse is higher than yr 9-13 boys resting pulse at census [G5, 4]
Difference: 2		
Student C	I wonder what level of fitness most teenage boys are at [A3, 0]	I wonder if 2009 NZ C@S boys tend to have a longer armspan than 2009 NZ C@S girls [H4, 3]
Mean pre-test: 0	I wonder what the average length of your index finger is for a teenage boy [A3, 0]	I wonder if 2009 NZ C@S girls tend to be more fit than 2009 NZ c@S boys [E4, 3]
Mean post-test: 3	I wonder what the average pulserest is for teenage girls [A3, 0]	I wonder if 2009 NZ C@S boys tend to have a longer index finger length than 2009 NZ C@S girls [H4, 3]
Difference: 3		

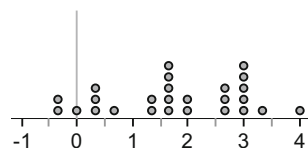
Fig. 8.7 Examples of student posed comparison investigative questions pre-test and post-test

thinking to extended abstract thinking. Student C initially was posing questions that were summary type questions, suggesting she did not understand what was meant by comparison questions. In the post-test student C was posing comparison questions and mostly ones that included the idea of tendency. This student still needed to work on her population descriptors because in all instances she was using the broad New Zealand student population rather than the target population.

8.7 Findings

From the class of 29 students, 26 students completed both the pre- and post-test. The findings are now discussed. Figure 8.8 shows the difference between students' pre-test mean score and their post-test mean score. A difference of two indicates that the student had a mean improvement of two points over their three comparison questions.

Fig. 8.8 Graph of difference between post-test mean score and pre-test mean score (Arnold 2013, p. 132)



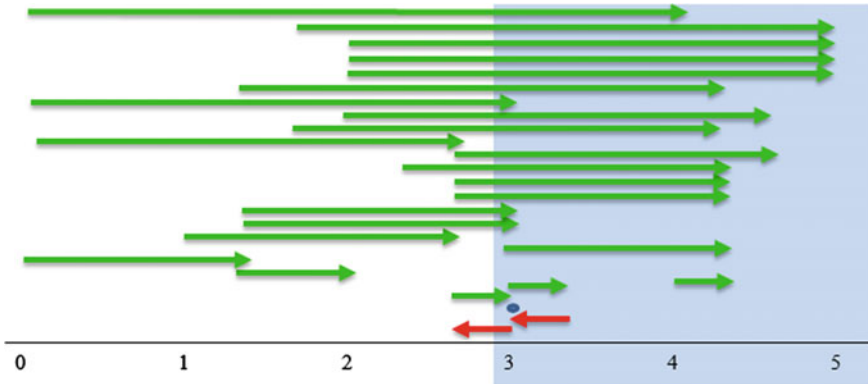


Fig. 8.9 Graph showing student movement from pre-test mean score to post-test mean score (red signals a negative movement—arrow pointing to the left, the circle signals no movement, green signals a positive movement—arrow pointing to the right) (color figure online)

Figure 8.9 shows the actual movement, from pre-test mean score to post-test mean score. The shaded grey area signals at least a multi-structural response on average. Of the 26 students 21 were posing at least at a multi-structural level on average in the post-test. The students in Fig. 8.9 are the same students as in Fig. 8.8. Working from left to right in Fig. 8.8 matches the students from bottom to top in Fig. 8.9.

Of the 26 students that sat both the pre- and post-tests, 23 improved their mean score (green/right pointing arrow in Fig. 8.9), one remained the same (circle in Fig. 8.9), and two lowered their mean score (red/left pointing arrow in Fig. 8.9). In the post-test, four students were working overall at extended abstract level, 10 at a relational level, seven at a multi-structural level; one at a pre-structural level, and four at a uni-structural level. The four uni-structural students all had a least one good question amongst their three, but were let down by a combination of the population category being low or one of the questions not being a comparison question. The pre-structural student asked questions that were about individuals (a boy, a girl) and also one non-comparison question. The students made significant improvement (p -value < 0.0001 , paired t -test) in their mean scores from pre- to post-test question posing and on average increased their mean grade by 1.78 points (95% CI = [1.29, 2.28]).

Analysis of the different types of questions the 2011 students posed in their post-tests showed that for the comparison question categories (Fig. 8.10a) there were a higher proportion of questions in category H (45 questions out of 87) than any of the other categories for 2011. The population categories were also analysed across all the post-test questions (Fig. 8.10b).

No students used the sample (category 1) as the population, and only one student for one question used people generally as the population (category 2), and that was in just one question. The proportion of questions using acceptable populations (category 5 and 6) was 67.7%. While the good population descriptor did not always line up

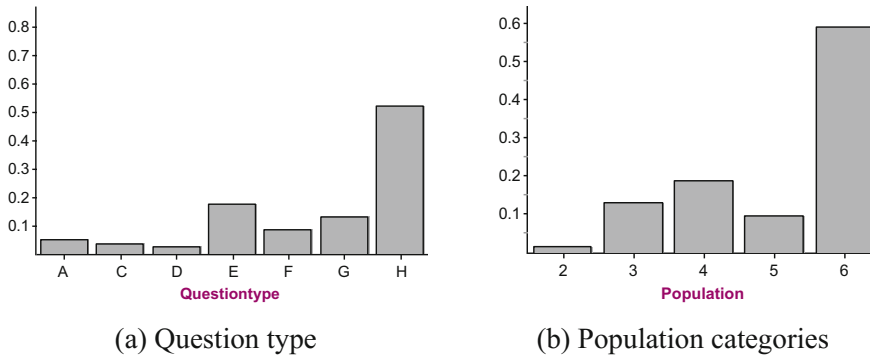


Fig. 8.10 Graphs of post-test analyses of comparison questions

with the good comparison question category, over half the question categories and two-thirds of the population descriptors were acceptable.

8.8 Discussion

The research question for this chapter was: What level of comparative investigative questions are year 10 (ages 14–15) students posing? The findings suggest that year 10 (ages 14–15) students are capable of posing comparative investigative questions that assume the idea of tendency (category H, Fig. 8.4) and have an acceptable population descriptor (Sect. 8.5.3); in other words they can pose “good” comparative investigative questions. For the 26 students who completed both the pre- and post-test, 54% were at least at this level. Most of the remaining students (27%) were posing comparative investigative questions, but their questions needed further refinement, mostly in terms of tidying up the population descriptor in the question.

There are considerations for statistics teaching and learning from the findings reported in this chapter. Firstly, the criteria (Fig. 8.2), the comparison question categories (Fig. 8.4) and the population descriptors (Sect. 8.5.3) provide structures to support teachers and students in improving their overall investigative question posing. If the quality of the question posed can be identified, for example, G4, then the improvements for the comparison question structure are given in Fig. 8.4 (G to H) and for the population descriptors (4 to 6). For students in particular, if they can become familiar with doing their own interrogation of their investigative questions against the criteria (Fig. 8.2), they will develop “thinking like a statistician” routines. Secondly, this chapter focuses only on investigative questions, the question that is asked of the data and implies a need for teachers to emphasise what a good investigative question is. Teachers also need to be discussing and highlighting the many other questions that are asked in statistical investigations, for example, survey questions, analysis questions, interrogative questions, inferential questions (Makar 2015). All

of these different question types make up the complex network of questions within the statistical investigation cycle. Thirdly, language used in investigative questions needs to be precise. Precise wording is critical (Biehler 1997; Pfannkuch et al. 2010) as “loose” or non-precise wording can cause confusion e.g. the use of *a* or *the* (see Sect. 8.4.2), and lead to poorly formed questions.

In addition, a number of statistical ideas and concepts should be developed concurrently. These include sample and population and the connection between the two, and ideas around tendency and typical. Finally, a statistical investigation is about more than just comparing or calculating simple measures; it is about students thinking distributionally, describing what they see in the sample(s) they have selected, and then making inferential statements about what may be happening back in the population(s) (Pfannkuch et al. 2010).

8.9 Implications

This research has identified gaps in the research knowledge base on posing statistical questions and consequently the big concepts underpinning the posing of good investigative questions that are needed for teaching and learning statistics at curriculum level 5 (ages 13–15) in New Zealand. This research into posing investigative questions has already had a huge impact in New Zealand classrooms and beyond year 10, curriculum level 5 (ages 13–15). Posing investigative questions is a key aspect of many of the statistics achievement standards in the national assessments, and the term investigative question is now widely used. Criteria for what makes a good investigative question, along with summary and comparison question categories, are available online as a support for teachers.

Implications for teachers include having the opportunity to experience the teaching and learning material in order to support their understandings of the research findings. Ideally this needs to happen before they take the material into their classrooms to use with their students. The sharing of the findings can support the teachers in the same way as it is hoped they will help their students. In addition supporting teachers to understand the different purposes of questions in statistics or how the use of precise language is important and making them aware of the potential confusions in language use for students would also be essential components of any work with teachers.

There is an urgent need to upskill teachers in their knowledge of the conceptual foundations required for posing good statistical questions. Many mathematics and statistics teachers are mathematics—not statistics—trained, or trained years ago. Either way, the statistics of today is not the statistics of their schooling or university days. It requires new knowledge and new ways of thinking. It also requires new ways of teaching, from a focus on the skills and calculations of the old statistics curriculum to a focus on the statistical reasoning and thinking that is inherent in the new New Zealand statistics curriculum as well as in the curricular guidelines for many other countries, such as GAISE in the United States (Franklin et al. 2005).

This chapter has addressed the problematic situation around what makes a good investigative question, including the underpinning concepts that are needed to support the teaching and learning. Suggested further research could include a focus on interrogating the statistical investigative cycle to find out what aspects should be a focus for students ages 13–15 or at other ages. Another suggestion for future research could be to explore students asking analysis questions. For example, what thinking prompts do students need to have when they are starting to analyse their data? Also because this research focused on comparison investigative questions and related research has explored summary investigative questions, three further areas of research could be posing relationship, time-series, and two-way table (two categorical variables) investigative questions.

References

- Arnold, P. (2008, July). *What about the P in the PPDAC cycle? An initial look at posing questions for statistical investigation*. Paper presented at the 11th International Congress on Mathematical Education (ICME-11), Monterrey, Mexico. <http://tsg.icme11.org/document/get/481>.
- Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* (Doctoral thesis). Retrieved from <https://researchspace.auckland.ac.nz/handle/2292/21305>.
- Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, The Netherlands: Freudenthal Institute.
- Bakker, A., & van Eerde, D. (2015). An introduction to design-based research with an example from statistics education. In A. Bikner-Ahsbals C. Knipping & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 429–466). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-017-9181-6_16.
- Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics. Proceedings of the International Association for Statistical Education Round Table Conference*, July, 1996. Granada, Spain (pp. 169–190). Voorburg, The Netherlands: International Statistical Institute.
- Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.
- Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2.
- Burgess, T. (2007). *Investigating the nature of teacher knowledge needed and used in teaching statistics* (Doctoral thesis). Retrieved from <http://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Burgess.Dissertation.pdf>.
- Cobb, P. (2000). The importance of a situated view of learning to the design of research and instruction. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 45–82). Westport, CT: Ablex.
- delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79–95). Dordrecht, The Netherlands: Kluwer.
- Franklin, C., & Garfield, J. (2006). The GAISE project. Developing statistics education guidelines for grades Pre-K–12 and college courses. In G. Burrill & P. Elliot (Eds.), *Thinking and reasoning*

- with data and chance: *Sixty-eighth yearbook* (pp. 345–375). Reston, VA: National Council of Teachers of Mathematics.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2005). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.
- Graham, A. (2006). *Developing thinking in statistics*. London, England: Paul Chapman.
- Hancock, C., Kaput, J., & Goldsmith, L. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist*, 2(3), 337–364. https://doi.org/10.1207/s15326985ep2703_5.
- Konold, C., & Higgins, T. (2002). Highlights of related research. In S. J. Russell, D. Shifter, & V. Bastable (Eds.), *Developing mathematical ideas: Collecting, representing, and analyzing data* (pp. 165–201). Parsippany, NJ: Dale Seymour.
- Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.
- Lehrer, R., & Romberg, T. (1996). Exploring children’s data modeling. *Cognition and Instruction*, 14(1), 69–108.
- MacKay, R., & Oldford, W. (1994). *Stat 231 course notes fall 1994*. Waterloo, Canada: University of Waterloo.
- Makar, K. (2015). *Informal inferential reasoning*. Keynote presentation to the Combined Hui, Auckland, New Zealand.
- Ministry of Education. (2007). *The New Zealand curriculum*. Wellington, New Zealand: Learning Media.
- Pfannkuch, M. (2006). Comparing boxplot distributions: A teacher’s reasoning. *Statistics Education Research Journal*, 5(2), 27–45.
- Pfannkuch, M., Arnold, P., & Wild, C. J. (2011). *Statistics: It’s reasoning not calculating* (Summary research report on Building students’ inferential reasoning: Levels 5 and 6). Retrieved from <http://www.tlri.org.nz/tlri-research/research-completed/school-sector/building-students-inferential-reasoning-statistics>.
- Pfannkuch, M., & Horring, J. (2005). Developing statistical thinking in a secondary school: A collaborative curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education 2004 round table* (pp. 204–218). Voorburg, The Netherlands: International Statistical Institute.
- Pfannkuch, M., Regan, M., Wild, C. J., & Horton, N. (2010). Telling data stories: Essential dialogues for comparative reasoning. *Journal of Statistics Education*, 18(1). <http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf>.
- Russell, S. J. (2006). What does it mean that “5 has a lot”? From the world to data and back. In G. Burrill & P. Elliot (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth Yearbook* (pp. 17–29). Reston, VA: National Council of Teachers of Mathematics.
- Schwartz, D., Chang, J., & Martin, L. (2008). Instrumentation and innovation in design experiments. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 45–67). New York, NY: Routledge.
- Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26(2), 114–145.
- Whitin, D. (2006). Learning to talk back to a statistic. In G. Burrill & P. Elliot (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 31–39). Reston, VA: National Council of Teachers of Mathematics.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–265. <https://doi.org/10.1111/j.1751-5823.1999.tb00442.x>.