DEVELOPING THE STATISTICAL PROBLEM POSING AND PROBLEM REFINING SKILLS OF PROSPECTIVE TEACHERS

AISLING LEAVY
Mary Immaculate College, Ireland
Aisling.Leavy@mic.ul.ie

DANIEL FRISCHEMEIER
University of Münster, Germany
dfrische@uni-muenster.de

ABSTRACT

Recent approaches to statistics education situate the teaching and learning of statistics within cycles of statistical inquiry. Learners pose questions, plan, collect, represent, analyse and interpret data. We focus on the first step – posing statistical questions. Posing statistical questions is a critical step as questions inform the types of data collected, determine the representations used, and influence the interpretations that can be made. We report on an investigation of 158 prospective elementary teachers as they design statistical questions to support group comparisons. Support was provided through implementation of three phases of question development (think, peer-feedback, and expert-feedback). We describe the features of initial statistical questions posed, examine refinements made to statistical questions, and evaluate the effectiveness of both peer and expert feedback. Our study reveals that generating adequate statistical questions is particularly complex and requires considerable time, targeted feedback, and support. With appropriate support, in the form of peer and expert feedback provided within a three-phase question design scenario, prospective elementary teachers could generate adequate statistical questions suitable for use in primary classrooms. While this study provides compelling evidence to support the use of expert feedback, further research is required to identify the best ways to support prospective teachers in both providing and implementing peer-feedback.

Keywords: *Statistics education research; Statistical inquiry; Teacher education; Statistical questions; Collaborative work; Peer feedback; Expert feedback; Prospective primary teachers*

1. INTRODUCTION

The PPDAC model (Wild & Pfannkuch, 1999), adapted from MacKay & Oldford (1994; as cited in Wild & Pfannkuch, 1999), is one dimension of a four-dimensional framework that attempts to capture statistical thinking during data-based enquiry. This model, developed from interviews with statisticians, incorporates attention to the actions and thoughts of statisticians during the course of a statistical investigation. The PPDAC (problem, plan, data, analysis and conclusion) model characterises the phases of an investigative cycle, which is ultimately concerned with "abstracting and solving a statistical problem grounded in a larger 'real' problem" (Wild & Pfannkuch, 1999, p. 225). Thus the statistical problem is a fundamental component of statistical investigations and constitutes the first step of the 5-phase PPDAC model: problem. Consequently, statistical questions drive the direction of the subsequent statistical inquiry and directly influence the type of data collected, the representations constructed, and the analyses and conclusions derived from the data. As such, statistical questions impact the nature and quality of the statistical thinking and reasoning that occurs and, when used in the classroom, the learning outcomes of the students. It makes sense, then, to invest resources in helping teachers develop the skills necessary to pose statistical questions. Posing statistical questions, however, is not a trivial task. The difficulties associated with posing adequate statistical questions have been observed not only with school students (Pfannkuch & Horring, 2004) but also with prospective teachers (Frischemeier & Biehler, 2018). This has motivated our interest in developing understanding of the types of experiences that support prospective teachers in designing statistical questions.

2. LITERATURE REVIEW

2.1. CHALLENGES FACED WHEN POSING STATISTICAL QUESTIONS

The focus on problem posing in mathematics has gathered momentum in recent years. In particular, there is a growing acknowledgement that efforts placed in improving the quality of problems posed will, in turn, influence the nature and types of problem solving activity that occurs in classrooms. This is evident in the proliferation of research (Cai et al., 2013; Cai & Hwang, 2020; Crespo & Sinclair, 2008; Crespo & Harper, 2020; Ellerton, 2013; Silver, 2013) and references to problem posing in curriculum standards and recommendations (National Council of Teachers of Mathematics, 1989, 2000). Furthermore, efforts to incorporate problem posing experiences in initial teacher education result in an accumulation of insights and recommendations around how to best support prospective teachers in improving their problem posing skills (Chapman, 2004; Crespo, 2003; Leavy & Hourigan, 2019, 2020). The field of statistics education is in the early stages of a somewhat similar journey. Parallels can be drawn between the problem solving process and the process of statistical investigation (see Wild & Pfannkuch, 1999), wherein learners are engaging in the authentic practises of statisticians. Indeed, akin to how the problem solving process is precipitated by the posing of a mathematical problem, the posing of a statistical question prompts the cycle of statistical investigation. For example, the *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) Report* (Bargagliotti et al., 2020), published by the American Statistical Association, highlights:

> the importance of asking questions throughout the statistical problem-solving process (formulating a statistical investigative question, collecting or considering data, analyzing data, and interpreting results), and how this process remains at the forefront of statistical reasoning for all studies involving data. (p. 2)

This report has been very influential and the use of the GAISE report as a guiding framework is visible across curriculum, evaluation and research landscapes within statistics education.

In contrast to mathematical problem posing, there is a dearth of research into statistical problem posing. Of those few studies completed, most focused on school level children with very little focus on prospective teachers. In their overview of this work, Watson and English (2017) pointed out that of those studies that incorporated attention to the question formulation phase, the focus shifted quickly to the construction of survey questions to stimulate the collection of data, as opposed to focusing on the broader statistical question. There is evidence, however, that the provision of supports can bring about substantial improvements in statistical question design. Also of interest is a study that examined the ways in which 9-year-olds construct relevant and reasonable questions that can be answered with a statistical investigation (Allmond & Makar, 2010). Initially, while students were able to write statistical questions, these questions tended to be closed in nature, non-investigative, or were not feasible to investigate through the collection of data. At the conclusion of the eight-lesson unit of instruction, there were noticeable improvements in the quality of statistical questions posed. The authors identified several features that supported improvement: notably, a focus on syntax and recognition of the utility of ambiguous words that support inferential reasoning, feedback from peers that lead to the critical evaluation of questions posed, and the opportunity to examine different structural contexts of investigations (e.g., comparison, prediction, evaluation etc.). Similarly, a study of 6th grade students found that 70% could engage meaningfully in all aspects of statistical investigation, including the formulation of statistical questions (Watson & English, 2017).

In their work with secondary school students and teachers, as part of a three-year project on developing a statistics curriculum for 15-year-old students, Pfannkuch and Horring (2004) observed that students tended to neglect the link between the statistical question and the data collected. They "looked at the data and just talked about the data instead of going back to the question" (Pfannkuch & Horring, 2004, p. 208). Switching focus to the teachers, Pfannkuch and Horring found that teachers tended to pose narrowly framed statistical questions. With regard to elementary preservice teachers, Frischemeier and Biehler (2018) investigated the development of the quality of statistical questions in a course on statistical thinking for prospective teachers. Participants were asked to complete the entire PPDAC-cycle including a component involving the generation of their own statistical questions. This required them to first generate their statistical questions in pairs (think phase) and then discuss them

with their peers (pair phase) and finally with the instructor (share phase). One fundamental conclusion of Frischemeier and Biehler (2018) was that the quality of statistical investigations depended on the statistical question posed and subsequently stimulated the investigation. Not surprisingly, poor statistical questions requiring merely a yes/no answer or including only one variable, led to short and non-sophisticated statistical explorations. Although feedback was provided during two stages (peer and expert feedback) of the study, the quality of statistical questions was found not to have improved in a considerable way.

Thus, it is evident from these studies that not only does statistical problem posing present a series of challenges for learners, but the design of poor questions has implications for the quality of subsequent statistical investigations. The emerging research suggests, however, that the provision of carefully designed experiences that focus on language and syntax, and provide opportunities for peer feedback, may bring about improvements in the design of statistical problems.

2.2. DIFFERENT TYPES AND CATEGORIES OF STATISTICAL QUESTIONS

In any discussion of statistical questions, it is important to distinguish between two fundamental types of questions—investigative questions and survey questions. Arnold (2013) stated that "investigative questions are the questions to be answered using data" (p. 19). As an example of an investigative question, Arnold presented the following: "What is the situation in relation to leisure time activities of grade 5 students?" Thus investigative questions are located within the problem phase of the PPDAC cycle. In contrast, survey questions are posed when collecting data and thus are located within the data phase of the PPDAC cycle. Arnold (2013) defined survey questions as a question that "is asked to get the data" (p. 19). An example for such a survey question might be "How much time do you spend watching TV on the weekend?". One typical misconception is posing survey questions when being asked to pose investigative questions (Arnold, 2013, p. 23). In this study, we concentrate on the generation of investigative questions primarily due to their influence on the nature of subsequent activities that occur within a cycle of statistical inquiry. For example, the investigative question directs the type and amount of data collected and informs the types of analyses carried out on the data. In this paper, we refer to investigative questions as *statistical questions*.

Several characteristics can be taken into account when examining the features of statistical questions. One characteristic relates to the types of variables involved and how many of each are present. Biehler (2001) distinguished between one and two-variable statistical questions (see Figure 1 for examples). A further distinction for two variable questions was provided by Konold et al. (1997, p. 7) who categorised questions as involving the investigation of the relationship of two categorical variables, two numerical variables, or group comparison questions involving one categorical and one numerical variable (see Figure 1). While the categories in Figure 1 are descriptive features of statistical questions, a more evaluative approach is that of Frischemeier and Biehler (2018) who distinguished different qualities and developed a rating system for statistical questions (see Figure 2). In this rating system, questions that take into account two variables are rated higher than questions that take into account only one variable. In the subset of one-variable questions, they distinguished questions leading to a yes/no answer, questions asking for a specific value, and more general questions with regard to the entire distribution. For two-variable questions a similar categorisation exists, which distinguished between questions requiring a yes/no response, questions aimed at working out differences in group comparison situations, and more sophisticated and complex statistical questions (called open and complex questions) involving two variables (see Figure 2 for examples).

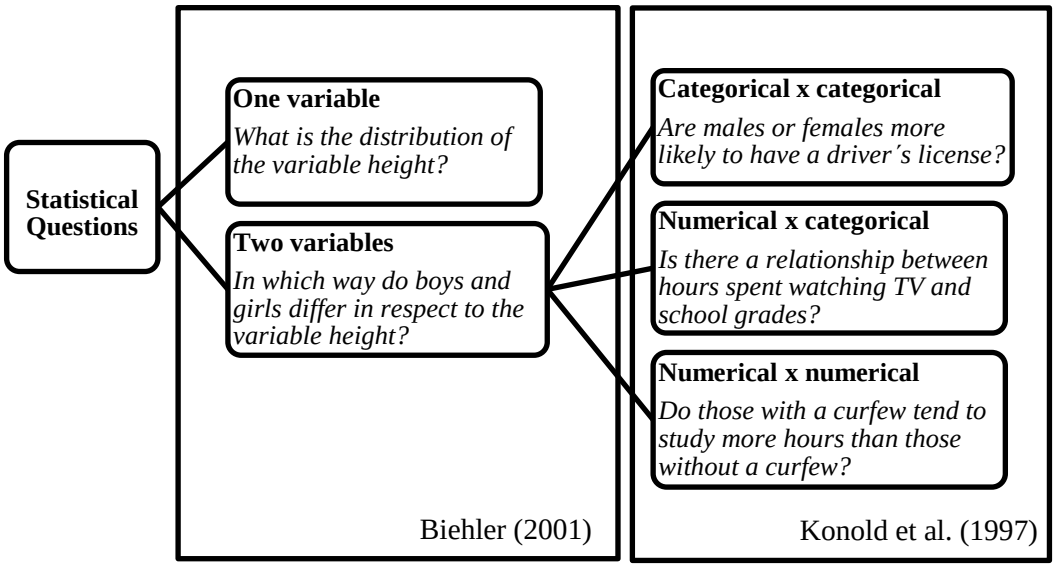

Figure 1: Categorising statistical questions according to type of variable

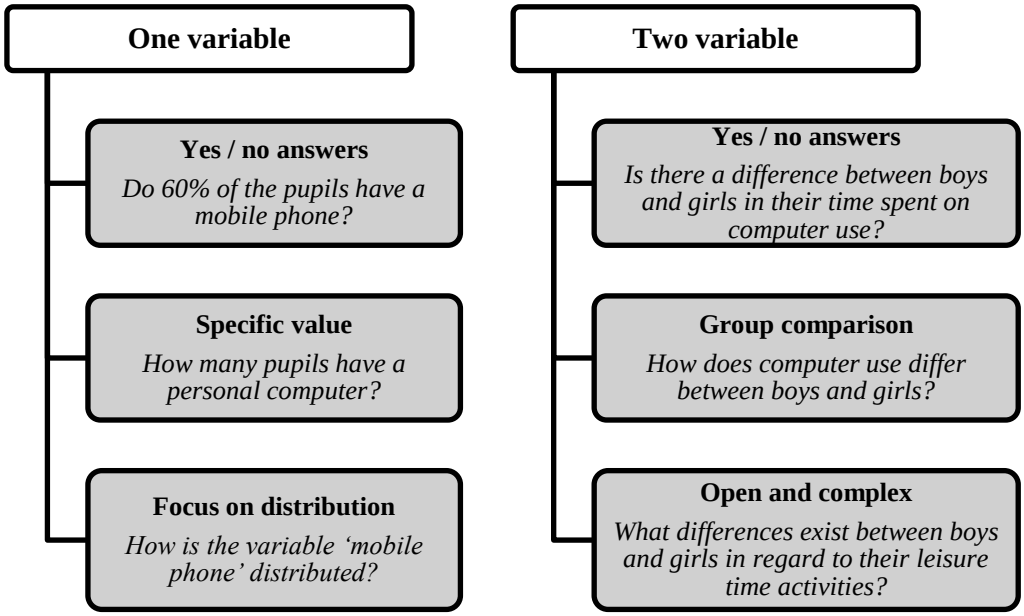

Figure 2: Hierarchy of one and two variable statistical questions (Frischemeier & Leavy, 2020)

Furthermore, Arnold (2013, p. 110-111) identified six fundamental criteria for what makes a good investigative question (in our terms, statistical question):

(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

Criterion (1) specifies clarity of the description of the variables, their availability to be measured, and correct identification. Criterion (2) focuses on the population of interest and whether learners focus on individuals, a sample, or the population. Criterion (3) emphasises clarity in the intent of the question in

terms of whether it is a summary, comparison, or relationship question. Consideration of whether the question can be answered with the given data is also part of Criterion (4). Criterion (5) highlights that the question be interesting and the information gathered from answering the question be both useful and serve a purpose. Finally, Criterion (6) investigates whether the question allows analysis with regard to a local view (single points, single aspects) or a global view of distributions (characteristics such as center, spread, skewness).

These criteria, categories, and ratings contribute to our understanding of the multiple considerations that need to be taken into account when posing statistical questions. In the context of this study, these considerations are incorporated into aspects of the design of instruction for prospective teachers (see sections 3.3 and 3.4) and inform the development of a category system for coding and evaluating statistical questions (see Table 6 and appendices A-D).

***Other question features*.** The way in which statistical questions are posed may impact how data are viewed. Bakker and Gravemeijer (2004), for example, distinguished between local and global views of distributions. Learners with a local view tend to focus on single values of a distribution whereas those developing a global view tend to take into account characteristics like center, variability, shape, and skewness when describing distributions. This view (local vs. global) might already be triggered by the statistical question posed (see Criterion 6 of Arnold, 2013, p. 111). In their study of nine year old children, Allmond and Makar (2010) referred to how children's initial use of syntax focused on the 'oneness' of the question and thus promoted a local view and closed down opportunities for inferential reasoning. They noted, for example, how subtle changes in language in the question 'how many peaches in *the* can?' to 'how many peaches in *a* can?' could shift the focus from 'oneness' to more inferential reasoning.

Questions promoting group comparisons of data. A fundamental emphasis of this study is the focus on group comparison. Comparisons of numerical data sets incorporate attention to several fundamental ideas such as data, representation, variability, viewing the data set as an entity (developing a statistical perspective), and inference (Ben-Zvi, 2004; Burrill & Biehler, 2011; Konold et al., 1997; Konold & Pollatsek, 2002). The types of understandings fundamental to group comparison have been described as critical for building the intuitive foundation for inferential reasoning (Ben-Zvi, 2004; Watson & Moritz, 1999) and for promoting a shift in student thinking toward thinking about propensities and thereby developing a statistical perspective (Konold et al. 1997). Group comparisons can be taught as early as primary school using, for example, modal clumps as pre-formal concepts of center to compare two numerical data sets (Bakker, 2004; Frischemeier, 2019; Konold et al., 2002; Watson & Moritz, 1999). Thus, as a result of the opportunities for development of statistical understandings when engaged in group comparison activities, our emphasis in this study is on the generation of statistical questions that lead to *group comparisons of data*.

2.3. THE BENEFITS OF FEEDBACK ON QUESTION DESIGN

Studies have found that opportunities to engage in conversations with peers and experts create spaces that support the ability to think critically about content. Allmond and Makar (2010) reported the benefits of peer collaboration with a class of 9-year-olds when designing statistical questions wherein "negotiations opened up new ideas, helped them to think more critically and purposefully about their investigative questions, and allowed them to refine the wording of the questions for improved clarity and meaning" (p. 5). Thus, receiving reliable feedback from peers has learning benefits both for those receiving the feedback (Falchikov & Goldfinch, 2000) and for those providing the feedback (Cho & Cho, 2011). The use of peer feedback is also emphasised by Frischemeier and Leavy (2020) in their description of three approaches—a checklist for improving statistical questions, a three-phase feedback activity with peers, and a matching game—that have been successful in helping pre-service teachers refine and improve the quality statistical questions for the comparison of data sets.

Although there is very little research examining the effect of feedback while posing statistical questions, research emanating from the field of writing education provides valuable insights into the specific features of effective feedback, both peer and expert. Feedback that incorporates summaries of work (Ferris 1997) and provides explanations to clarify the feedback's purpose (Bitchener et al., 2005)

promote feedback implementation. The incorporation of specific rather than general comments is also effective, in particular with regard to identifying the problem explicitly (Matsunura et al., 2002), offering solutions early in the task (Sugita, 2006), and locating the source or location of the problem and/or solution (Nilson, 2003). Interestingly, studies examining the role of affective language in influencing the implementation of feedback indicate that praise rarely leads to changes being implemented and generally has small effect size (Ferris, 1997; Kluger & DeNisi, 1996). The influence of mitigating language, used to make criticisms seem less harsh, is less clear with some studies indicating positive outcomes (Tseng & Tsai, 2006) and other studies indicating little positive effect (Sugita, 2006; Ferris, 1997). Insights into effective peer feedback with college level students are provided by Nelson and Schunn (2009) who examined the relationship between various types of feedback, potential internal mediators, and the likelihood of implementing feedback. Their analysis, of the peer feedback and subsequent revisions made to 50 undergraduate papers, found that feedback was more likely to be implemented if it provided a solution to perceived problems, presented a summary statement which condensed and reorganized the information presented, pinpointed the location of the problem or solution, and clarified the purpose of the feedback.

In summary, research provides some indicators of the features of effective feedback albeit not in the context of statistics education. Considered alongside expert feedback, peer feedback is an efficient and logistical support to engage students in discourse and provide feedback on course content (Cho & McArthur, 2010; Topping, 2009). Nonetheless, it appears that despite the uptake in use of peer feedback across a range of instructional settings and with a range of learners, there is a recognised lack of direction on how to facilitate students' learning of peer feedback skills and on how to address the gap between current and ideal peer feedback performance.

This research examines the statistical problem posing skills of preservice primary teachers as they pose and refine statistical problems (i.e. investigative questions) that promote group comparisons of data. Thus we focus our attention on the 'problem' phase of Wild & Pfannkuchs' (1999) investigative cycle within which statistical problem posing is located. Furthermore, we explore the influence of feedback on the evolution of statistical questions through the incorporation of opportunities for peer and expert feedback. The research questions are:

> What are the features of the statistical questions designed by prospective teachers for use within primary classrooms? In what ways do these statistical questions evolve over the course of the study?
>
> To what extent does peer feedback support the development of statistical questions?

3. METHODS

3.1. PARTICIPANTS

There were 158 preservice primary teachers involved in this study. The 118 Irish participants were 3rd year students in a 4-year undergraduate degree in primary teacher education. The remaining 40 German participants were 1st year students in a 3.5-year undergraduate degree in primary teacher education.

3.2. COURSE DELIVERY

Ireland. Irish participants were enrolled in a 12-week compulsory course on the teaching of statistics and probability: the first eight weeks focused on statistics. Participants had previously completed four compulsory mathematics education courses focused on the pedagogy of number, algebra, geometry, and measures. They were taught in groups of 30-40 students, once a week for 60 minutes. Four such groups taught by the first author were invited to participate in the research (n=160) and 118 agreed to participate. They were informed that participation involved allowing access to their coursework for the purposes of research. They were also informed that their level of participation would not influence their grade in the course and they could withdraw at any time without penalty. Ethical clearance was granted by the college ethics research committee.

Germany. German preservice teachers participated in a 14-week compulsory course on the teaching of geometry and statistics. The first 11 weeks focused on geometry and the last three weeks focused on statistics. The course consisted of a weekly lecture (90 minutes) taught by the second author and a weekly seminar (60 minutes) also taught by the second author. The participants were also enrolled in a parallel course on algebra and probability that consisted of a weekly lecture (90 minutes) and a weekly seminar (60 minutes). The 104 students enrolled in the lecture and seminar were invited to take part in the study. They were informed that their participation would allow the second author access to their course work. Furthermore, they were told that their level of participation would not influence the grade; 40 students provided consent to participate. All guidelines set forth in the informed consent agreement have been followed.

3.3. CONTENT OF THE STATISTICS COURSE

Courses in both settings were designed around the cycle of statistical investigation and focused attention on the problem, plan, data, analysis and conclusion (PPDAC) components (Wild & Pfannkuch, 1999) and its implementations in primary classrooms. A core assessment task at both institutions was the design of a statistical question (i.e. the problem phase of the PPDAC cycle), focusing on group comparisons of data and that would stimulate and drive a cycle of statistical investigation targeted for use with upper elementary grade children (age 10-12). The content and sequence of the statistics pedagogy component taught in Ireland can be viewed in Table 1 and in Germany in Table 2. Note that while the Irish course met once a week for 60 minutes, the German course consisted of a weekly lecture (90 minutes) and seminar (60 minutes). Also, while all participants were required to design a statistical question at the conclusion of the course, only the Irish group collected/produced a set of data in answer to the posed statistical question.

Table 1. Alignment of statistics course and phases of study design (Ireland)

Week	Content and focus	Study phase: Lectures in Weeks 2-4
1	Statistics versus mathematics; Designing classroom statistical investigations; Children as data detectives	*Phase 1:* Design of statistical research question [SQV1]
2	Formulating statistical questions; Designing investigations; Samples and populations; Variability in statistics	
3	Types of data; The concept of distribution	
4	Representing and analysing data using graphical displays	
5	Exploring distributions of data: formal and informal measures, a focus on shape, skew	*Phase 2:* Revision of statistical research question [SQV2]
6	Analysing distributions of data I: measures of central tendency, measures of variability	
7	Analysing distributions of data II: measures of central tendency, measures of variability	*Phase 3:* Design of final statistical research question [SQV3]
8	Comparing distributions of data: motivating questions, designing investigations, sample sizes, representing distributing, analysis, and conclusions	

Table 2. Content of the German course

Week	Content and focus	Study phase: Seminars in Weeks 1-2
1	Basics of descriptive statistics Types of data: categorical vs. numerical The PPDAC cycle: Statistical questions, Designing data collection: Population vs. sample, Data analysis: Distributions of categorical and numerical variables, Comparing distributions	*Phase 1:* Design of statistical research question [SQV1] *Phase 2:* Revision of statistical research question [SQV2]
2	Data analysis II: Creating, reading and interpreting graphical visualizations: Bar graphs, Pie charts, Histograms, Boxplots and Dot plots.	
3	Data analysis III: Calculating, reading and interpreting summary statistics - measures of central tendency, measures of variability: Mean, Median, IQR.	*Phase 3:* Design of final statistical research question [SQV3]

3.4. STUDY DESIGN

This study focused on posing and refining statistical questions. Participants were introduced to the PPDAC model (Wild & Pfannkuch, 1999) and reminded that statistical problem posing is located in the 'problem' phase of the model. They were placed in pairs and requested to design a statistical question they would refine and develop over the course of the semester. The statistical questions were targeted for use with upper elementary grade children (age 10-12) and were required to lead to *group comparisons of data*. The Irish students were also required to plan and collect data that addressed their statistical question (thus completing the first three parts - problem, plan and data - of the PPDAC cycle).

A three-phase design was used to support participants when posing and refining statistical questions. The scheduling of the phases in relation to the delivery of the course is shown in Table 1 and Table 2. The Irish group implemented the three-phase design between weeks 2-4 of the 8-week semester (see Table 1). In the German group the three-phase design was implemented in one session (60 minutes) in the seminar between weeks 1 and 2 (Table 2). While scheduling of the phases differed between both groups, the fundamental components were very similar.

Phase 1 (Think). The purpose of Phase 1 was the design of an initial statistical question (Statistical Question Version 1, SQV1). Participants worked in pairs on the design of a statistical question (SQV1) that promoted data comparison of two numerical data sets. Course content in both institutions aligned with this phase, provided examples of strong and weak statistical questions and explored four components relating to question design (see Table 3). Examples of statistical questions were developed by the course instructors, which showcased the four question components outlined in Table 3. These examples illustrated to varying degrees, attention to or neglect of these question components. Phase 1 culminated with the design of the initial statistical question (SQV1). Each pair recorded their question on a Statistical Problem Posing (SPP) form. This form had a space allocated for recording questions, receiving feedback and recording revisions to questions.

Phase 2 (Peer-feedback). The purpose of Phase 2 was to provide peer feedback on initial research questions. Each pair received an SPP form containing a statistical question (SQV1) of a different peer pair. They used a set of question prompts (see Table 3), directly addressing the four question components, to guide their analysis and feedback. Question prompts were developed taking into account the work of Arnold (2013) and Frischemeier and Biehler (2018). For example, the item "Is it meaningful?" is derived from Arnold´s Criterion 5. Participants recorded their feedback on the SPP form, beneath SQV1, in the allocated space. SPP forms were then returned and pairs of participants were given the opportunity to refine their statistical questions (SQV2).

Table 3. Question components and associated prompts when designing and providing feedback on statistical questions

Components	Question prompts
Look at the question	Is it meaningful? Will the question sustain interest and curiosity of primary children? Is the intent clear and unambiguous?
Think about the variables of interest	Is the variable described clearly? Is the variable available/possible to measure?
Look at the relationship between the question and the data it will generate	Can the question be answered with a simple 'yes/no' response [avoid these type of questions]? Will the question generate quantitative data (i.e., numbers)? Will the question motivate a focus on two data sets? Does the question promote group comparison of data?
Look at (or imagine) the data	Can you answer the question with the given data? Is there sufficient data collected to answer the question? Is there sufficient variability in data collected (is there the potential for a wide range of possible data values)?

Phase 3 (Expert-feedback). The endpoint for Phase 3 was the production of the final research question (SQV3) based on the provision of expert feedback from the course instructor. Instructors examined the SPPs containing the original statistical questions (SQV1), peer feedback on the SQV1, and the revised statistical questions (SQV2). Questions were examined according to the same four component structure, and question prompts, that guided the peer feedback (see Table 3). Experts provided feedback verbally during class time, identifying the common weaknesses observed in the SQV2s. Feedback was organised using the four question component structure and examples of SQV1 and SQV2 questions were provided as illustrations. Table 4 is an example of expert feedback presented to one of the pairs. Participants once again revised their statistical questions taking into account the expert feedback. The revised questions (SQV3) were recorded on the SPP form.

Table 4. Examples of expert feedback

Components	Feedback *Sample question demonstrating weakness*
Look at the question	The intent of the question isn't always clear. *How long have 3rd grade children had their pets?*
Think about the variables of interest	The question may not sustain curiosity as the answer is obvious. *Will a class of 30 children run a 100m distance faster when carrying their schoolbags or when not carrying their schoolbags?*
	There is no indication of the unit of measurement *Do grade 4 spend more time outside playing at the weekend or on their screens?*
Look at the relationship between the question and the data it will generate	The question is a survey question and not an investigative question *How many pairs of shoes do you own?*
	The question wording doesn't require a comparison of two data sets. *What were the winning times in the 100m sprint in the Olympics finals in the last 20 years?*
Look at (or imagine) the data	Sample size is too small to support comparison *5 different types of seeds have been planted. Each seed has equal access to sunlight and water. How will the growth of each seed differ from week 1 and week 2?*
	There is no indication of sample size, i.e., how many people are involved *Compare the amount of sweets bought on a Friday after school to the amount bought on a Monday after school.*

3.5. DATA ANALYSIS

Due to the large volume of data, one question from each of the 79 subject pairs in each of the three phases of the study totaling 79 initial statistical questions (SVQ1) and 158 revised statistical questions (SVQ2 and SVQ3), we used a structuring qualitative content analysis method (Mayring, 2015) for rating the statistical questions. The advantage of qualitative content analysis is the possibility of the reduction of a large amount of data in the form of category systems. Furthermore, coding rules and key examples are provided, which carefully guide the assignment of a precise code to each question component.

We constructed our categories mainly using a deductive approach, but also took into account inductive elements from our data as a kind of "mixed approach" (Kuckartz, 2012, p. 69). A mixed approach means, in this sense, we first took into account categories already established from existing research studies (deductive perspective). In the next step (inductive perspective), these categories were refined with regard to our findings in our data (the questions we analyzed from our participants). This also required adding new categories not mentioned previously by other research studies, but which arose out of the analyzed data. Since there had already been fundamental theoretical work (see especially, Arnold, 2013) carried out on the generation of statistical questions (categories, criteria, etc.), we mainly used the deductive perspective to take and adapt categories identified from existing research studies (Arnold, 2013; Biehler, 2001; Frischemeier & Biehler, 2018; Konold et al., 1997). The process model of this kind of structuring qualitative content analysis can be seen in Figure 3 (adapted from Mayring, 2015, p. 378).

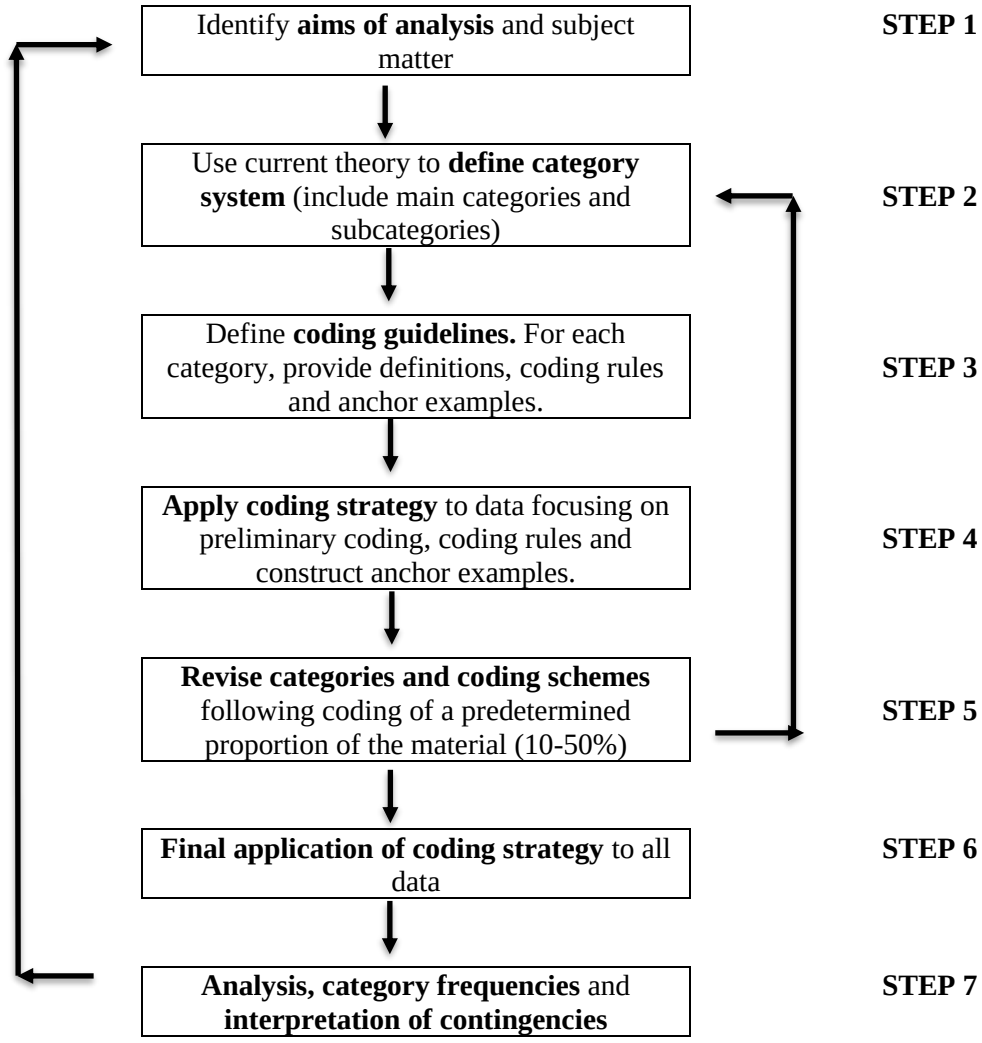

Figure 3: Model of structuring qualitative content analysis (adapted from Mayring, 2015, p. 378)

This process model displays the data analysis procedure used in this study: We first (Step 1) analyzed the subject matter and the underlying research (e.g., Arnold, 2013; Biehler, 2001) and then generated research questions (aims of the analysis). From the theory and existing frameworks (Arnold, 2013; Biehler, 2001; Frischemeier & Biehler, 2018; Konold et al., 1997) we derived the category system Step 2) and defined the coding guidelines presenting definitions of categories, anchor examples, and coding rules (Step 3). The creation of the specific categories is exemplified below and in Table 5. In a fourth step (Step 4) both authors read a sample of 30 statistical questions independently, assigned preliminary codes, and then revised our category system based on our shared observations (Step 5). We then (Step 6) applied the final version of the category system to all statistical questions [SQV1, SQV2, SQV3] collected from our participants. The analysis (Step 7) was then carried out in the form of counting the occurrence of categories and conducting interpretations about the frequencies of the occurrences with the intention of answering the research questions generated in Step 1.

We rated the quality of each statistical question [SQV1, SQV2, SQV3] for each pair of participants with regard to the four main components used in the peer and expert review feedback stage of instruction (see Table 3):

Component 1: Look at the question
Component 2: Variables in the question
Component 3: Relationship between data and question
Component 4: Look at the data

Each of the four components and their categories is described and presented in Table 5 alongside the research that guided their formation. Definitions and examples of the statistical questions generated in the study are presented in Appendices A–D.

Component 1: Look at the question. This component considers general characteristics of statistical questions and consists of five categories (Table 5). The first two categories consider whether the question is meaningful and interesting for primary school students. The third category ascertains whether the statistical question is posed in a clear and unambiguous way. As an illustration, a counter-example for a non-clear statistical question found in our data is the example “What is the probability that a boy has glasses in your class?” This question is not aimed at a statistical investigation and is neither a summary, comparison or relationship question. The penultimate category examines which kind of variables are embedded in the statistical question. The final category focuses on whether the language used in the statistical question focuses on an individual case, on a sample, or on a population. Each of the categories, their values, definitions and examples are presented in Appendix A.

Component 2: Variables in the question. The analytic framework for this component has two aspects: clarity and measurability of the variable. Variables are described clearly when the variable and its unit (in the case of numerical variables) are explicitly mentioned. A variable is considered measurable if it is present in a dataset and/or quantifiable. The variable is both clear and measurable in the question which examines the data set on freetime activities of the 3rd and 4th graders and asks “Do 3rd graders spend more time (in hours per week) on sporting activities than the 4th graders?” In contrast to this example in the question “Do 3rd graders spend more time on sporting activities than 4th graders?” the variable (time) is measurable but not clear, because the unit of time is not mentioned clearly (minutes, hours per week?). In addition to that, in the question “Is there a relationship between intelligence and eye color?”, the variable is clear, but not measurable. These categories, their values, definitions and examples are presented in Appendix B.

Component 3: Relationship between data and question. The analytic framework for this component examines the type of data generated by the posed question. To facilitate this, each statistical question is coded as belonging to one of five category types: survey/investigative, yes/no, non-binary, specific value, sophisticated. The first category discriminates survey questions from investigative questions. Survey questions, such as “What eye color do you have?”, are classified as a question type with the poorest quality. Other statistical questions such as “Are male students spending more time on the computer than female students?” can be answered with a single word like “yes” or “no”. These kinds of questions are qualitatively better than survey questions but nevertheless do not allow a

sophisticated and well-elaborated data exploration. Very similar to "yes/no" questions are "specific value" questions wherein the answer to the question is a specific value; an example of this question is "What is the percentage of tablet users in the sample?" Other questions such as "Which team has a better score?" are those posed in group comparison situations which investigate whether group A is better/larger/smaller/higher/etc. than group B. We called these kinds of questions "Non binary" questions. Finally, questions were coded as "sophisticated" if they either aim at a relationship between variables and an answer exceeding Yes/no ("How much more pocket money do 4th graders get than 3rd graders?") or if students pose a question which allows a deeper open exploration ("In which way do 3-graders and 4-graders differ in their free time activities?". We regarded "Yes-No", "Non binary" and "Specific value" questions as having equivalent quality and for the purposes of coding, we merged these questions to one category called "reasonable", so that we have at least three categories in this component "survey", "reasonable" and "sophisticated", and coded them all equivalently. Appendix C present these categories, their coding values and exemplar questions.

Component 4: Look at the data. This component focuses more closely on the statistical investigation by examining the nature of the data motivated by the statistical question. The first category examines whether the statistical question can be answered with the given data; an example is the question described earlier which asks "Do 3rd graders spend more time on sporting activities than the 4th graders?" and presents the data set on freetime activities of the 3rd and 4th graders. The next two categories evaluate whether the sample size is sufficiently large to facilitate comparison and whether there is a sufficiently wide range of data values (variability) presented or not. The sample size is sufficiently large in the data set accompanying the question "Is there a difference in weights of 6th grade school bags and 5th grade school bags in Forest elementary school?" compared to situations where small data sets were provided (in some cases consisting of less than 10 values). The final category examines whether the statistical question triggers a local or global view of the data. An example of a question motivating a local view is "What is my armspan?" compared to "How wide would an entrance have to be to allow all elementary school students pass through with their arms spread out?" which triggers a more global view of a data set. The categories, values, definitions and examples are presented in Appendix D.

Table 5. Analytic framework question prompts for providing feedback on statistical question

	Look at the question	Variables in the question	Relationship between data and question	Look at the data
Categories for examination	Is the question meaningful? [(1)]	Are the variables described clearly? [(1)]	Is a survey question posed? [(1)]	Can the question be answered with the given data? [(1)]
	Will the question sustain interest and curiosity of primary children? [(1)]	Are the variables available or possible to measure? [(1)]	Does the question require a Yes/No answer? [(4)]	Are sufficient data collected?
	Is the intent clear and unambiguous? [(1)]		Does the question require the selection of one group as the answer (i.e., as better/faster etc.)?	Is there sufficient variability in the collected data?
	Which variables are embedded in the question? [(2)] [(3)]		Does the question require identification of a specific value (ratio/percentage/mean)?	Does the question require a global view of the data?[(1)] [(5)]
	Is the population of interest clear? [(1)]		Does the question require exploration of a relationship between variables or require deep exploration of the data?	

Note: The superscripts denote the categories were developed based on work from [(1)]Arnold (2013), [(2)]Konold et al. (1997), [(3)]Biehler (2001), [(4)]Frischemeier and Biehler (2018), and [(5)]Bakker and Gravemeijer (2004).

To measure whether the peer-feedback was effective with regard to the improvement of the statistical question, we identified - after the analysis of the questions in the first step - the aspects of the questions that showed the most improvement. We examined, in terms of the four question components (see Table 3), the most mentioned component (we identified the distribution of components in which feedback was provided) and then analyzed the peer-feedback that accompanied the questions. In this sense we first analyzed whether the statistical question improved because of the peer-feedback and then we identified the specific components where improvements were visible.

4. RESULTS

This study reports on the statistical problem posing skills of preservice primary teachers as they pose and refine statistical problems (i.e. investigative questions) that promote group comparisons of data. In doing so, we also examine the influence of peer and expert feedback on the evolution of statistical questions. The results are organised using the two research questions as a structuring framework.

4.1. WHAT ARE THE FEATURES OF THE STATISTICAL QUESTIONS DESIGNED BY PROSPECTIVE TEACHERS FOR USE WITHIN PRIMARY CLASSROOMS? IN WHAT WAYS DO THESE STATISTICAL QUESTIONS EVOLVE OVER THE COURSE OF THE STUDY?

In order to address this research question, the results are organised and discussed according to the categories established in the category system. A total of 79 initial and 158 revised questions were analysed. The proportion of questions classified into particular categories and subcategories, and across the three phases, is presented in Table 6.

Table 6. Categorisation of questions across the three phases

		Phase1 SQV1 (%)	Phase 2 SQV2 (%)	Phase 3 SQV3 (%)
Component 1: Look at the question				
Meaningful		100	100	100
Interesting		96	98	99
Clarity of Question		62	75	85
Variables	Non-statistical	9	8	3
	One variable	39	32	9
	Two variables	52	60	88
Clarity of population of Interest		62	74	90
Component 2: Variables in the question				
Clarity of Variables		42	46	89
Measurability of Variables		86	86	95
Component 3: Relationship between data and question				
Question Type	Survey	16	9	4
	Reasonable	68	71	63
	Sophisticated	16	20	33
Component 4: Look at the data				
Answer with given data			n/a	95
Sufficient data collected			n/a	95
Sufficient variability in the data			n/a	92
Local vs global view		59	69	88

Look at the question. The majority of questions posed were considered *meaningful* and of *interest* with almost all questions meeting the criteria in the first phase of the study (Table 6). In contrast, ensuring question *clarity* presented a greater challenge as only 62% of the initial questions [SQV1] were categorised as possessing clarity. For example, as seen in the first effort by Pair 6 to construct a statistical question, while they presented a relevant and interesting context, the question posed was non-statistical, in particular insufficient guidance was provided on whether to explore the distribution of data within each group or analyse data between the groups.

> It is the summer time and the weather is lovely to play outside. Children in 6th class complain they get too much homework. Can you help the student council ensure that children (in 2nd and 4th class) have enough free time in the evening after completing homework? [Pair 6, SQV1]

Note that improvements in clarity were evident across all subsequent phases indicating the benefit of both peer feedback (+13%) and expert feedback (+10%) in improving the clarity of questions posed (see Table 6). Examination of the number of *variables* identified in the questions revealed that by the end of the study 88% of questions involved two variables (Table 6). The assignment required the construction of a statistical question involving analysis of numerical data across two groups; hence, two variable questions were the most suitable type. The figure of 88% in Phase 3 represents a marked improvement to the 52% of two-variable questions constructed in Phase 1. As is evident from Table 6, this increase in two-variable questions at the end of the study was concomitant with a decrease in the less desirable non statistical questions (decreasing from 9% in Phase 1 to 3% in Phase 3) and one variable questions (decreasing from 39% in Phase 1 to 9% in Phase 3) across the study. Furthermore, an examination of changes that occurred across the three phases provides interesting insights into the relative efficacy of peer and expert feedback in influencing changes in the number of variables incorporated into questions. The large number of one-variable questions (39%) posed in Phase 1 remained relatively stable following peer feedback in Phase 2 (32%); similarly, there was a small increase in two variable questions from Phase 1 (52%) to Phase 2 (60%). This suggests that peer feedback had only a small effect in shifting a focus from the distribution of one numerical variable to examining the relationship between two variables. Expert feedback, in contrast, appeared more effective in precipitating the required shift in focus to two variable questions with a decrease in one variable questions (-23%) and an increase (+28%) in two variable questions from Phases 2 to Phase 3. Given the limited statistical understanding of peers as compared to the instructors, it is not surprising that peer feedback appeared less effective than expert feedback in this regard.

Only 62% of participants successfully identified a *population of interest* in the first phase of the study (Table 6). Examination of the pattern of responses suggests that the lack of question clarity evident in the initial stages of the study may also have contributed to the lack of success in identifying the population of interest. There was improvement in identifying a population of interest across subsequent phases indicating the effectiveness of both peer (+12%) and expert (+16%) feedback in shifting the focus from an individual to a sample or population.

Variables in the question. The task that presented greatest difficulty for participants in the initial stages of question design was *clarity in the variable*, ensuring that the variable was clear and the unit of measurement clearly defined (Table 6). Only 42% of questions met the criterion of clear and unambiguous in Phase 1. For example, examination of the question posed by Pair 11 reveals the lack of clarity in the variable 'sporting activity' in terms of what will be measured and the units of measurement. Following completion of the feedback phases, improvement in the clarity of variables was evident with 89% of questions being clear in terms of variables. The relatively weak influence of peer feedback is evident in the small increase from Phase 1 to Phase 2 (+4%). In stark contrast to this is the large (+43%) increase in questions categorised as possessing clarity in the variables following the provision of expert feedback. Examination of the revision made below by Pair 11 reveals the improvements in clarity in their question at the end of the study.

> Are there differences in the sporting activity of boys and girls? [Pair 11, SQV1]
>
> Are there differences in the *weekly* sporting activities *in hours* of boys and girls? [Pair 11, SQV3]

While comparison across the national settings is not a focus of this study, it is informative to note that those participants who were not required to produce a data set (as part of the course requirements) still struggled for clarity in variable description at the end of the study. One possibility is that the requirements to produce the sets of data may have focused participants' attention on the nature and features of the data such as unit of measurement, thereby forcing clarity in the variable description. Nonetheless, while advantages may have been afforded by constructing data sets, the construction of data sets posed a challenging task for some. Even in cases where the unit of measurement was identified and reasonably large data sets were provided with sufficient variability between values, in some cases, the actual values provided were not realistic. In one case (Pair 6, SQV3), the data referred to the number of hours of homework for two different groups of children. Data were presented in hours and the values were very high with several values indicating 7 or 8 hours of homework per night – unrealistic values in the real world. Similarly, another investigation [Pair 18, SQV3] examined the weight of schoolbags and data were presented in kilograms. Again, there were some very unrealistically high values with the majority of schoolbags weighing between 9 and 11 kgs.

From the initial phases, 86% of participants were successful in identifying variables available to *measure* (Table 6). This is one of the situations where there was no evidence of the effectiveness of peer feedback (+0%) and small differences arising from expert feedback (+9%), possibly due to a ceiling effect arising from the strong performances in identifying variables available to measure in Phase 1.

Relationship between the data and the question. When exploring the relationship between the *type* of question posed and the data generated by the question (Table 6), insights were gleaned into the ways in which the question opened (or closed) opportunities for a rich exploration of the data. Questions were broadly categorised into survey, reasonable (consisting of yes/no, non binary and specific value question) and sophisticated. While there are qualitative differences between all three, the latter types were desirable categories from our perspective. Overall, by the completion of the study, 96% of questions posed motivated the types of statistical reasoning considered desirable in school contexts.

The number of "simple" survey questions showed a decrease from 16% in Phase 1 and represented 4% of questions in Phase 3. Examples of a survey question is the probability-type question (Pair 2) and estimation question (Pair 35) below that were posed in Phase 1. Correspondingly the proportion of "sophisticated" questions showed a two-fold increase across phases from 16% in Phase 1 to 33% in Phase 3. Examples of questions categorised as sophisticated in Phase 3 are those of Pair 3 and Pair 1 below.

> What is the probability that a boy in your class has glasses? [Pair 2, SQV1]
>
> How many cream crackers would you eat with and without water in 1 minute? [Pair 35, SQV1]
>
> By how much do the shoes sizes of boys and girls in class 5 differ? [Pair 3, SQV3]
>
> Based on your examine of the sugar content of the fruit juices and minerals provided here, explain which drinks are more suitable for consumption. [Pair 1, SQV3]

The category of question identified as "reasonable" fluctuated across phases and was the only situation for which proportions of questions showed an increase and a decrease across different phases. Analysis reveals that the peer feedback brought about a modification of some questions categorised as 'survey' questions in Phase 1 to be recategorised as "reasonable" in Phase 2; thus indicating a small but appreciable positive influence of peer feedback. Of note then, is the impact of the expert feedback which followed and precipitated revision of many of those 'reasonable' questions (particularly those of the 'yes/no' type) to be categorised as "sophisticated" in Phase 3.

Look at the data. This component focuses on whether the statistical question posed can be adequately addressed by the type of data generated. Participants in only one of the settings (Ireland, n=118) were required to present the data sets. As these data sets were presented in Phase 3 alongside the final statistical question [SQV3], there were no opportunities to provide peer or expert feedback on the data sets that were constructed.

In Phase 3, 95% of the Irish participants were able to *answer* the question with the given data. Similarly, 95% collected *sufficient* data to answer the question. For the remaining 5%, the data sets

presented were small consisting of approximately 10 data values and did not present any sense of the distribution of the variable. An example is Pair 23 (below), who posed a statistical question but provided too few data values to support group comparison adequately.

> Fidget spinners are replacing the old style 'Spin Topper' - but in a test against time, who will be the last ones spinning? 5th class children worked in groups of 3 and observed and recorded (in seconds) how long their fidget spinners would spin. 6th class carry out the same activity for the spin tops.
>
> Fidget Spinner (in seconds): 72, 78, 86, 64, 93, 80, 52, 53, 58, 61
>
> Spinning Top (in seconds): 56, 65, 60, 69, 56, 78, 78, 73, 75, 81 [Pair 23, SQV3]

When the data sets were examined, the majority (92%) had sufficient structure or *variability* in the data. For the remaining 8%, some pairs provided data in intervals (e.g., Pair 7 who presented heart rates in intervals of 10 beats per minute) or rounded the data values (Pair 45 who rounded volume of water drank in half litres, see example below). These interval and rounded data values were problematic as they reduced the variability in the data and limited the insights that can be provided into the distribution of data. Also, the task was to construct an investigation for elementary grade children, and the presentation of interval-type data is not suitable for the skill sets of these children.

> Are we H_20 healthy 2 operate? Examine the data on the amount of water drank by a class of students in a normal day and on a sports day. Do our bodies really need more water when we are engaged in activity?
>
> Normal day (in mls): 500, 1000, 1000, 500, 1500, 1500, 1500, 1000, 500, 500, 1000, 1500, 1500, 2000, 2000, 1500, 0, 0, 500, 1000, 1500, 1000, 0, 500, 500
>
> Sports day (in mls): 1000, 1500, 1000, 1000, 1500, 1500,1000, 500, 1000, 500, 1500, 1500, 2000, 2000, 2500, 1500, 0, 500, 1000, 1500, 1500, 0, 500, 500 [Pair 45, SQV3]

Questions were examined to ascertain whether they would support and encourage analysis not only for an individual but for a whole group (Table 6). There was a steady increase in questions that supported a focus on the whole group across all phases indicating the benefit of both peer (+10%) and expert (+19%) feedback in supporting a global view of the data.

In summary, by the end of the course our students were relatively successful in writing questions that addressed component 1, with the biggest gains being the ability to write about two variables, followed by the clarity of the population of interest (interesting since this was not on the SPP) and of the question. For component 2, the variables were clear (in the sense that the units for example were defined in a clear way) in less than half of the original questions, but the variables were measurable in 86% of the questions. Summarizing the main outcomes of the analysis of component 3, while students initially demonstrated difficulty generating sophisticated questions, they developed greater skills across the phases of the study. Participants performed very well with regard to all aspects of component 4 - "answer with given data", "sufficient data collected", "sufficient variability in the data" and the development from the local towards a global view on data.

4.2. TO WHAT EXTENT DOES PEER FEEDBACK SUPPORT THE DEVELOPMENT OF STATISTICAL QUESTIONS?

In order to answer question 2, peer feedback from a sample of 40 SPPs (50%) was analysed. A selection of initial questions (SQV1), the associated feedback and subsequent revised questions (SQV2) are presented in Table 7. When examining the focus of the peer feedback and how it was distributed across the four question components (Table 3), we noted that peer feedback emphasised:

- Look at the question (40%)
- Variables in the question (27%)
- Relationship between data and question (25%)
- Look at the data (8%)

The most mentioned component was 'Look at the question,' which constituted 40% of the feedback. A large proportion of this feedback praised the question context and commented favourably on the levels of interest and curiosity invoked by the question design. Some feedback advised changes to the

question wording to ensure the answer to the question would not be obvious (see Pair 41, Table 7). Feedback focusing on 'Variables in the question' represented 27% of mentions. Feedback in this category focused predominantly on the need to specify a unit of measurement (Pairs 6, 7, 15), to describe the variable more clearly (Pair 8), and requested clarification about how data would be collected (Pair 15). Examination of the 'Relationship between data and question' was mentioned in 25% of the feedback. Approximately half of this feedback provided praise for correctly identifying two groups of numerical data (Pairs 14, 29) while the remainder focused on the need to identify two comparable groups (Pairs 7, 8) or commented on the question limiting the degree of investigation (Pair 13). The component 'Look at (or imagine) the data' represented 8% of the feedback and focused mainly on the need to specify the amount of data to be collected (Pairs 12, 15).

Our analysis of the peer feedback provided on the initial questions (SQV1), and the subsequent revisions made to questions (SQV2) (see Table 7), generated two observations.

Observation 1: The focus and quality of feedback varied greatly. Some feedback focused on praising noteworthy aspects of questions. Due to the absence of targeted feedback for improvement, in these instances there was little appreciable improvement made to the revised statistical questions. For example, the feedback provided to Pair 10 (Table 7) consisted only of praise and provided no guidance for revisions. There was, however, a revision made to the specification of sample size. This revision, which was not suggested by their peer feedback group, may be attributed to learning for Pair 10 arising from acting in the role of providing feedback to another pair. More prevalent, however, was the use of mitigating language, which highlighted strengths of questions while also providing constructive feedback on question components in need of attention. Again, this feedback aimed at improving questions varied in its' focus and quality. Some feedback referred to aspects of the question that were not part of the four-component framework and were not motivated by the question prompts that were provided. In such instances, peers referred to the impact of involvement in the imagined PPDAC cycle on the study participants (Pairs 14, 29). In the remaining cases, the feedback was very focused and effectively utilised the framework and prompt questions. For example, the feedback provided to Pair 15 identified positive question aspects and then focused on sample size and the units of measurement. In the case of Pair 13, the feedback addresses the lack of an investigative aspect to the question and encourages development of an investigative component.

Observation 2: The degree of implementation of peer feedback varied greatly. The effectiveness of peer feedback is difficult to establish partly due to a lack of an obvious relationship between the quality of feedback and the degree of implementation. For example, the feedback provided to Pair 15 had the potential, if carefully implemented, to bring substantial improvements to the question. As evidenced in Table 7, however, the participants constructed an entirely new statistical question rather than modifying the initial question. On some occasions when this happened, the new questions represented an improvement (Pair 8, Table 7), and on other occasions a disimprovement (Pair 41, Table 7), on the initial question. Regardless, these situations prevented us from ascertaining the effectiveness of, what we considered, good quality peer feedback on the refinement of statistical questions. Analysis of the questions that were revised indicates again that the degree of implementation varied. Some pairs who received appropriate feedback did not implement any of the suggestions (see Pair 14) or made very minor changes that appeared of little appreciable value (Pair 12). Others, however, implemented the feedback leading to appreciable improvement in the statistical questions (Pairs 6, 7).

Examination of Table 6 suggests a lack of effectiveness of the peer-feedback, as opposed to expert feedback, in improving some aspects of question design. Our analysis identifies possible reasons for this. Firstly, many of the German participants constructed very good questions in Phase 1; hence, for these participants, there may have been a ceiling effect and little room for improvement arising from peer feedback. Second, analysis of the types of feedback provided by peers between Phases 1 and 2 indicates that the Irish participants struggled to provide constructive feedback to their peers. Rather, there was the tendency to identify the strengths of questions and provide praise for specific question aspects such as the use of interesting contexts. Fewer pairs identified weaknesses in questions or provided suggestions for improvements.

Table 7. Examples of Statistical questions, peer-feedback and revised questions

Pair	Initial Statistical Question (SQV1)		Peer feedback provided		Revised Statistical Question (SQV2)
6	Do 6th class pupils spend more time on Instagram or snapchat over the course of the week?	→	The question is interesting for most students. But, it excludes children who don't have social media. Some suggestions are that you could think of an alternative to social media. Or, carry out on a different age group (e.g., secondary school students) so that everyone will have a phone. Identify a specific unit of time - are you measuring in minutes or hours?	→	Over the course of the weekend (Friday to Sunday), do 3rd year students (15-year-olds) spend more time on Instagram or Snapchat?
7	What wingspan is the best for making paper airplanes?	→	There are not two sets of data. You need to be more specific about the length of wingspan.	→	Does a paper airplane fly better with a wingspan above 10cm or below 10cm?
8	How many children get nine or more hours of sleep per night?	→	There aren't two sets of data here – it's just all children! We don't know what the variable is – it might be children? Or it might be sleep? We don't know how many of them we should look at. And it is worded badly. Maybe see the difference in groups of children who get more than nine hours of sleep and those who get less.	→	In the Olympic games over the past 100 years, is the height reached by high jumpers more than the distance jumped by long jumpers?
10	What are the age differences between Olympic gymnasts and 100m sprinters?	→	The question is associated with sport so it is interesting and lots of people can relate to it. Athletes can peak at different times in their careers.	→	What are the age differences between male and female gymnasts and 100m sprinters picked from 25 countries in the Olympics?
12	Does reading before bed or using technology before bed impact the number of hours slept? Measure hours slept per night using fitbits.	→	Really interesting question. But, people already know that using technology leads to disrupted sleep. Maybe the question should read 'Does reading before bed or using technology before bed lead to less or more sleep? Which has more of an impact?' How many people will you study?	→	Which has a greater impact on the number of hours children slept using technology before bed or reading before bed?
13	How many All-Ireland football titles have Munster teams won compared to Leinster teams?	→	This is good because everyone would be interested in it. But, it is a one-dimensional question. You just have to go online or to a sport magazine and look it up and you'll find out. So there isn't much investigating. Maybe you should investigate 'why' something is happening in sports?	→	There has been much debate about the comparison of skills between the county football teams in Leinster and Munster. Can you figure out the difference between the scores in football finals in the last 10 years?
14	How much time do you spend on your homework? Compare with now (Mondays in October) and later in the year (Mondays in another month).	→	These are two comparable groups. The data are numerical. The answers should remain anonymous so that the children don't feel uncomfortable sharing. What unit will you use to measure? But, there is too	→	Students complain that in October they receive too much homework. Students feel they get a more manageable amount of homework in April. Compare how much time spent on homework in

			long a time period to wait for the data collection. Suggestion: who gets more homework 1st class or 5th class?		April to homework in October. Evaluate the results.
15	What is the difference in sugar content between banana bread and chocolate cake?	→	This is a good question because there are two different groups – banana bread and chocolate cake. It is also interesting. There need to be about 25 different values for each of the groups. You need to name a unit of measurement like spoons of sugar or weight. Where would you obtain the data?	→	Disney has designed a new roller coaster X metres high and Y metres long. Where should Disney put the roller coaster and why? Find the heights and lengths of 25 roller coasters in Disney-Paris and 25 in Orlando. Based on these data, decide where the new roller coaster should be located.
29	It is active schools week. There is a competition for push ups in a minute between 4th and 6th classes.	→	There are two groups and the number of pushs up is numerical. But, there is no question asked. Unfit children might feel selfconscious about doing this task. The 6th class children are more likely to win cause they are stronger – so you need to make it apparent who will win.	→	
41	Do you read more books or watch more movies? How many books have you read in the last month versus how many movies have you watched in the last month?	→	There is no need to collect data cause the answer is obvious – it takes much longer to read a book than watch a movie. It also isn't a good context for children cause it won't intrigue or engage them. It isn't clear how many people are in the study- is it just one? It should be more.	→	Who scored more points per game in the last 3 rugby world cups – Ireland or England?

Efforts on the part of the instructor to emphasize that the purpose of the feedback was to improve the statistical question, and that improvements can only occur by helping pairs identify the shortcomings of their statistical question, did bring about noticeable improvements in the quality of feedback. This may suggest that participants were not comfortable providing constructive criticism, thus, explaining their tendency to praise desirable question features rather than suggest areas for improvement. The third point is that 36% of Irish participants entirely changed (rather than revised) their questions between Phase 1 and Phase 2; this limited our ability to see the benefit of feedback because the actual question changed. There is also the possibility that useful feedback on the statistical problem was provided but the pairs could not implement a solution to the identified problem. There is research to indicate that feedback is more likely to be implemented if a solution was provided (Nelson & Schunn, 2009); our analysis reveals that the majority of feedback did not identify solutions and this may have been a contributing factor to the lack of efficacy of peer feedback.

5. DISCUSSION

The challenges experienced by prospective teachers when posing statistical questions in the first phase of this study are similar to those described by Frischemeier and Biehler (2018). Some participants struggled with these initial questions and, similar to the findings from Pfannkuch and Horring (2004), they grappled at first with the meaning of the word question within a statistical context (i.e. how would you ask a question that would drive the collection of data and the statistical inquiry process?). As evidenced by the improvement in quality of statistical questions over time, however, our study reveals that when provided the appropriate structured support, prospective teachers can develop the skills and understandings to develop rich statistical questions suitable for use in primary school contexts. Appropriate structured support took several forms. Working in pairs provided opportunities for ongoing discussion and revision of questions. Peer feedback provided commentary in the form of a 'fresh pair of eyes', structured according to a four-component framework (Table 3), targeted towards each pairs' own specific question. Pairs, having had the experience of reviewing the question of another pair, then had the opportunity to revise and refine questions in response to the feedback provided by their peers. Finally, expert feedback was structured and delivered according to the same four-component organising framework as the peer feedback. While expert feedback was given verbally in a whole class setting, experts provided examples that illustrated question weaknesses and allowed participants to evaluate their own research questions in light of this feedback.

It is evident from the study that the process of developing good statistical questions is particularly complex and takes considerable time and support. Prospective teachers have to consider multiple features when designing questions. These relate to the *question* and address the meaningfulness, interest, and clarity of questions in addition to the variables and populations elicited by the questions. They also involve consideration of the *variable* and require attention to the clarity of the variable described and the measurement of those variables. Furthermore, prospective teachers must attend to the *relationship* between the question and the type of data generated by the question in an effort to ensure the data support subsequent analysis. Finally, the *data themselves* require consideration when designing a statistical question. Attention must be paid to the potential to answer the question from the amount and type of data generated, whether there is sufficient variability in the data and whether the question supports a global view of the data.

Some of these categories posed lesser or greater difficulty than others from the outset of the study. As early as Phase 1, participants showed great expertise in addressing the meaningfulness and interest of questions and in identifying the variables available to measure. In contrast, all participants struggled in Phase 1 to provide clarity when describing the variables, incorporate two variables for consideration in the question, and support a global view of the data. For those participants who struggled, it is apparent that their efforts in the early stages of the study focused largely on the descriptions, rationale and design of contexts that underpinned the questions. It appears that they invested their energies in the broader contextual dimensions and in framing questions divorced from attention to the necessary statistical and content dimensions of the questions posed. Future studies need to examine how to marry the focus on context alongside the statistical content dimensions from the early phases of question design.

The benefits of the peer-feedback process are not clear from this study. As found in other studies, peer review brought about improvements (Allmond & Makar, 2010; Frischemeier & Biehler, 2018)

between Phase 1 and Phase 2 on some aspects of question design. It appears to have supported improvements in clarifying the question, identifying the population of interest and supporting a global view on the data. For other question components, however, peer-feedback did not support any substantial improvement in question quality particularly when selecting and identifying the number of variables embedded in the question and in bringing about improvements in the question type. The nature of feedback provided by participants was similar to that found in other research. Specific feedback which identified the source of the problems alongside feedback that provided suggestions or solutions for improvement was more likely to be implemented and lead to improvements in the statistical questions thus supporting the research from Matsunara et al. (2002) and Nelson and Schunn (2009).

Several recommendations for improvements to practice around peer feedback arise from this study. Firstly, some participants appeared reluctant to provide critical feedback to their peers. Consequently, from an affective and relationship perspective, greater effort needs to be made to provide clarity around the purposes of providing peer feedback. A second recommendation relating to this point is to provide examples of desirable feedback; this may further assist prospective teachers in identifying the features of good feedback. This may also lead to the incorporation of effective feedback identified in the literature which did not appear in the feedback provided in this study, notably the provision of summaries of work (Ferris, 1997) and clarification of the purpose of the feedback (Bitchener et al. 2005). Thirdly, the provision of a scoring rubric may be useful in directing attention to the quality of question components and thus support the peer feedback process. The use of a rubric may also allow the identification of strengths of the questions alongside isolation of problematic features in need of revision. Finally, as research identifies a direct relationship between implementation of feedback when a solution to the problem is provided (Nelson & Schunn, 2009), it is recommended that the structured feedback provided by peers should be accompanied by suggested solutions to that problem.

There are a number of suggestions for further research arising from this study. Firstly, there is an opportunity to investigate the efficacy of a rubric for prospective teachers to self-assess the quality of statistical questions. Secondly, the efficacy of peer feedback was difficult to determine due to the unanticipated decision of many participants to write an entirely new question in response to peer feedback. Future studies might either incorporate interviews with pairs to provide insights into decision making in these situations and/or include a stipulation preventing the construction of new statistical questions in Phase 2.

REFERENCES

Allmond, S., & Makar, K. (2010). Developing primary students' ability to pose questions in statistical investigations. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. *Proceedings of the 8th International Conference on Teaching Statistics* (ICOTS8), Ljubljana, Slovenia, July 11–16. International Statistical Institute. http://iase-web.org/documents/papers/icots8/ICOTS8_8A1_ALLMOND.pdf

Arnold, P. M. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* [Doctoral dissertation, University of Auckland]. https://researchspace.auckland.ac.nz/handle/2292/21305

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools* [Doctoral dissertation, University of Utrecht]. https://dspace.library.uu.nl/bitstream/handle/1874/893/full.pdf?sequence=2

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2278-6

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63. https://iase-web.org/documents/SERJ/SERJ3(2)_BenZvi.pdf?1402525004

Biehler, R. (2001). Statistische kompetenz von schülerinnen und schülern: Konzepte und ergebnisse empirischer studien am beispiel des vergleichens empirischer verteilungen. In M. Borovcnik, J. Engel, & D. Wickmann (Eds.), *Anregungen Zum Stochastikunterricht* (pp. 97–114). Franz Becker.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14,* 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics: Challenges for teaching and teacher education.* A joint ICMI/IASE study (pp. 57–69). Springer. https://doi.org/10.1007/978-94-007-1131-0

Cai, J., Moyer, J. C., Wang, N., Hwang, S., Nie, B., & Garber, T. (2013). Mathematical problem posing as a measure of curricular effect on students' learning. *Educational Studies in Mathematics, 83,* 57–69. https://doi.org/10.1007/s10649-012-9429-3

Cai, J., & Hwang., S. (2020). Learning to teach through mathematical problem posing: Theoretical considerations, methodology, and directions for future research. *International Journal of Educational Research, 102*. https://doi.org/10.1016/j.ijer.2019.01.001

Chapman, O. (2004). Helping pre-service elementary teachers develop flexibility in using word problems in their teaching. In D. McDougall & A. Ross (Eds.) *Proceedings of the 26th North American Chapter of PME, 3*, 1175–1182. http://www.pmena.org/pmenaproceedings/PMENA%2026%202004%20Proceedings%20Vol%203.pdf

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–43. https://doi.org/10.1007/s11251-010-9146-1

Cho, K., & MacArthur. C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–38. https://doi.org/10.1016/j.learninstruc.2009.08.006

Crespo, S. (2003). Learning to pose mathematical problems: Exploring changes in preservice teachers' practices. *Educational Studies in Mathematics, 52*(3), 243–270. http://dx.doi.org/10.1023/A:1024364304664

Crespo, S., & Sinclair, N. (2008). What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. *Journal of Mathematics Teacher Education, 11*(5), 395–415. http://dx.doi.org/10.1007/s10857-008-9081-0

Crespo, S., & Harper, F. K. (2020). Learning to pose collaborative mathematics problems with secondary prospective teachers. *International Journal of Educational Research, 102*. https://doi.org/10.1016/j.ijer.2019.05.003

Ellerton, N. F. (2013). Engaging pre-service middle-school teacher-education students in mathematical problem posing: Development of an active learning framework. *Educational Studies in Mathematics, 83,* 87–101. http://www.jstor.org/stable/23434198

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. https://doi.org/10.2307/1170785

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31*(2), 315–339. https://doi.org/10.2307/3588049

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

Frischemeier, D. (2019). Primary school students´ reasoning when comparing groups using modal clumps, medians and hatplots. *Mathematics Education Research Journal, 31*(4), 485–505. https://doi.org/10.1007/s13394-019-00261-6

Frischemeier, D., & Biehler, R. (2018). Stepwise development of statistical literacy and thinking in a statistics course for elementary preservice teachers. In: T. Dooley, & G. Gueudet (Eds.), *Proceedings of the 10th Congress of the European Society for Research in Mathematics Education* (pp. 756–763). DCU Institute of Education and ERME. https://hal.archives-ouvertes.fr/hal-01927856

Frischemeier, D., & Leavy, A. (2020). Improving the quality of statistical questions posed for group comparison situations. *Teaching Statistics, 42*(2), 58–65. https://doi.org/10.1111/test.12222

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://psycnet.apa.org/doi/10.1037/0033-2909.119.2.254

Konold, C., & Pollatsek, A. (2002). *Data analysis as the search for signals in noisy processes. Journal for Research in Mathematics Education, 33*(4), 259–289. https://doi.org/10.2307/749741

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference* (pp. 151–167). International Statistical Institute. https://iase-web.org/documents/papers/rt1996/13.Konold.pdf?1402524984

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In *Proceedings of the Sixth International Conference on Teaching Statistics* (ICOTS6), Cape Town, South Africa. International Statistical Institute. https://iase-web.org/documents/papers/icots6/8b2_kono.pdf?1402524963

Kuckartz, U. (2012). *Qualitative inhaltsanalyse. Methoden, praxis, computerunterstüzung*. Beltz Juventa.

Leavy, A. M., & Hourigan, M. (2019). Posing mathematically worthwhile problems: Developing the problem posing skills of prospective teachers. *Journal of Mathematics Teacher Education, 23,* 341–361. https://doi.org/10.1007/s10857-018-09425-w

Leavy, A. M., & Hourigan, M. (2021). Enhancing the mathematical problem posing skills of prospective teachers through a mathematical letter writing initiative. *Journal of Mathematics Teacher Education.* https://doi.org/10.1007/s10857-021-09490-8

Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*(1), 3–25. https://doi.org/10.1086/499713

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365–380). Springer. https://doi.org/10.1080/10986065.2016.1151294

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

National Council of Teachers of Mathematics. (2000). *The principles and standards for school mathematics* [PSSM]. Author.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37* (4), 375–401. https://psycnet.apa.org/doi/10.1007/s11251-008-9053-x

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51*(1), 34–38. https://doi.org/10.1080/87567550309596408

Pfannkuch, M., & Horring, J. (2004). 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 Statistics Education 2004 Roundtable* (pp. 204–218). International Statistical Institute. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.4633&rep=rep1&type=pdf

Silver, E. A. (2013). Problem-posing research in mathematics education: looking back, looking around, and looking ahead. *Educational Studies in Mathematics, 83,* 157–162. http://www.jstor.org/stable/23434203

Sugita, Y. (2006). The impact of teachers' comment types on students' revision. *ELT Journal, 60*(1), 34–41. http://dx.doi.org/10.1093/elt/cci079

Topping, K. J. (2009). Peer assessment. *Theory into Practice, 48*(1), 20–7. https://doi.org/10.1080/00405840802577569

Watson, J. M., & English, L. D. (2017). Statistical problem posing, problem refining, and further reflection in grade 6. *Canadian Journal of Science, Mathematics and Technology Education, 17*(4), 347–365. https://doi.org/10.1080/14926156.2017.1380867

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168. https://www.jstor.org/stable/3483313

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

AISLING LEAVY
Mary Immaculate College
South Circular Road
Limerick, Ireland

Appendix A: Categories in the component *Look at the question* with their values, definitions and examples

Category	Sub categories	Possible values	Definition	Examples from the data
Question_meaningful		Yes (1 Pts)	A question is meaningful if it is connected to a context.	Do female students spend more time doing homework than male students?
		No (0 Pts)	A question is not meaningful if it does not have a context and is set in abstract terms.	Compare the means of the two sets of data.
Question_interest		Yes (1 Pts)	A question will sustain interest and curiosity of primary children, if a topic in the frame of the experiences of their environment is chosen (e.g., pets, free time activities)	Do children with pets spend fewer time with their friends than children without pets?
		No (0 Pts)	A question will not sustain interest and curiosity of primary children, if a topic is chosen which does not fit the interests and experiences of primary school children.	Do male employees have a larger hourly income than female employees?
Question_clear		Yes (1 Pts)	A question is clear and unambiguous, if it can be identified as a summary question, comparison question or relationship question.	Are backpacks of the 4th graders heavier than backpacks of the 3rd graders? (relationship question)
		No (0 Pts)	A question is not clear and unambiguous, if it cannot be identified as a summary question, comparison question or relationship question.	What is the probability that a boy has glasses in your class?
Question_variables	Cat	Distribution of one categorical variable (1 Pts)	Question aimed at the distribution of a categorical variable.	What is the distribution of male and female students in the class?
	Num	Distribution of one numerical variable (1 Pts)	Question aimed at the distribution of a numerical variable.	What is the distribution of shoe size in this class?
	Cat x Cat	Relationship between two categorical variables (2 Pts)	Question aimed at relationship between two categorical variables.	Do male students more likely to have a tablet rather than female students?
	Num x Cat	Relationship between a categorical and a numerical variable (2 Pts) → group comparison question	Question aimed at relationship between a numerical and a categorical variable.	How do male and female students differ in regard to their height?
	Num x Num	Relationship between two numerical variables (2 Pts)	Question aimed at relationship between two numerical variables.	Are students who spend much time in reading (in hours per day) also spending much time in homework (in hours per day)?
Pop_interest		Focus_individual (0 Pts)	"A boy" indicates an individual	"a boy"
		Focus_sample (1 Pts)	"the boys" suggests the boys	"the boys"
		Focus_population (1 Pts)	"boys" most likely indicates that it is about the population.	"boys"

Appendix B: Categories in the component *Variable in the question* with their values, definitions and examples

Category	Possible values	Definition	Examples from the data
Variables_clear	Yes (1 Pts)	The variable(s) is/are described clearly.	Time_computer (in hours per week)
	No (0 Pts)	The variable(s) is/are only mentioned, but the unit of measurement is not clearly defined.	Time_computer (without referring to a unit)
Variables_measure	Yes (1 Pts)	The variable(s) is/are available/possible to measure.	Do 3rd graders spend more time on sporting activities then the 4th graders? (when working on a data set on freetime activities of the 3rd and 4th graders, including the variable Time_for_sporting_activities)
	No (0 Pts)	The variable(s) is/are not available/possible to measure.	Is there a relationship between Intelligence and eye color? [when working on a data set on freetime activities of 3rd and 4th graders which does not include the variable intelligence]

Appendix C: Categories in the component *Relationship between data and question* with their values, definitions and examples

Category	Possible values	Definition	Examples from the data
Question_Type	Survey question (0.5 Pts)	Student poses a survey question.	What eye color do you have?
	Yes_No Question (1 Pts)	Student poses a question which answer will be yes/no.	Are male students spending more time on the computer than female students?
			Which team is more accurate (3rd class or 4th class)?
	"Non binary" Question (1Pts)	Student poses a question aimed at investigating whether a Group A is better/larger/smaller, higher/…than Group B	Which team has a better score?
	Specific value question (1 Pts)	Student poses a question focusing on a specific value (like a relative frequency, mean, median, etc.).	What is the relative frequency of tablet users in the sample?
	Sophisticated questions (2 Pts)	Student poses a question which aims at exploring a relationship between variables and an answer exceeding Yes/no	How much more pocket money do 4th graders get than 3rd graders?
		Students poses a question which allows a deeper exploration of the data set	In which way do 3-graders and 4-graders differ in their freetime activities?

Appendix D: Categories in the component *Look at the data* with its values, definitions and examples

Category	Possible values	Definition	Examples from the data
Answer_question_with_ _given_data	Yes (1 Pts)	The question can be answered with the given data.	Do 3rd graders tend to spend more time on sporting activities then the 4th graders? [when working on a data set on freetime activities of the 3rd and 4th graders, including the variable] Time_for_sporting_activities
	No (0 Pts)	The question cannot be answered with the given data.	Is there a relationship between Intelligence and eye color? [when working on a data set on freetime activities of 3rd and 4th graders]
Sufficient_Data_collected	Yes (1 Pts)	The question aims at an investigation in the sample and identifies a sample size sufficiently large so as to facilitate comparison (n>25).	Is there a difference in weights of 6th grade school bags and 5th grade school bags in Forest elementary school?
	No (0 Pts)	The question aims at an investigation which goes beyond the sample or generates a sample too small to facilitate comparison.	Is there a difference in the weight of all boys and girls' school bags?
Sufficient_Variability_in_data	Yes (1 Pts)	A wide range of possible data values is given.	Who is most successful in the bottle flip competition in two minutes – students or staff? [data values raged from 0 to 18]
	No (0 Pts)	No wide range of possible data values given.	No example in the data
Local_vs_global_view	Local view (0 Pts)	The question is referring to an individual case (armspan of a boy).	What is my armspan?
	Global view (1 Pts)	The question is referring to the aggregate picture rather than an individual case (distribution of armspan of a whole class).	How wide would an entrance have to be to allow all elementary school students pass through with their arms spread out? [Distribution of armspan of a whole school]