Investigating Computer Aided Assessment of Mathematical Proof by Varying the Format of Students' Answers and the Structure of Assessment Design by STACK

Maryam Khalid H Alarfaj

Doctor of Philosophy

University of Edinburgh

2023

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

Maryam Alarfaj

Acknowledgement

First and foremost I would like to thank my supervisor, Prof. Christopher Sangwin, for the considerable amount of work that he has put in in assisting me with both the work which comprises this thesis, and the thesis itself. I will be forever grateful for your guidance and support in producing this thesis. I would also like to thank my second supervisor Dr Toby Bailey for all his support. I would like to thank the entire Mathematics teaching theme at The University of Edinburgh. It has been a privilege to be surrounded by such a welcoming, passionate and diverse group of researchers.

Several people in particular need to be acknowledged for always being there when I needed them. I would like to thank Salma Alarfaj for all her time and help. Words cannot express my gratitude and appreciation to Dr. Maarya Sharif for all her support and advise.

I have left the most important people in my life until last and they are of course my family. To my husband Mohammed, and my daughters Sarah and Lulu who have been a constant source of support and encouragement during the challenges of graduate school and life.

Finally , this work is dedicated to my parents who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

Abstract

Students are increasingly being expected to use Computer Aided Assessment (CAA) systems as support for traditional courses. Assessing a full mathematical proof in an educational context and providing feedback and other outcomes to students is currently well beyond the capabilities of CAA systems. One possible approach to assessing students' answers has been to break up larger tasks into smaller individual steps to which automatic assessment can then be applied. However, the method of marking depends on the format of a mathematical response. This thesis aims to investigate the effectiveness of different formats in computer aided assessment of mathematics. A format effect occurs when the format of an exercise affects the rate of successful outcomes of the exercise.

Having established the need to study the format effect when writing mathematical arguments particularly online, I consider three formats for writing open-ended questions: two-column, typing, and Separated Concerns. The first of three studies explored the impact of the two-column format in writing simple mathematical arguments. In conducting this research, I developed a coding scheme to describe and analyse the structure of individual mathematical arguments. The second study focused on the difference and the format effects between uploading handwritten and typing in writing mathematical responses. Another outcome of this study is to provide a further application of using the coding scheme on analysing students' arguments. The third study focused on updating STACK potential response tree based on Separated Concerns. STACK is a System for Teaching and Assessment using a Computer algebra Kernel, is an open source computer aided assessment system for mathematics, and other STEM subjects. Separated Concerns is a phrase used to describe materials in which potential misconceptions are addressed directly. In this study, I focused exclusively on students' responses, and misconceptions in learning proof by induction using STACK. Mathematical induction is used as a vehicle to illustrate the idea of Separated Concerns. The main goal of the third study is to understand how engagement with learning materials packaged into online quizzes to replace live lectures, a "lecture quiz", related to success on the weekly assessed quiz, and the course total. A second goal of the third study is to explore the common mistakes made by students when using online materials to prepare for mathematical induction. This study also illustrates how to use research to update the algorithms which assess students' answers, known as "STACK potential response trees", in questions written to support learning mathematical induction based on Separated Concerns.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The primary aim of this thesis is to investigate the effectiveness of different formats in computer aided assessment of mathematics. A *format effect* occurs when the format of an exercise affects the rate of successful outcomes of the exercise. Previous work, such as Sangwin and Jones (2017), found a format effect between multiple choice and constructed response questions when students work with reversible mathematical processes, and so it is reasonable to expect a format effect might exist between the traditional and other formats when writing complete mathematical arguments.

Having established the need to study the format effect when writing mathematical arguments particularly online, I consider three formats for writing open-ended questions: two-column, typing, and Separated Concerns. The underlying goal of the research reported in this research is to inform understanding of how to effectively assess students' proof construction, particularly online.

1.1 Outline of thesis

In Chapter 2, the literature review can be broadly split into three main categories; proofs, cognitive load theory and formats used for assessing open-ended question for assessing mathematics online. First, I survey the literature on proof, beginning with a brief history of proof and its role in mathematics. I look at the different ways one could present a proof text. Second, I review the literature on cognitive load theory and discuss expertise reversal effect when using instructional design. Third, I discuss three different formats for writing open-ended questions: two-column, typing, and faded worked examples. I also provide examples of using these formats in the contemporary online assessment. The final part of the literature review covers the computer aided assessment system (STACK).

Chapter 3 discusses the literature on logical reasoning and argumentation, based mainly on the works of Toulmin (1958) and Aberdein (2005). I then introduce a "unit coding scheme" and explain how the work of Toulmin (1958), Hodds (2014) and R. J. Back, Mannila, and Wallin (2010) were integrated to create the scheme. Toulmin's argumentation scheme provided a model for only one simple argument, therefore R. J. Back, Mannila, and Wallin (2010) has also been used to model structured derivations for mathematical proofs. The coding scheme of Hodds (2014) was adapted to code the written justification generated when writing mathematical arguments. I then defined what I mean by a "proof unit". I provided an example that illustrates how to apply the unit coding scheme. The unit coding scheme has been applied to code students' responses for the first two studies in this thesis. The chapter presents the unit coding scheme as a significant methodological contribution, offering an approach to assessing written work in online environments by evaluating both the quantity and quality of student explanations and proof construction. This innovative tool has the potential to revolutionize online assessment practices by providing deeper insights into student understanding and fostering more effective learning experiences. Further research could explore the scheme's applicability to additional subjects and its effectiveness in promoting deeper learning strategies.

Following the coding scheme, the three studies are reported. Each study has: methods, where the methods specific to each study addressed, results, where the statistical results are reported, and discussion.

Chapter 4 presents the first of three studies focused on a potential format effect when writing mathematics. The study aims to explore the impact of the two-column format in writing simple mathematical arguments. In this study, I focus exclusively on the relationship between formats and types of justifications given by students. I undertook an experiment to compare students' responses between traditional arguments and arguments in a two-column format. Indeed, this experiment seeks to consciously constrain the format. A secondary outcome of this study was the use and further development of the coding scheme which I provided in Chapter 3.

Chapter 5 presents the second study focused on the difference and the format effects of typing vs handwriting mathematical responses. The results of the first study into the format effects when writing mathematics arguments are a promising indicator. Students are increasingly moving away from paper submission of assignments to working online, a trend accelerated in 2020/21 by the global pandemic. Online submission includes both automated online assessment and online submission of written work for human making. A natural question is therefore: is there a difference in performance and justifications between uploading handwritten and typing in writing mathematical responses? A secondary outcome of the second study was the use and further development of the coding scheme which I provided in Section 3.2. I conducted an experiment in which participants responded to an online task containing equivalent typing and uploading handwritten items, and students' reactions immediately after the task were obtained. Factors explored included number of steps or "units", and the overall score awarded.

Chapter 6 starts by discussing some difficulties students have when learning mathematical induction identified by previous educational research. Four concerns in learning mathematical induction were identified. I introduced the phrase "Separated Concern" which I used to describe materials in which potential misconceptions are addressed directly.

Chapter 7 presents the third of three empirical studies focused on updating STACK potential response tree based on Separated Concerns. In this study, I focus exclusively on students' responses, and misconceptions in learning induction proofs. Mathematical induction is used as a vehicle to illustrate the idea of Separated Concerns. The main goal of this study is to firstly illustrates how engagement with the lecture quiz related to success on the weekly assessed quiz, and the course total. Second, to explore the common mistakes made by students when using online materials (i.e., STACK questions) to prepare for mathematical induction? Secondary to the above research goals, this chapter illustrates how to use research to update STACK potential response trees in questions written to support learning mathematical induction based on Separated Concerns. Assessing a full mathematical proof is currently well beyond the capabilities of computer systems, but one possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. One of the contributions of this study is to improve our general understanding of how to design and use STACK potential response trees.

The thesis concludes by providing a general summary of the research conducted with some of the main limitations of the studies. Some possible directions for future work are suggested.

Chapter 2

Literature Review

In this chapter, the literature review can be broadly split into three main categories; proofs, cognitive load theory and formats used for assessing open-ended question in CAA of mathematics. First, I survey the literature on proof, beginning with a brief history of proof and its role in mathematics. I look at the different ways one could present a proof text. Second, I review the literature on cognitive load theory and discuss expertise reversal effect when using instructional design. Third, I discuss three different formats for writing open-ended questions: two-column, typing, and faded worked examples. I also provide examples of using these formats in the contemporary online assessment.

The first format is two-column which is discusses in Section 2.3.1. The constraints in the interface might help students in introductory classes to understand the basic elements of mathematical proof. This two-column format requires students to explicitly express conditions for each step and combine them in order to connect assumptions with conclusions. However, at a technical level typing proofs provides a way to capture students' expression but typing is a proof with meaning. Therefore, human marking is needed for typed responses in which students have to show their work in reasoning by expressing their work in longer argumentation and derivations. For that reason, I turn my attention to discuss the mechanism of typing mathematical response and compare the format effect of typing versus uploading photographed mathematical responses. The study is presented in Section 2.3.1. Since COVID-19, online learning has been the primary method to keep learning going, requiring substantial support. Students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). One possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. I was influenced by faded worked examples to introduce and investigate a new way of designing STACK question which is termed "separated concerns". The idea of separated concerns is discussed in Section 6.4.

The final part of the literature review covers the computer aided assessment system (STACK).

2.1 Proofs

Proof fulfills a number of purposes, in the mathematics community and mathematics classrooms. The literature provides many definitions to what exactly is mathematical proof? For example, Rav (2007) suggested that "... a mathematical proof is perhaps just a sequence of logical steps, following some implicit rules, to be judges of the legitimacy of these steps"(p.239). A. Selden and J. Selden (2003) stated that proofs are "... arguments that prove theorems"(p.4) whilst Pelc (2008) simply stated that "by proofs we ... mean that arguments used in mathematical practice in order to justify the correctness of theorems" (p.85). Indeed "... proofs really are devices that mathematicians use to convince one another ..." (Azzouni 2004).

Proofs can help to confirm a student's understanding of theorems, axioms, rules, givens and hypotheses. However, proofs are used not merely to convince students that a claim is true but also to give students

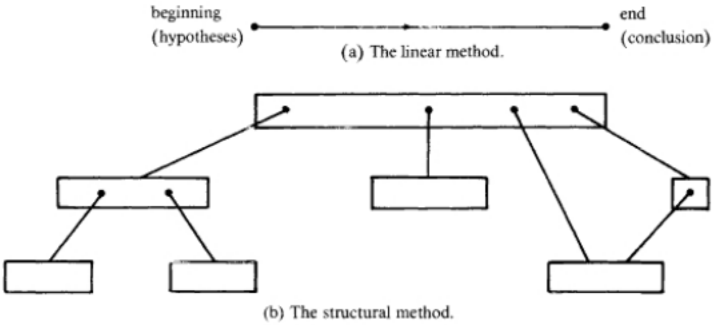


Figure 2.1: Leron (1983) comparison of linear proofs and structured proofs

some form of mathematical intuition (e.g., Hanna (1991); Hersh (1993)). For instance, Stylianides (2007) discusses a notion of proof in elementary school mathematics. In that work a proof was defined as a mathematical argument, which is to say a connected sequence of assertions used to justify a mathematical claim.

Another issue to consider is the nature of proof; whether the proof is *formal* or *informal*. Mathematicians have conflicting views on which proofs come under each heading. Philosophers argued that a formal proof has a logical structure and each line logically follows from the next and always require the full use of available axioms(Panza 2003). In contrast, an informal proof may not have clear linear structure. For example, a proof might include or be entirely made from diagrams (Hodds, Alcock, and Inglis 2014). Some mathematicians also argue that a textbook proof is a actually formal (Aberdein 2009) as it gives the learners all of the information to obtain a basic understanding of the proof, although some of the logical links might be not explicitly mentioned in the text. Since the proofs used in the studies reported in this thesis are taken from lecture-notes and textbooks, this distinction is important as readers from different disciplines will view the proofs described in different ways.

Some mathematics educators attribute students' difficulties in understanding proofs to the writing style of the traditional paragraph proof (Rowland 2002). To deal with these difficulties, alternative formats of writing proofs were suggested by several researchers. For example, Leron (1983) argued that presenting a proof in linear style is unsuitable for mathematical communication. He suggested a format for writing a proof that is arranged into *levels*. The top level gives the main line of the proof in general terms, while the bottom levels provides more details. Figure 2.1 shows a diagram taken from Leron (1983) comparing the linear method to the structured method. As Leron (1983) claims, the teacher's explanation and presentation are what provide the structure for proof. Although, the traditional linear approach still provides difficulty for learners in understanding proofs. However, transferring the explanations into the proof structured presentation itself might help learners (Hodds, Alcock, and Inglis 2014).

Other suggestions include generic proofs Rowland (2002), which explain the proof using a carefully chosen generic example which illustrates why a mathematical claim is true. There are several advantages for considering the format of generic proofs as a potential method for improving proof understanding.

Providing counterexamples are often used to show that a mathematical definition or claims does not hold for a particular mathematical objects. Using a counterexample is common in advanced level mathematics, and generally, using examples is unavoidable (Alcock and Inglis 2008). On the other hand, (Iannone et al. 2011) stated that "...simply asking students to generate examples about a concept may not substantially improve their abilities to write proofs about that concept, at least not more so than providing students with examples to read." Although it seems the use of examples has caused some disagreement, examples are considered important for the learning process (Hodds, Alcock, and Inglis 2014).

The research continues to show that students at both secondary and post-secondary levels have difficulty with proof (Hanna and Yan 2021). Some students impulsively apply useful strategies to understand proofs: they identify the structure of the proof, break the proof into sub-proofs, use examples to illustrate difficulties in the proof, and compare the proof they read with their own attempts (Weber 2015). According to the survey of mathematics professors reported by Weber (2015), over 63.9% of the professors preferred that mathematics students use these strategies when reading proofs. However, it is a common experience that many students reason without any justification or logic and that students find it difficult to distinguish the overall logical structure from the details of individual steps.

Some researchers suggested that changing the way students engaged with a proof might develop proof understanding. For example, Hodds, Alcock, and Inglis (2014) used self-explanation in a clinical study with students, and found that self-explanation helped students engage with proof, and hence develop deeper proof understanding.

Self-explanation training has also been shown to have successful results when learning probability (Renkl 1997) and geometry (Aleven and Koedinger 2002). Further, Renkl, Atkinson, and Gross (2004) showed that self-explanation is triggered by using faded worked examples. Self-explanation has also been used in other fields, such as physics problem solving (Chi et al. 1989), biology (Ainsworth and Burcham 2007) and history (Wolfe and Goldman 2005). Renkl (2002) found that participants who received self-explanation training with material that involved instructional explanation learned more and performed better.

Alcock and Simpson (2009) suggested using electronic proofs to improve proof comprehension. The electronic proof is designed as a mixture of audio and text explanations, (Hodds, Alcock, and Inglis 2014). The presentation is a series of slides which highlight one line at a time and provide added explanations on the proof structure. The audio is re-playable and used to clarify the explanations that a lecturer might give when presenting a proof (Alcock and Wilkinson 2011). According to Alcock and Simpson (2009), electronic proofs are designed to help students to understand proofs by stating the proof reasoning and structure explicitly without providing too much details on the screen at a time. Unfortunately, recent reviews of the literature have concluded that some of these formats are ineffective for improving proof comprehension. Electronic proofs provide extra information to enhance students' understanding without considering individuals' experience in reading mathematical proofs, reversal effects might occur and perhaps only novice learners benefit from electronic proofs (Roy, Alcock, and Inglis 2010). This is a reminder that even with the best intentions a teaching innovation may not actually prove to be helpful, indeed it may turn out to be counter-productive or even harmful.

A multidimensional framework for assessing proof comprehension presented by (Mejia-Ramos, Fuller, et al. 2012) in undergraduate mathematics, based upon the work of Yang and Lin (2008). The framework provides ways to assess students' understanding by seven different aspects of a proof. The first three types of assessment address students' comprehension of only one, or a small number, of statements within the proof. These types of assessment, which they called *locals*, are

1. Meaning of terms and statements: understanding the meaning of terms, symbols and definitions.
2. Logical structure: understanding the logical relationship between lines of a proof.
3. Justification of claims.

The remaining four items of assessment, which are called holistic, address students' understanding of the proof as a whole. These assessment types are:

1. Higher level ideas: identifying a good summary of the proof.
2. Identifying modular structure: understanding the main components and modules within a proof and the relationship between them.
3. General method: transferring the general ideas or methods to another task.
4. Illustrating with examples.

The framework is helpful for teachers and course designers when trying to write questions to assess students' understanding of a given proof, as it gives ideas for different types of questions can be asked.

One challenge in my research is defining ”formal” proofs. Philosophers emphasize strict logic, while mathematicians sometimes consider proofs in textbooks with missing steps as ”formal.” This ambiguity could affect how readers understand the proofs I analyze in this thesis, mainly drawn from lecture notes and textbooks. To address this, I'll explicitly state my definition for the proofs used in this thesis. In this thesis, I use the phrases *proof* and *mathematical reasoning* in an inclusive manner: at one end this is a purely algebraic calculation (with lines linked with equivalence statements, often omitted/assumed in written work) at the other end this is highly structured logic, working from formal definitions, e.g. as in an epsilon-delta argument in real analysis. The proofs used in this thesis are ”standard calculation proofs” which represent a prevalent format in introductory university mathematics. These proofs often involve manipulating well-established formulas and algorithms within specific domains. This choice stems from their frequent occurrence in undergraduate courses of mathematics and the relative ease with which students can engage in such manipulations. Furthermore, I have developed a coding scheme specifically designed to analyze this type of proof, enabling me to systematically investigate the format effect on students' writing of mathematical arguments. Through this approach, I aim to gain valuable insights into how to effectively assess students' proof construction, particularly online. This will contribute to understanding how different formats support students' overall understanding and construction of mathematical arguments. Indeed, Brown (2008) argues that pictures within proof don't automatically lack rigour.

2.2 Cognitive Load Theory

Cognitive Load Theory (Cognitive load theory) is a psychological theory about learning built on the premise that since the brain can only do so many things at once, one should be intentional about what questions it could be asked to do. The theory was developed in the 1980s by psychologist John Sweller to improve the teaching of mathematics and science. The assumptions of CLT are based on a model of human cognitive architecture, which is characterised by a limited working memory and unlimited long term memory. Working memory has a significant role in the acquisition of new skills and the storage of information into long term memory (Van Gerven et al. 2002). During learning, the working memory mainly process information while the long term memory stores the processed information.

CLT focuses on how constraints on our working memory affect what kinds of instruction are effective (Renkl, Atkinson, and Gross 2004). To avoid taxing the working memory's limited capacity, CLT suggests minimizing processing and/or storing activities that are not directly relevant to learning. The theory differentiates cognitive load into three types: intrinsic, extraneous, and germane (Sweller, Ayres, and Kalyuga 2011). Intrinsic load refers to load due to the complexity of the learning materials. Germane load refers to load due to mental activities that contribute immediately to learning. For example, when learning through faded worked examples, the self-explanation activities would be considered as germane load. An extraneous load is caused by mental activities that do not relate directly to the learning process. For example, if instructional materials contains graphics and text that are difficult to be integrated together, extraneous load is imposed as the learners will utilize much of their cognitive capacity to establish a coherence between the two information. As a result, little or no room will remaine for germane load.

Since learning is strongly effected by the potential of working memory, it has been argued that working memory should be filled by task-relevant operations, especially in learning complex material (Van Gerven et al. 2002). Therefore, researchers in cognitive load theory suggested ways to reduce the total cognitive load (Sweller, Ayres, and Kalyuga 2011) by reducing the intrinsic and extraneous cognitive load and increasing the germane cognitive load.

The intrinsic cognitive load of a learner is determined by two factors: the complexity of the information and the knowledge of the learner (Sweller, Merriënboer, and Paas 2019). In general, the intrinsic cognitive load can not be reduced but can be influenced by the prior knowledge of the learner (Sweller, Merriënboer, and Paas 2019). The extraneous cognitive load is not determined by the complexity of the information, but by how the information is presented and what the learner is required to do (Sweller, Merriënboer, and Paas 2019). Therefore, the extraneous cognitive load could be reduced by designing effective materials (Sweller, Merriënboer, and Paas 2019). Germane cognitive load was described as the cognitive load necessary to learn, which refers to the working memory resources used to manage intrinsic cognitive load rather than extraneous cognitive load (Sweller, Merriënboer, and Paas 2019). As extraneous cognitive load increases, less will be learned since there will be fewer resources available to deal with intrinsic cognitive load (Sweller, Merriënboer, and Paas 2019).

2.2.1 Cognitive load and expertise reversal

In instructional settings, it is usually assumed that if some instructional material is effective with novices, it should also work with higher knowledge learners (Kalyuga, Rikers, and Paas 2012). However, providing instructional guidance, which may be essential for novices, may have negative consequences for more experienced learners (Kalyuga, Ayres, et al. 2003). The *expertise reversal effect* refers to the reversal of the effectiveness of instructional techniques on learners with higher levels of prior knowledge (Kalyuga, Ayres, et al. 2003).

Expertise has been shown to alter the relative effectiveness of instruction (Kalyuga, Ayres, et al. 2003). Working memory is primarily influenced by the level of experience learners have in a domain (Kalyuga, Ayres, et al. 2003). The novice does not have sophisticated schemas associated with a given situation or task. The long-term memory of these inexperienced learners does not provide guidance about how to handle a given situation or task. Occasionally, instructional guidance can replace missing schemas and construct schemas if effective. Working memory load is minimized through effective instruction that directly provides instructional guidance (Tuovinen and Sweller 1999). Experts, on the other hand, construct mental representations of situations and tasks using their activated schemas. They may not need additional guidance because their schemas provide complete guidance. However, when instruction provides information to help learners construct appropriate mental representations, and experts cannot avoid attending to this information, an overlap occurs between the schema-based and redundant instruction-based components of guidance. This additional cognitive load can also occur when the learner recognizes that the instructional material is redundant and therefore chooses to ignore that information as best he or she can. Redundant information is often difficult to ignore (Kalyuga, Ayres, et al. 2003).

Some of the studies on the expertise reversal effect investigated cases of expertise reversal effect with instructional guidance provided to learners prior to independent problem-solving (Ericsson et al. 2006). For example, comparing verbal explanation of technical diagrams to diagrams without textual explanation; or worked examples of how to program industrial equipment to minimal guided problem solving. These studies found that studying worked examples produced better post-test results for novices than minimally guided forms of instruction (Ericsson et al. 2006). However, the difference reversed or disappeared for higher knowledge learners. According to Ericsson et al. (2006), unguided learning required higher cognitive demands for low knowledge learners than guided instruction. On the other hand, for higher knowledge learners, studying information that included guided instruction could be harmful or redundant.

Feedback provided to learners during problem solving is a common form of instructional support. Explicit feedback in practice problems has been shown to be beneficial to students with low prior knowledge when learning statistics (Krause, Stark, and Mandl 2009). However, no effect of feedback was found for students with higher prior knowledge. In using computer networking training simulation with university students Nihalani, Mayrath, and Robinson (2011) found that novices learned more effectively when they learned individually with feedback rather than collaboratively with similar feedback. Learners with higher prior knowledge, however, benefited more from collaborative feedback than from individual feedback (Nihalani, Mayrath, and Robinson 2011).

Ericsson et al. (2006) reported expertise reversal effects by comparing traditional product-oriented worked examples that include only step-by-step solutions without explaining why each step is used and process-oriented worked examples that provided such explanations. Results indicated that only low prior knowledge students benefited from the added explanation during the initial stages of training. As the learners become more knowledgeable, the explanations became redundant causing an expertise reversal effect.

One of the popular pedagogical designs that is recommended by researchers in Cognitive load theory is worked examples instruction. Worked example might help in manipulating the working memory resources during learning to reduce extraneous cognitive load (Renkl 1997). However, providing worked out steps for learners who already know how to solve the problem imposes extraneous load without benefits (Renkl, Atkinson, and Gross 2004).

A theoretical framework suggested by Koedinger, Corbett, and Perfetti (2012) to organize the development of instructional theory for guiding the design, development and continual improvement of effective academic course materials, technologies and instructor practices. Koedinger, Corbett, and Perfetti (2012) defined a knowledge component (KC) as an acquired unit of cognitive structure or function that can be implied from performance on a set of related tasks. A knowledge component is used broadly to describe pieces of cognition or knowledge, such as schema or misconception as well as more common terms like concept, principle, fact or skill. According to Koedinger, Corbett, and Perfetti (2012), there are many types of knowledge components, where some are simpler and faster to use and others are complex and slower to use. Recognizing types of knowledge components and their complexity is critical not only because deep analysis of domain knowledge is an important source for innovative instructional design, but also because different types of learning processes and instructional treatments may be more or less useful depending on the complexity and nature of the target knowledge.

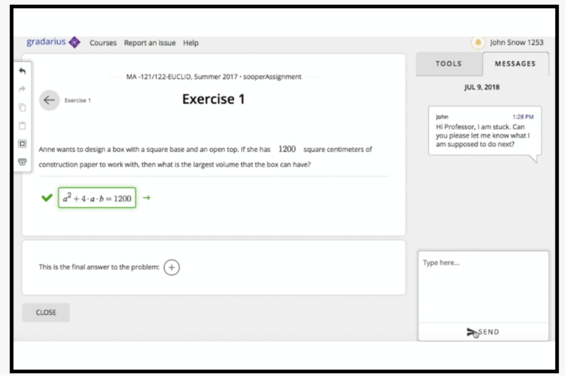
2.3 Online assessment

Assessment is a central part of the teaching and learning process in all educational contexts, including higher education institutions. The improvement of technology and e-learning systems, has resulted in a high demand for ways and means of assessing students in such a system. Systems have been developed by universities, companies and as part of virtual learning environments. Introducing computerised marking of mathematics assessments benefits not only students but lecturers as well. Computer aided Assessment can promote different tasks, such as access to large problem databases and test prototypes, marking facilities, immediate feedback or access to statistical packages to analyse students' data (Engelbrecht and Harding 2005). This technology certainly reduces the time taken for marking, but it also provides personalised, detailed and immediate feedback (Delius 2004). Moreover, teachers have the opportunity to be directly involved in designing questions (Sangwin 2004). More sophisticated computer aided assessment systems have enabled mathematical questions to be broken down into steps and have provided feedback. The online assessments systems with focus on separating a question into steps (for example, CALM, CUE nad PASS-IT) were reviewd by (Jordan 2013).

Using technology for facilitating mathematical tasks promotes a series of advantages. Students appear to enjoy the virtual environments as these provide the chance to practise as much as needed and to receive feedback (Pitcher, Goldfinch, and Beevers 2002). Jordan (2013) reported that such assessment tools enhance students' confidence and changes in their attitudes towards mathematics. Technology contributes to increasing the appeal of mathematics as students find the subject more difficult. Technology-based assessment assist students in self-evaluation. On the other hand, alternative assessment methods help students and teachers to recognize difficulties early on in the semester. For example, Gradarius provides immediate feedback for each solution step. The feedback is also supported by instruction, i.e. if a student makes a mistake, the student also receives layered instruction that is available with the problem, including hints, notes, concept review, and video. Instructors can also add their own instruction. Furthermore, at any point in the process of entering a solution, students can reach out to the instructor (or TA) using the Gradairus messaging functionality (see Figure 2.3.) The messaging is linked directly to the student's solution in progress, making quick work of providing guidance or redirection.

Gradarius problems are associated with "reference solution(s)" that serve as the standard against which student solutions are automatically graded. Each "reference solution" has identified "essential steps" which are deemed critical to the reference solution. Instructors can choose to offer partial credit for essential steps or offer credit based on only the final answer to the problem.

Gradarius allows line by line (or we call it row by row) solving. Every step (or cell) within that row is analyzed independently, and Gradarius assesses a student's work algorithmically to determine when an error has been made. This feedback and analysis happens at every step of the solution, not just the final answer, and is tracked within the system so that professors can see, for instance, mistake profiles of what types of errors students make as they are solving problems, and not just what types of problems students are making mistakes on. In addition to the algorithmic feedback, problems may also have "Essential Steps' assigned, which help students know when they are on the right track and are also used to offer partial credit grading, should an instructor desire it. Feedback is available instantaneously, if desired, but



MA -121/122-EUCLID, Summer 2017 • sooperAssignment

Exercise 1

Anne wants to design a box with a square base and an open top. If she has 1200 square centimeters of construction paper to work with, then what is the largest volume that the box can have?

✓ $a^2 + 4 \cdot a \cdot b = 1200$ →

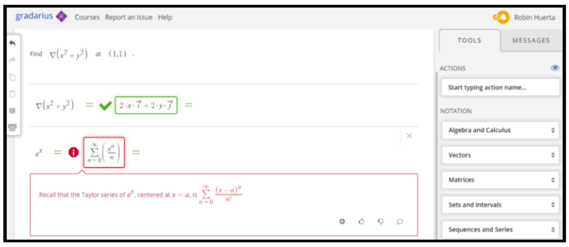
This is the final answer to the problem: ⊕

JUL 9, 2018

John 1:28 PM
Hi Professor, I am stuck. Can you please let me know what I am supposed to do next?

Type here...

Figure 2.2: Screenshot of the Gradarius platform adapted from https://www.gradarius.com



Find $\nabla(x^2+y^2)$ at $(1,1)$.

$\nabla(x^2+y^2) = $ ✓ $2 \cdot x \cdot \vec{i} + 2 \cdot y \cdot \vec{j} =$

$e^x = $ $\sum_{n=0}^{\infty}\left(\frac{x^n}{n}\right) =$

Recall that the Taylor series of e^x, centered at $x = a$, is $\sum_{n=0}^{\infty} \frac{(x-a)^n}{n!}$

Start typing action name...

Figure 2.3: Gradairus messaging functionality adapted from https://www.gradarius.com

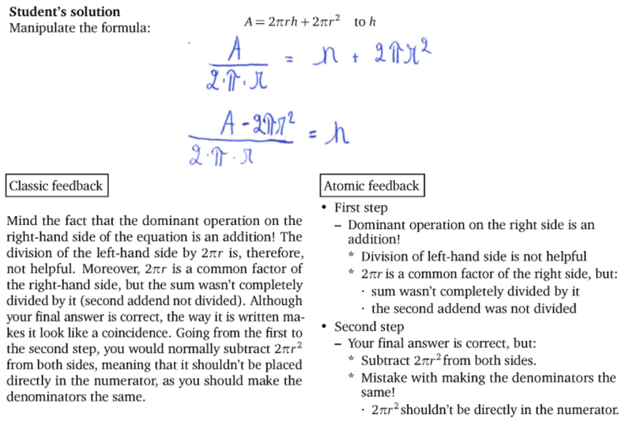
comes with multiple levels of support and assistance so that students receive the least amount of support needed to solve the problem themselves. First, for instance, they will just be informed that they've made an error of some kind, and then they will be informed about what sort of error they've made, and finally, they will see, for instance, a full explanation of the chain rule and how they've misapplied it.

An older example is AIM (Alice Interactive Mathematics). AIM does not penalise wrong mathematical language errors (Sangwin 2004). AIM provides students the possibility to validate answers before being marked. Moreover, students have the opportunity to review and practise as much as necessary.

The potential effect of using a computer-aided assessment like Maple T.A. on students' ways of working with mathematics was discussed by (Rønning 2017). He reported when only the final answer is evaluated, and only with right or wrong, seems to encourage students to pay less attention to how the answer is reached and turns into a “random hunt for the correct answer”.

According to Rønning (2017), when comparing Maple T.A. problems to written hand-ins, the students are focusing on the difference between presenting the answer to the computer and presenting the answer to a person. They expect the person to look at the whole solution in the problem and not just the final answer and therefore they take greater care with the presentation. An important impact of this is that they realize that they learn more from the actual procedure of presenting the argument to an assumed reader (Rønning 2017).

Mixing between the automatic and human marking (semi-automatic marking) is expected to contribute to better balances and allow both students and teachers to take advantage of the strengths of both methods of marking. By automatic assessment, students will be able to receive immediate feedback, doing more attempts and also have the flexibility of working anywhere and anytime. While using human marking will let students work with the well-understood paper-based format, with fewer constraints to explaining their reasoning. In addition, the advantage of mixing between the two methods of marking for teachers are also enormous. Moons, Vandervieren, and Colpaert (2022) proposed to re-use feedback in their semi-automated system: teachers write feedback items, and the computer saves these items so they can easily be re-used when other students make similar mistakes. In order to use this approach, a teacher needs to firstly identify the independent errors occurring and then write small, independent feedback items for each error. Figure 2.4 illustrates a comparison between classic and atomic feedback. Atomic



Student's solution
Manipulate the formula: $A = 2\pi rh + 2\pi r^2$ to h

$$\frac{A}{2\cdot\pi\cdot r} = h + 2\pi r^2$$

$$\frac{A - 2\pi r^2}{2\cdot\pi\cdot r} = h$$

Classic feedback

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it shouldn't be placed directly in the numerator, as you should make the denominators the same.

Atomic feedback

- First step
  - Dominant operation on the right side is an addition!
    - Division of left-hand side is not helpful
    - $2\pi r$ is a common factor of the right side, but:
      - sum wasn't completely divided by it
      - the second addend was not divided
- Second step
  - Your final answer is correct, but:
    - Subtract $2\pi r^2$ from both sides.
    - Mistake with making the denominators the same!
      - $2\pi r^2$ shouldn't be directly in the numerator.

Figure 2.4: A comparison between classic and atomic feedback adapted from (Moons, Vandervieren, and Colpaert 2022)

feedback consists of a set of formulation requirements that makes feedback significantly more reusable; instead of writing long comments describing lots of different mistakes at once (Moons, Vandervieren, and Colpaert 2022).

Moons, Vandervieren, and Colpaert (2022) suggested that the advantage of dividing the atomic feedback items into two separate sub-items is that the sub-items can be reused independently for a potentially much larger group of students: those who only got the unknown wrong, those who got only the initial equation wrong, and those who got both wrong. Moons found that the word frequencies is similar in both feedback types, while atomic feedback contains fewer abbreviations, more section titles, and more concrete instructions.

The online assessment facilitates quick and clear reports on students results and progress. This makes it easier to give useful feedback on how students are doing, areas where they are strong, and what areas of learning require attention. Printing and circulating work on paper and organizing shipments of completed scripts to markers is a time-consuming and costly process.

Students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). Since assessing a full proof online is currently impossible, one possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which e-assessment can then be applied. One draw-back of breaking up questions is that it requires a significant investment of time and experience to develop a suitable question bank. In this section, I discuss three different formats for writing open-ended questions: two-column, typing, and faded worked examples. I also provide examples of using these formats in the contemporary online assessment.

2.3.1 Formats in online assessment

Moodle offers around thirty different types of questions, according to T. J. Hunt (2012). The range of question types (e.g. 'multiple-choice', 'drag-and-drop', 'calculated', 'numerical', 'true-false') adds variety, but Hunt's survey of more than 50,000,000 questions from around 2,500 Moodle sites found that about 90% of the questions in use were selected-response questions i.e. question types like multiple-choice or drag-and-drop where options are presented for a student to select, in contrast to 'constructed-response' where students construct their own response. Procedural multiple-choices items typically present a mathematical object, such as an equation, and an instruction to transform the object into a specified form. The equivalent constructed response item would contain the same question stem, but the answer options would be removed and replaced with a space to write the answer, or a text box if administered as a computer-based assessment. Removing the options from multiple-choices items in this way creates what are called stem-equivalent constructed-response items. Even the simplest of multiple-choice quizzes can enable students to check their understanding of a wide range of topics, whenever and wherever they choose to do so.

Jordan (2013) reported the pros and cons of selected-response and constructed-response questions. By using selected-response questions, some issues of data-entry could be eliminated, particularly problematic in constructed-response questions when symbolic notation is required, for example in mathematics (Sangwin 2013). Selected-response can assess a wide range of knowledge whereas a test comprising constructed-response questions is likely to be more selective in its scope. Furthermore, selected-response questions avoid problems with wrong or incomplete answer. However, constructed-response answers may be incorrectly marked.

Constructed-response questions provide a method for gaining insight into the knowledge of a student since students have to provide the solution to a question themselves. However, multiple-choice questions provide students with all the possible answers. Constructed-response questions still preferable by test developers for mathematics tests in schools (B. Cooper and Dunne 2000). On the other hand, the main drawback of constructed-response questions is the opportunity for subjectivity in the marking (e.g.(Laming 2004)). There are some limitations of using selected response questions to assess higher order learning outcomes. Some researchers have reported a gender effect on question format. Mazzeo, Schmitt, and Bleistein (1993), Hassmén and D. P. Hunt (1994) and Livingston and Rupp (2004) found that performance for males is higher than performance for females when multiple-choice items are used. Kuechler and Simkin (2003) found that students for whom English was a second language sometimes had difficulty determine the wording implication of multiple-choice questions. In some multiple-choice questions, students can work back from the options to select the right option, so the question is not assessing the learning outcome that it claims to be assessing. For example, a question that asks students to integrate a function can be answered by differentiating each of the options provided (Sangwin and Jones 2017). Previous work, such as Sangwin and Jones (2017), found a format effect between multiple-choice and constructed-responses when students work with reversible mathematical processes. The study proved that students solve a multiple-choice version by verifying the answers by the direct method, not by undertaking the actual inverse calculation. Students may be guessing when answering selected-response questions, so their teacher has no way of telling what the student really understands (Crisp 2007). So it is reasonable to expect a format effect might exist between the traditional and other formats when writing complete mathematical arguments. A *format effect* occurs when the format for exercises affect the rate of successful outcomes of the exercise. The format effect is important to consider because the method of marking depends on the format of a mathematical response.

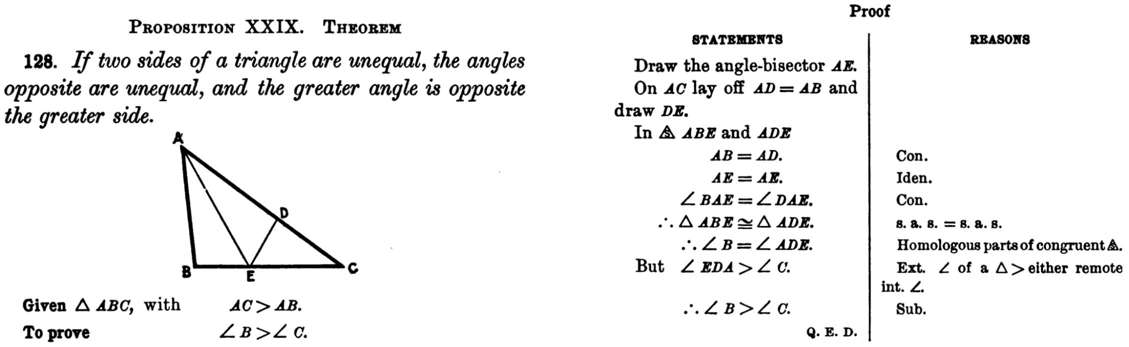
There are some issues that need to be considered when choosing the format of a mathematical response; for example: extended working and justification, logical reasoning, symbolism, diagrams and graphs. Two variables which might influence marks on a writing test: (1) mode of presentation, which is the format of responses presented to scorers; and (2) mode of composition, which is the medium (paper or computer) used by respondents for producing responses (Russell and Tao 2004).

Two-column format

In this section of the literature review, I define two-column proofs. I provide examples of using the two-column format in the contemporary online assessment. There are cognitive aspects linked to two-column proofs. The constraints in the interface might help students in introductory classes to understand the basic elements of mathematical proof. This proof format requires students to explicitly express conditions for each step and combine them in order to connect assumptions with conclusions. In this way two-column proofs might be helpful in creating an opportunity for students to self-explain what they are doing.

To begin, there is a brief discussion on the history of two-column proof and the development and use of two-column proof. A two-column proof is a method of presenting a mathematical proof or argument by using a tabular layout with two-columns.

> A proof in two-column form is written out in a chronological sequence that correlates with numbered assertions: line 1 is written first, line 2 is written second, and so on. Moreover, the two-columns are written in alternating sequences: first the statement of line 1, then the reason for line 1; then the statement of line 2, then the reason for line 2; and so on. (Weiss,



Proposition XXIX. Theorem

128. *If two sides of a triangle are unequal, the angles opposite are unequal, and the greater angle is opposite the greater side.*

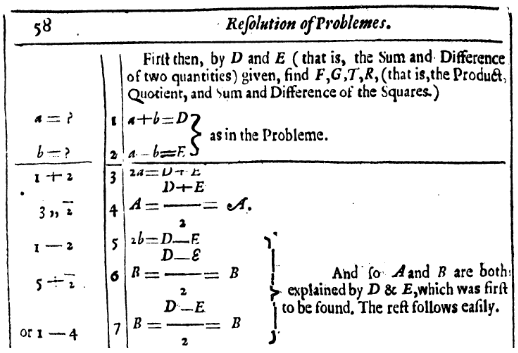
Given $\triangle ABC$, with $AC > AB$.

To prove $\angle B > \angle C$.

Proof

STATEMENTS	REASONS
Draw the angle-bisector AE.	
On AC lay off $AD = AB$ and draw DE.	
In ⚠ ABE and ADE	
$AB = AD$.	Con.
$AE = AE$.	Iden.
$\angle BAE = \angle DAE$.	Con.
$\therefore \triangle ABE \cong \triangle ADE$.	s. a. s. = s. a. s.
$\therefore \angle B = \angle ADE$.	Homologous parts of congruent ⚠.
But $\angle EDA > \angle C$.	Ext. $\angle$ of a $\triangle >$ either remote int. $\angle$.
$\therefore \angle B > \angle C$.	Sub.
Q. E. D.	

Figure 2.5: A school geometry proof in two-column format



58 Reſolution of Problemes.

Firſt then, by D and E (that is, the Sum and Difference of two quantities) given, find F,G,T,R, (that is, the Product, Quotient, and Sum and Difference of the Squares.)

$a = ?$	1	$a+b=D$	as in the Probleme.
$b = ?$	2	$a-b=E$	
$1+2$	3	$2a=D+E$	
$3 \div 2$	4	$A=\dfrac{D+E}{2}=A$.	
$1-2$	5	$2b=D-E$	And ſo A and B are both explained by D & E, which was firſt to be found. The reſt follows eaſily.
$5 \div 2$	6	$B=\dfrac{D-E}{2}=B$	
or $1-4$	7	$B=\dfrac{D-E}{2}=B$	

Figure 2.6: An algebraic derivation in two-column format

P. Herbst, and Chen 2009)

The two-column proof has been used in the USA to teach students how to prove since the early twentieth century (P. G. Herbst 2002), mostly in school geometry. P. G. Herbst (2002) was interested whether the custom of using two-column proofs developed as a practical way to ensure that every student should be able to write proofs. In particular, an early example is Schulze and Sevenoak (1913) who aimed to *"give to the student mental training instead of teaching him mere facts; to develop his power instead of making him memorize"*. An example from Schulze and Sevenoak (1913), pg 67, is shown in Figure 2.5 with the theorem to the left, and the two-column proof to the right. Such two-column proofs have become somewhat routine, and in this sense the two-column proof format has been criticised, e.g. see (Weiss, P. Herbst, and Chen 2009).

The use of page layout to structure mathematical argument is not new, indeed the two-column idea can be traced back at least to the algebra book of (Brancker, Pell, and Rahn 1668), but the format is not currently popular. An algebraic example from (Brancker, Pell, and Rahn 1668) is shown in Figure 2.6. Note that Figure 2.6 contains somewhat archaic algebraic 17^{th} Century notation: see (Stedall 2002) for a detailed discussion.

The two-column proof format separates out the individual steps from the justification for the legitimacy of individual steps. Using the format to consciously separate the steps from their justification is highly likely to reduce cognitive load, particularly for novices, by providing a structured and explicit format to work within. Hence when measuring the effectiveness of using such proofs with instructional designs, it is expected to see the expertise reversal effect (Kalyuga, Rikers, and Paas 2012) which refers to the reversal of the effectiveness of instructional techniques on learners with differing levels of prior knowledge. In particular, the two-column format might well significantly help early proof attempts, but hinder experts for whom the scaffolding is a distraction. Previous research on students' conceptions of proof, and changes in contemporary technology, provide a number of reasons why I believe a fresh re-evaluation of the two-column format is needed. In particular, the constraints in the interface might help students in introductory classes to understand the basic elements of mathematical proof. The two-column proof format requires students to explicitly express conditions for each step and combine them in order to connect assumptions with conclusions.

The two-column proof creates a systemic environment in which errors may be easier to spot, which a traditional rhetorical style does not.

The two-column proof format perhaps makes it easier for students to self monitor their own work after each step. Indeed, the two-column format can be both a resource and a constraint in engaging students in proving (Weiss, P. Herbst, and Chen 2009). In the past, the two-column proof may have led students to perform routines, i.e. do proofs, without necessarily understanding what they were doing, or why they were proving at all. As P. G. Herbst (2002) concludes

> The two-column proving custom was an accomplishment of geometry instruction in the sense that it helped comply with a mandate. But that accomplishment did not come for free. It brought to the fore the logical aspects of a proof at the expense of the substantive role of proof in knowledge construction. Questions about the relevance or the strength of the propositions proved, or about the accomplishments and potential use of the theories being developed, were left in the background. (P. G. Herbst 2002)

The recent interest in using a two-column format for algebraic derivations with contemporary online tools risks not learning the lessons from prior experience of using this format in geometry. I am certainly not advocating for a pedagogical practice, but I am trying to understand how constraints on students' input (particularly with reference to potentially designing an interface to technology) might affect students' responses. Furthermore, the two-column format potentially separates steps in a proof, from justification of those steps. The format has the potential to help us better understand the nature of mathematical arguments, and how students write these arguments, by comparing constrained (two-column) arguments from unconstrained free form arguments.

There are cognitive aspects linked to using constraints, which I believe are potentially beneficial. The constraints in the interface might help students in introductory classes to understand the basic elements of mathematical proof. For example, in the the two-column proofs, the format requires students to explicitly express conditions for each step and combine them in order to connect assumptions with conclusions. The two-column proof creates a systemic environment in which errors may be easier to spot, which a traditional rhetorical style does not.

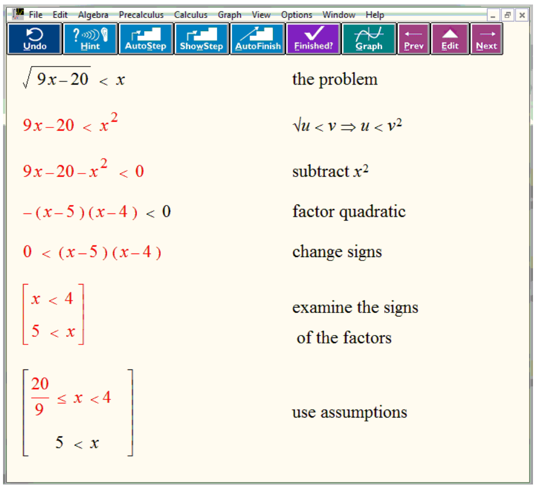
The two-column proof lends itself to online assessment, with a number of possible options. For example, teachers might expect students to provide (1) the actual steps in the proof, (2) the justification, or both.

Our first example is by Prof. Michael Beeson at San Jose State University, who has been developing educational software since 1985.

Mathpert, MathXpert, and more recently an online system `http://www.helpwithmath.com/` are tutoring systems for mathematics, more specifically solving equations. Mathpert and MathXpert allow its user to construct a step-by-step solution to a wide range of problems in elementary algebra, and more generally in calculus, trigonometry and simple inequalities.

> *Mathpert* is intended to replace paper-and-pencil homework in algebra, trig, and calculus, retaining compatibility with the existing curriculum while at the same time supporting innovative curriculum changes; to provide easy-to-use computer graphics for classroom demonstration in those subjects, as well as for home study; to replace or supplement chalk-and-blackboard in the classroom for symbolic problems as well as graphs. (Beeson 1998)

Students pick a topic and then use a *calculation window* to solve the problem in a step-by-step fashion. By design, users select part (or all) of an expression and the software provides a menu of operations which can be performed on that selection. MathXpert actually performs that operation automatically. MathXpert also a wide range of options, including hints, or it can even complete the whole problem automatically.


File Edit Algebra Precalculus Calculus Graph View Options Window Help
Undo
Hint
AutoStep
ShowStep
AutoFinish
Finished?
Graph
Prev
Edit
Next

$\sqrt{9x-20} < x$ the problem

$9x-20 < x^2$ $\sqrt{u} < v \Rightarrow u < v^2$

$9x-20-x^2 < 0$ subtract x^2

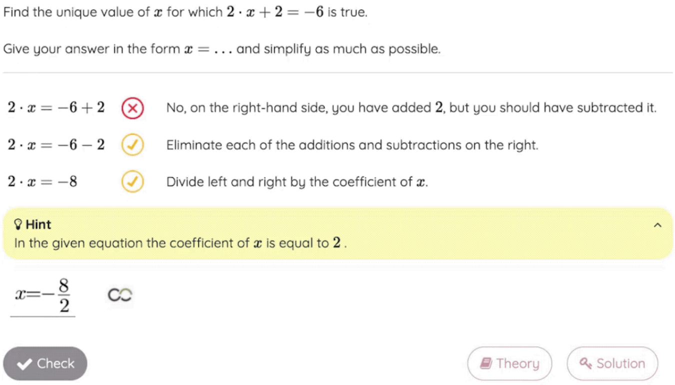
$-(x-5)(x-4) < 0$ factor quadratic

$0 < (x-5)(x-4)$ change signs

$\left[\begin{array}{l} x < 4 \\ 5 < x \end{array}\right]$ examine the signs of the factors

$\left[\begin{array}{l} \frac{20}{9} \le x < 4 \\ 5 < x \end{array}\right]$ use assumptions

Figure 2.7: The user interface in the MathXpert system



Find the unique value of x for which $2\cdot x+2=-6$ is true.

Give your answer in the form $x=\ldots$ and simplify as much as possible.

$2\cdot x=-6+2$ ⊗ No, on the right-hand side, you have added 2, but you should have subtracted it.

$2\cdot x=-6-2$ ✓ Eliminate each of the additions and subtractions on the right.

$2\cdot x=-8$ ✓ Divide left and right by the coefficient of x.

Hint
In the given equation the coefficient of x is equal to 2.

$x=-\frac{8}{2}$

Figure 2.8: The user interface in the SOWISO system

The user interface from the desktop version of the MathXpert system is shown in Figure 2.7. Notice this interface is fundamentally a two-column format, but the emphasis is not on the students performing the calculations themselves, rather students are required only to make decisions on which direction to go.

A more recent system is SOWISO, `https://sowiso.nl/`. This online learning environment also expects students to perform calculations, and the software analyses students' input and attempts to identify what they have done. The software seeks to identify any mistakes and provides specific hints and feedback. The design provides students with this feedback regardless whether their input is judged correct or incorrect, again in a fundamentally two-column format, see Figure 2.8.

MathXpert and SOWISO represent opposite extremes in design between (i) decision making (MathXpert) and (ii) user computation (SOWISO). Clearly there are many design decisions required when implementing such software.

In this thesis, what I am reporting is paper-based foundational research about what justifications students actually provide, and how the two-column format might influence these justifications. The study is reported in Chapter 4.

Typing

Typing is the process of inputting text by pressing keys on a typewriter, keyboard, cell phone, or a calculator.

Writing long mathematical responses traditionally in pen and paper style might be easier for students to freely explain their reasoning, draw sketches and using mathematical symbols. On the other hand, typing mathematical symbols or equations can be very difficult and is not an issue isolated to just assignments applications (Sangwin and Ramsden 2007). Students often struggle with typing mathematical symbols and graphing online (Howard and Beyers 2020). However, typing provides an opportunity to revise and edit the work without starting all over as when writing traditionally using pen and paper which is time consuming.

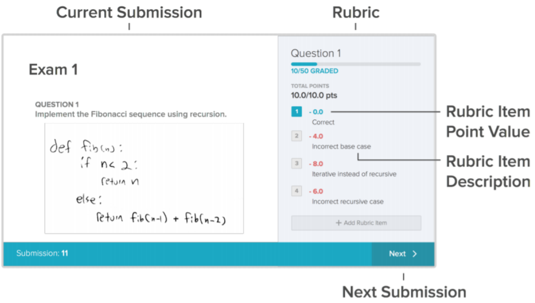
Overall familiarity with technology also seems to play an important role in student performance (Mogey, Paterson, et al. 2010). Lack of experience may oblige students to use inefficient techniques to add or remove electronic text. Using the undo, redo, erase, cut, or paste functions while typing may actually assist writers with their organization. Such manipulation of blocks of text is not as easy or clean when handwriting.

Two variables might influence marks on a writing test: (1) mode of presentation, which is the format of responses (handwritten or typed text) presented to scorers; and (2) mode of composition, which is the medium (paper or computer) used by respondents for producing responses (Russell and Tao 2004). Several studies have found that a type-written essay will be marked more harshly than an identical handwritten text, although the difference in grades is not always large. The reason for the difference is not known for certain but seems likely to be related to an expectation that handwritten text is a first draft standard whereas typed text would normally have been revised more thoroughly (Mogey, Paterson, et al. 2010). On the other hand, the early studies by James (1927) and Markham (1976) found that quality of handwriting significantly influenced grades given to essays. Markham (1976) investigated the influences of handwriting quality on teacher evaluation of written work and reported that papers with better handwriting consistently received higher scores than did those with poor handwriting regardless of the quality of the content. According to Markham (1976), the analysis of variance indicated that the variation in scores explained by handwriting was significant. Many researchers have consistently studied the association of poor handwriting with lower marks from graders (e.g., (Chase 1986) and (Markham 1976). (Chase 1968), and (Soloff 1973) have reported that secondary school teachers give significantly higher ratings to those student papers which have handwriting of good quality, regardless of content. Chase (1968), and Soloff (1973) have reported that secondary school teachers give significantly higher ratings to those student papers which have handwriting of good quality, regardless of content. Thus, although there is little support for the supposition that good handwriting always corresponds to a good paper, these research reports suggest that handwriting of good quality may lead to higher marks while handwriting of poor quality may lead to lower marks (Markham 1976).

Mogey, Cowey, et al. (2012) reported how students opt to respond to examination questions when permitted to handwrite or word process. Students from all years in the Divinity school were invited to contribute. As part of the routine examination diet, participants were offered the choice to handwrite or type their examination. A few weeks prior to the examination, students were given a demonstration of the software to be used, and were allowed unlimited access should they wish to familiarize themselves with the technology. The majority of students declined the offer to type their essays. Only 16 of the 204 candidates for the examination chose to type; 125 of the survey respondents chose to handwrite in the examination.

Perhaps this will change over time, if typing becomes a norm and handwriting subsides in popularity. students are much more familiar with photographing and uploading responses. So, if students who choose to type are more likely to go back to edit their text, and since text which has been more thoroughly reviewed tends to earn higher marks, this all suggests that it will be to the advantage of all but the slowest typists to choose to type (Mogey, Cowey, et al. 2012). Thus, although there is little support for the supposition that good handwriting always refer to a good paper, these research reports suggest that handwriting of good quality may lead to higher marks while handwriting of poor quality may lead to lower marks (Markham 1976).

Since COVID-19, online learning has been the primary method to keep learning going, requiring substantial support. Consequently, students have become more comfortable typing or uploading handwritten responses via online assessment systems. However, typing mathematical arguments is all about meaning and thus human marking is needed. An example for online assessment system that support human marking is Gradescope. Figure 2.9 illustrated an example for the grading interface in Gradescape adapted



Exam 1

QUESTION 1
Implement the Fibonacci sequence using recursion.

```
def fib(n):
    if n< 2:
        return n
    else:
        return fib(n-1) + fib(n-2)
```

Question 1
10/50 GRADED
TOTAL POINTS
10.0/10.0 pts

1 - 0.0 Correct
2 - 4.0 Incorrect base case
3 - 8.0 Iterative instead of recursive
4 - 6.0 Incorrect recursive case

Submission: 11

Figure 2.9: An example illustrated Gradscope's grading interface.

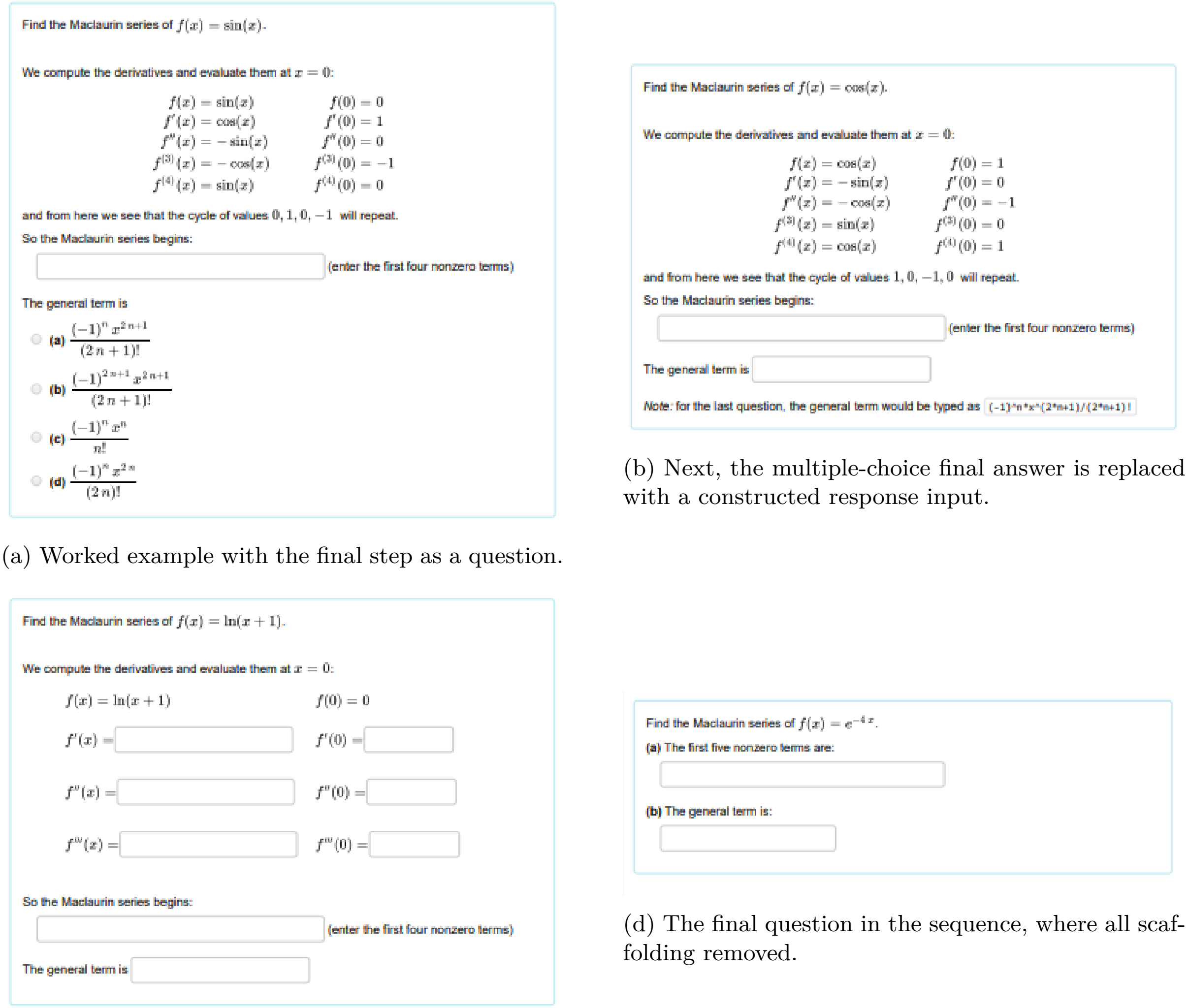
from (Singh et al. 2017). The interface is divided in two parts. In the left part, a single student's submission can be seen by the grader to the question they're grading. On the right part, the rubric which is composed of multiple rubric items that each have point values and descriptions. After grading, graders can navigate to the next submission for the same question (Singh et al. 2017). The instructor can add any feedback to students in the rubric description or as comments. Students can view their grades and feedback on their submission when the instructor publishes the grades from the assignment's Review Grades page. One of the advantages of Gradescope is that it can be linked individual course to courses in the virtual learning environment of choice, such as Blackboard or Moodle with no additional log in requirements.

Online assessment systems are varied depend on the functionalities or the purpose for each systems. In term of freedom, SOWISO has a constraint to work in line by line, while in Gradscope students have more freedom by uploading a picture of their work. In addition, if we consider time, human marking takes time. However, automatic assessment can save time but is a difficult technology to set. It has not been possible to mention all online assessment systems and in particular the systems described in this research are given as exemplars. Where choices have been made between systems and papers to include, those that have been used or evaluated in the context of mathematics science or related disciplines have been favoured.

Faded worked examples

As students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). One possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. In this case, a format such as faded worked example is effective to design mathematical online task. I discuss faded worked example in this Section. Faded worked examples heavily influence the idea of *separated concerns* which I developed to design STACK questions and which are introduced in Chapter 6.

The worked example effect was first demonstrated by Sweller and G. Cooper (1985), who found that algebra students learned more studying algebra worked examples than solving the equivalent unguided problems. A worked example consists of three components: a problem formulation, the solution steps, and the final solution itself. In worked examples instruction, students are provided with an instructional sheet which presents in a step by step procedure to solve a problem. A finding from the cognitive science suggested that learners are more likely to benefit from studying worked examples than from unguided problem solving (Kirschner, Sweller, and Clark 2006). The effect of worked example instruction occurred when students can transfer and apply their understanding of the worked examples to solve a similar problem. Based on cognitive load theory, solving problem requires problem-solving search and search only occurs using our limited working memory. Worked examples reduce working memory load since search is reduced or eliminated, and attention (i.e., working memory resources) are directed towards learning the essential relation between problem-solving moves (Kirschner, Sweller, and Clark 2006). For novices, studying worked examples is more beneficial than discovering or constructing a solution to a problem (Kirschner, Sweller, and Clark 2006).



Find the Maclaurin series of $f(x) = \sin(x)$.

We compute the derivatives and evaluate them at $x = 0$:

$$\begin{aligned} f(x) &= \sin(x) & f(0) &= 0 \\ f'(x) &= \cos(x) & f'(0) &= 1 \\ f''(x) &= -\sin(x) & f''(0) &= 0 \\ f^{(3)}(x) &= -\cos(x) & f^{(3)}(0) &= -1 \\ f^{(4)}(x) &= \sin(x) & f^{(4)}(0) &= 0 \end{aligned}$$

and from here we see that the cycle of values $0, 1, 0, -1$ will repeat.

So the Maclaurin series begins:

(enter the first four nonzero terms)

The general term is

○ **(a)** $\frac{(-1)^n\, x^{2n+1}}{(2n+1)!}$

○ **(b)** $\frac{(-1)^{2n+1}\, x^{2n+1}}{(2n+1)!}$

○ **(c)** $\frac{(-1)^n\, x^n}{n!}$

○ **(d)** $\frac{(-1)^n\, x^{2n}}{(2n)!}$

(a) Worked example with the final step as a question.

Find the Maclaurin series of $f(x) = \cos(x)$.

We compute the derivatives and evaluate them at $x = 0$:

$$\begin{aligned} f(x) &= \cos(x) & f(0) &= 1 \\ f'(x) &= -\sin(x) & f'(0) &= 0 \\ f''(x) &= -\cos(x) & f''(0) &= -1 \\ f^{(3)}(x) &= \sin(x) & f^{(3)}(0) &= 0 \\ f^{(4)}(x) &= \cos(x) & f^{(4)}(0) &= 1 \end{aligned}$$

and from here we see that the cycle of values $1, 0, -1, 0$ will repeat.

So the Maclaurin series begins:

(enter the first four nonzero terms)

The general term is

Note: for the last question, the general term would be typed as `(-1)^n*x^(2*n+1)/(2*n+1)!`

(b) Next, the multiple-choice final answer is replaced with a constructed response input.

Find the Maclaurin series of $f(x) = \ln(x+1)$.

We compute the derivatives and evaluate them at $x = 0$:

$f(x) = \ln(x+1)$ $f(0) = 0$

$f'(x) =$ $f'(0) =$

$f''(x) =$ $f''(0) =$

$f'''(x) =$ $f'''(0) =$

So the Maclaurin series begins:

(enter the first four nonzero terms)

The general term is

(c) The fully scaffolded version of the question.

Find the Maclaurin series of $f(x) = e^{-4x}$.

(a) The first five nonzero terms are:

(b) The general term is:

(d) The final question in the sequence, where all scaffolding removed.

Figure 2.10: A sequence of faded worked examples of calculating Maclaurin series

A *faded worked example* is a worked example in which some of the solution steps have been removed so that a student can complete the missing steps. A progressive sequence of faded worked examples is a pedagogic device in which the scaffolding provided by the solutions steps within a worked example are systematically removed, requiring students to take progressive responsibility for completing the problem. Typically, a student will be provided a progressive sequence of problems of a particular, similar type in which amounts of the solution already worked out changes. Fading worked examples leads to higher transfer performance and enhanced student learning (Renkl, Atkinson, and Gross 2004).

For enhancing cognitive skill acquisition, Renkl, Atkinson, and Gross (2004) found that it is useful to use faded worked examples before starting to solve problems independently. This is particularly useful where there is a *model worked solution* which a student is expected to learn. Removing steps from the end of the problem, i.e. first removing the last step, has been found to be most favourable for learning (Renkl 2002). Figure 2.10 presents an example which illustrated a sequence of faded worked examples. The example is adapted from Kinnear (2019) which is used to introduce the procedure for computing terms of Maclaurin series in Fundamentals of Algebra and Calculus (FAC). The first sub-figure (a) starts with a worked example where the final step is missing, then it follows up with a sequence of STACK questions which are faded worked examples and ended up with unguided problem, with all scaffolding removed. STACK is discussed in 2.4.

Solving conventional problems yields low learning results with high cost (Van Gerven et al. 2002). Worked examples, in contrast, are more likely to lead to an efficient construction of cognitive load, because they focus the attention of learners on problem states and operators, rather than on goals and sub-goals (Van Gerven et al. 2002).

Renkl, Atkinson, and Gross (2004) reported that, at the beginning of the learning process, a learner's with low-level domain of prior knowledge is correlated with two important characteristics: 1) the learners are unable to apply specific solution procedures so, instead, they utilize general strategies for problem-solving; and 2) high intrinsic load. Hence, this strategy requires an extraneous load and, as a result, leaves little or no room for germane load such as producing self-explanation. On the other hand, when studying using a format of constrains such as faded worked examples, learners will focus more on understanding and they will be free from performance demands (Renkl, Atkinson, and Gross 2004). Solving conventional problems yields low learning results with high cost (Van Gerven et al. 2002). Worked examples, in contrast, are more likely to lead to an efficient construction of cognitive load, because they focus the attention of learners on problem states and operators, rather than on goals and sub-goals (Van Gerven et al. 2002).

Previous research suggests that it is important not to generate high extraneous load, especially when it is connected with high intrinsic load, since no cognitive capacity may remain for germane load. As Renkl, Atkinson, and Gross (2004) suggested, it is important to find ways of fostering germane load. According to Van Gerven et al. (2002), worked examples are more likely to lead to an efficient construction of cognitive load, because they focus the attention of learners on problem states and operators, rather than on goals and sub-goals.

Sweller (2006) outlined some of the cognitive principles that underlie cognitive load theory and indicate the relation between those principles and the worked example effect. It is unlikely that a single worked example per instructional area will result in a worked example effect (Sweller 2006). As a result of studying a worked example, learners need a procedure, usually a problem, to assess how well they have learned it (Sweller 2006).

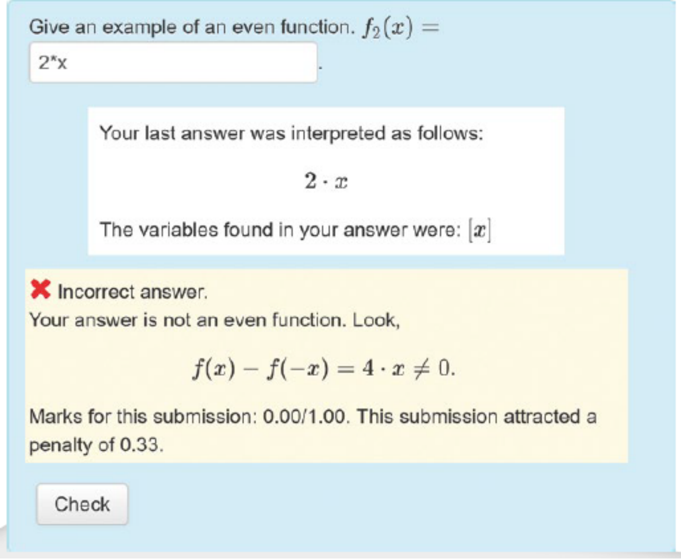
Richey and Nokes-Malach (2013) compared the effectiveness of partial or complete explanation of the solution procedures in worked examples instruction. They found that providing students with partial instruction is more beneficial than providing complete instruction guide. In partial worked examples instruction, students are encouraged for active learning while with complete instructions students have a passive role. On the other hand, Wittwer and Renkl (2010) meta analytic review of the effect of instructional explanations on example based learning suggested that providing students with full instructional explanation is essential for understanding the worked examples. They found that, providing students with full instructional worked examples yielded significant benefits for learning (Wittwer and Renkl 2010).

There are also discrepancies among researchers regarding when to use worked examples with students during learning. According to Sweller, Ayres, and Kalyuga (2011), the best way to use worked examples is to present the worked examples to students before asking them to solve a similar problem.

In contrast, it is not totally clear whether the position of the faded step is really crucial. Renkl, Atkinson, and Gross (2004) showed that the position of the faded steps did not influence learning outcomes; instead, individuals learn most about those specific principles that were faded.

2.4 STACK

This section of the thesis focuses on STACK, a System for Teaching and Assessment using a Computer-algebra Kernel, is an open source Computer Aided Assessment (CAA) system for mathematics, and other STEM subjects. The first version of STACK was developed in 2004 by Chris Sangwin in collaboration with Laura Naismith at the University of Birmingham. Since its first release, STACK has been continuously developed and is in widespread use particularly in higher education, notably by The University of Edinburgh, The Open University and Loughborough University. STACK focuses on accepting algebraic input from students. Figure 2.11 illustrates an example of STACK question with a student's



Give an example of an even function. $f_2(x) =$

2*x

Your last answer was interpreted as follows:

$$2 \cdot x$$

The variables found in your answer were: $[x]$

✗ Incorrect answer.

Your answer is not an even function. Look,

$$f(x) - f(-x) = 4 \cdot x \neq 0.$$

Marks for this submission: 0.00/1.00. This submission attracted a penalty of 0.33.

Figure 2.11: An example STACK question

response and feedback. STACK marks and provides feedback based on the mathematical properties of the student's answer.

Student answers in STACK will often need to be both algebraically equivalent to the correct answer and in the appropriate format. Equivalent expressions are expressions that work the same even though they look different. If two algebraic expressions are equivalent, then the two expressions have the same value when we plug in the same value(s) for the variable(s).The property of an answer does not have to be unique to be correct, and STACK uses these properties to test a student's answer objectively. STACK's feedback may include calculations that are directly linked to the answer a student entered. Figure 2.11 provides an example of this type of feedback.

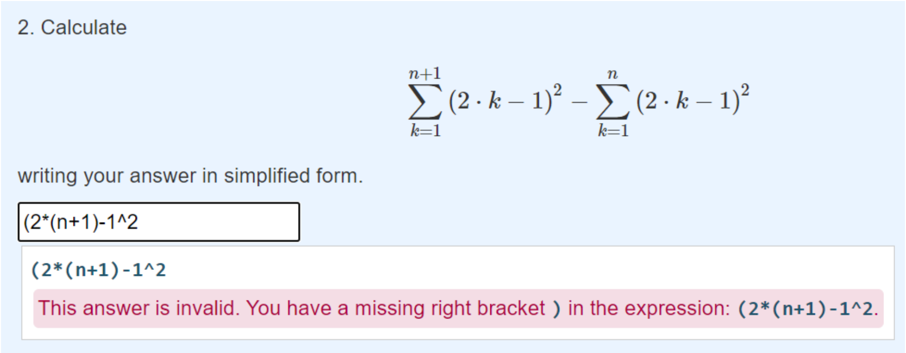
STACK uses the computer-algebra system (CAS) Maxima to establish the properties of the student's answers and provide feedback. Tests of correctness are based on establishing algebraic equivalence between student and teacher answers, but this can only be achieved when there is only one correct answer. A student's answer can be assessed using more than one answer test through STACK's library of answer tests. The STACK question type is now available in Moodle, a sophisticated online learning platform with advanced reporting capabilities.

The motivation for the work documented here was the desire to introduce my work in Chapter 7 to firstly illustrates how engagement with the lecture quiz related to success on the weekly assessed quiz, and the course total. Second, to explore the common mistakes made by students when using online materials (i.e., STACK questions) to prepare for mathematical induction? Secondary to the above research goals, this chapter illustrates how to use research to update STACK potential response trees in questions written to support learning mathematical induction based on separated concerns. Separated concerns is a phrase used to describe materials in which potential misconceptions are addressed directly. Assessing a full mathematical proof is currently well beyond the capabilities of computer systems, but one possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. In particular, we want to be able to develop online assessments by transforming existing (largely paper-based) problem sets into online assessments.

I begin by discussing the way in which STACK determines the correctness of an answer in general.

2.4.1 Validity and correctness

To reduce the problem of students being penalised on a technicality, the student's response is displayed immediately in the form of *validation feedback*. Separating out *validity* from *correctness* is a core part of



2. Calculate

$$\sum_{k=1}^{n+1}(2\cdot k-1)^2-\sum_{k=1}^{n}(2\cdot k-1)^2$$

writing your answer in simplified form.

```
(2*(n+1)-1^2
```

```
(2*(n+1)-1^2
```

This answer is invalid. You have a missing right bracket `)` in the expression: `(2*(n+1)-1^2`.

Figure 2.12: An example of STACK question with an invalid student's answer

the design of STACK. The validation feedback is (almost) always shown, and when needed will provide feedback about missed brackets, and other validation issues. E.g. when a student misses out a bracket it becomes impossible to make sense of the student's answer, and immediate feedback is required before any marking algorithm is applied. An example of this kind of feedback is illustrated in Figure 2.12.

Once the student has valid expressions then the system can decide whether the answer is *correct*. The difference between validity and correctness is somewhat subtle. Some issues which render a student's answer *invalid* are clear, and do not vary between questions. In other situations the teacher has choices about validity.

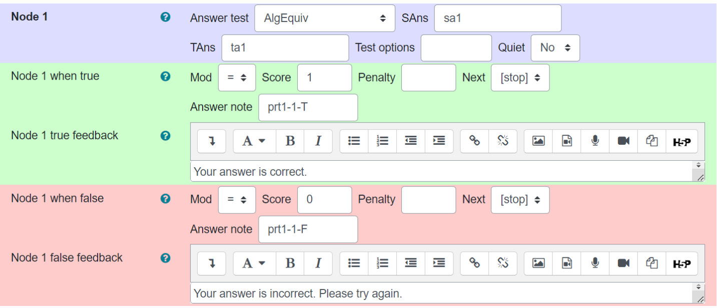
In other situations, where students need to type in an equation, a validation feedback could be generated to ensure the student has an equation before it is assessed. These validation choices have two effects: (i) they prevent students being penalised on a technicality and (ii) it significantly increases the reliability of the marking algorithm. Mostly, the validation does not vary between questions but the teacher does still have some choices to make.

STACK runs the required validity tests on student input, and if their answer passes all the tests, it is presented to the student as typeset mathematics. In this way, the student can verify that they have entered exactly what they intended - as their answer are submitted via Maxima syntax, but presented back to them as typeset mathematics - and that any uncertainties have been resolved. Students may change their answers and submit them for validation as many times as they like without losing marks.

When a student's answer has passed the relevant validity tests, they submit it again for marking. For illustration, consider the student and teacher each having a single answer. The comparison between student and teacher answer will be done using an *answer test*. An answer test is a predicate function that returns `true` if the student's and teacher's answers are "the same" to a certain extent and `false` if they are not. The prototype test compares two answers— `SA` from the student and `TA` from the teacher — to see if they are mathematically equivalent, i.e.,

```
    simplify(SA - TA) = 0.
```

Only mathematical expression without equal signs are expected for this test. Answer tests yield two more bits of data in addition to `true` or `false`. The first is *feedback*, which the teacher may choose to show the pupil. The second is a *note*, which captures the properties of the student's answer that the system has identified. The note is saved in the system in case the teacher wants to spot patterns in students' answers. Therefore, in addition to determining whether or not a student's answer is correct, it is desirable if an answer test can provide useful information about an answer that is incorrect, just like a teacher may when grading work.



Node 1

Answer test: AlgEquiv SAns: sa1

TAns: ta1 Test options: Quiet: No

Node 1 when true

Mod: = Score: 1 Penalty: Next: [stop]

Answer note: prt1-1-T

Node 1 true feedback

Your answer is correct.

Node 1 when false

Mod: = Score: 0 Penalty: Next: [stop]

Answer note: prt1-1-F

Node 1 false feedback

Your answer is incorrect. Please try again.

Figure 2.13: Potential Response Tree node

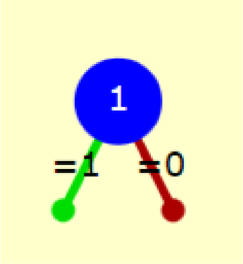


Figure 2.14: Potential Response Tree

2.4.2 STACK potential response tree (PRT)

Each question in STACK may contain a number of part-questions, each of which requires the student to provide an answer. These part-questions are all on the same page and, while mathematically connected to one another, are distinct from one another in terms of how the system views them. Each part-questions will contain a box for students' input ,which we refer to as *interaction elements*. The answer will be processed via STACK Potential Response Tree after validation and resubmitted.

A *potential response tree* consists of an arbitrary number of linked nodes which are called potential response nodes. While the word *tree* is used, strictly speaking we have an acyclic directed graph of potential response nodes. In each node two expressions are compared using a specified answer test, and the result is either true or false. The outcome of this answer test determines what happens next. Each true/false branch of the potential response node can (i) assign/modify the numerical mark, (ii) add feedback for the student, (iii) record an *answer note* and (iv) determine whether any further potential response nodes should be executed next or the assessment process should stop at this point. The directed graph is acyclic preventing an infinite assessment algorithm loop.

A teacher can use a potential response tree to establish separate properties of an answer by comparing it with different tests or a range of possible correct answers.

In the PRT, each node includes the student's expression SA, the teacher's expression TA, and the required answer test.

As a result of applying the answer test to SA and TA, the result determines whether SA is sent to another node of the answer test, or if the outcome is to end the test. Each of these options includes a per-node option for providing feedback or adding notes.

Figure 2.13 shows a node of a PRT as it appears to someone authoring a question. The student's answer is a variable, `sa1`, while the teacher's answer is a variable, `ta1` (it too could be entered directly, or as a Maxima function). The answer test being applied is *algebraic equivalence* `AlgEquiv`. Additional feedback and notes can be provided. This feedback can be suppressed using the `Quiet` option in the PRT node. Recall, `prt1-1-T` is read as "potential response tree 1, node 1, returned true." and `prt1-1-F` is read as "potential response tree 1, node 1, returned false."

The feature of the tree-based approach is as that, by creating one answer test to compare `SA` with `TA`, and provide specific feedback if the answer test returns `true`. If the answer test returns false, either the student's answer is correct or they have made some other error in their answer. The answer test returning `false` will lead to another node of the PRT which compares SA with the correct TA. The corresponding potential response tree for Node1 is illustrated in 2.14. As shown in the PRT, two different situations are branched from Node1 as follow:

1. If the answer test returning true (prt1-1-T): student's answer is correct (1 mark), and if so stop.
2. If the answer test returning false (prt1-1-F): students' answer is wrong, feedback will be provide.

A potential response tree gives a teacher the ability to compare a student's answer to a variety of correct answers or to multiple test in order to determine the properties a student's answer. Updating the PRTs in STACK questions, based on research findings, is discussed in Chapter 7.

2.5 Summary

Before moving on to describe the unit coding scheme that has been applied to code students' responses in two studies in this thesis. A list of the main points arising from the literature review are provided for the reader.

- Some mathematics educators attribute students' difficulties in understanding proofs to the writing style of the traditional paragraph proof (Rowland 2002). Consequently, alternative formats of writing proofs were suggested by several researchers.
- In instructional settings, which may be essential for novices, may have negative consequences for more experienced learners (Kalyuga, Ayres, et al. 2003).
- No study had previously investigated the potential format effect with two-column proofs at the undergraduate level. In this thesis, two-column is employed in Study 1 to determine the effects of two-column format on writing mathematical arguments.
- Human marking is needed for typed responses in which students have to show their work in reasoning by expressing their work in longer argumentation and derivations, and this provided my motivation to study the mechanism of typing and compare the format effect of typing vs. uploading photographed mathematical responses. The study is presented in Chapter 5.
- As students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). One possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. In this case, a format such as faded worked example is effective to design mathematical online task. Faded worked examples heavily influence the idea of *seperated concerns* which are introduced in Chapter 6. The phrase separated concerns is used to describe materials in which potential misconceptions are addressed directly.

Chapter 3

Unit Coding Scheme

Proofs are mathematical arguments and being able to understand the logic that supports the proof is essential to understand the structure of mathematical arguments. In order to understand the structure of a mathematical argument, it is crucial to study how to analyse mathematical arguments. Therefore, this chapter starts with a review of the literature on logical reasoning and argumentation, based mainly on the works of Toulmin (1958) and Aberdein (2005).

In this chapter, I introduce a “unit coding scheme” and explain how the work of (Toulmin 1958), (Hodds 2014) and (R. J. Back, Mannila, and Wallin 2010) were integrated to create the scheme. Toulmin’s argumentation scheme provided a model for only one simple argument, therefore (R. J. Back, Mannila, and Wallin 2010) has also been used to model structured derivations for mathematical proofs. The coding scheme of Hodds (2014) was adapted to code the written justification generated when writing mathematical arguments. I then defined what I mean by a “proof unit”. I provided an example that illustrates how to apply the unit coding scheme.

The unit coding scheme has been applied to code students’ responses in two studies in this thesis. The first study investigated a potential format effect with two-column proofs which is discussed in Chapter 4. The second study is to investigate the format effects of typing vs handwritten mathematical responses which is presented in Chapter 5.

This chapter presents the unit coding scheme as a significant methodological contribution, offering an approach to assessing written work in online environments by evaluating both the quantity and quality of student explanations and proof construction. This innovative tool has the potential to revolutionize online assessment practices by providing deeper insights into student understanding and fostering more effective learning experiences. Further research could explore the scheme’s applicability to additional subjects and its effectiveness in promoting deeper learning strategies.

3.0.1 Toulmin’s Model of Argument

Among mathematics educators, Toulmin’s scheme has become a popular tool for analysing the structure of arguments. Toulmin’s scheme was developed in the 1950s by Stephen Toulmin (Toulmin 1958), but started to take place in mathematics education in the 1990s e.g. by (Krummheuer 1995). It has since been used to analyse arguments in primary schools (Evens and Houssart 2004), secondary schools (Arzarello and Sabena 2011), undergraduate mathematics (Rasmussen and Stephen 2007) and also for arguments from postgraduate mathematicians (Inglis, Mejia-Ramos, and Simpson 2007).

The simplest layout of Toulmin’s model is shown in Figure 3.1, an argument starts from Data (D), ends with a Conclusion (C), and requires a Warrant (W) which connects the data with the conclusion.

Toulmin describes warrants as “...hypothetical statements, which can act as bridges, and authorise the sort of step to which our particular argument commits us”. The warrant may have a further supporter called the backing (B) and the qualifier (Q) provides the degree of confidence we have about the conclu-

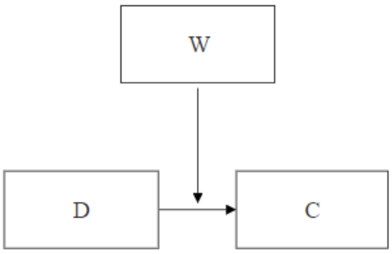


Figure 3.1: Toulmin's argumentation scheme

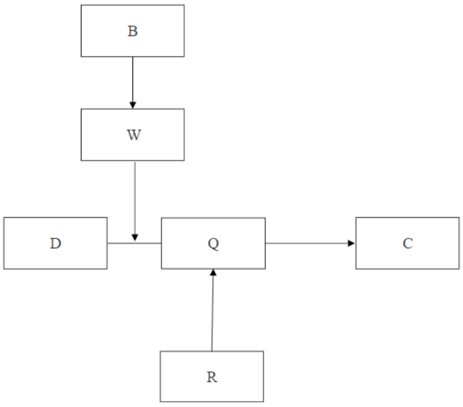


Figure 3.2: Toulmin's enhanced scheme for a general argument. The argument would read 'D, and since W (given B) we can Q conclude C, unless R

sion. Hence the full framework may be understood as 'Given that D, we can Q claim that C, since W (on account of B), unless R'. For example: 'Given that Harry was born in Bermuda, we can presumably claim that he is British, since anyone born in Bermuda will generally be British (on account of various statutes ...), unless his parents were aliens, (Toulmin 1958). This simple layout can be applied to one argument. Figure 3.2 shows a diagram of the enhanced scheme suggested by Toulmin.

While some researchers have criticised Toulmin's scheme for its unreliable definition of some of its components (Weinstein, 1990), the extensive use of the scheme indicates that it has become a useful tool for many researchers. Many authors adapt Toulmin's scheme in undertaking their analysis. In the field of mathematics education, Krummheuer (1995) started using Toulmin's scheme by analysing classroom-based mathematical arguments. However, he applied a reduced version of the original scheme by omitting the use of the rebuttal and the modal qualifier, considering them irrelevant to mathematical arguments.

Authors from other disciplines have adopted a different approach to adapting Toulmin's scheme of informal logic to formal mathematics. Aberdein (2005) retained all of Toulmin's components including modal qualifiers and rebuttals when analysing formal mathematical proofs. However, most mathematical proofs consist of a sequence of logical steps to connect the given data to the final conclusion. Accordingly, a mathematical proof consists of more steps, more warrants and a more complex structure. Aberdein (2005) developed the simple scheme of Toulmin to be applicable to more complicated mathematical proofs. Aberdein (2005) stated that, in a multi-step proof, there are usually several conclusions that clarify each step on the path to the main conclusion. Often the conclusion of one step is the data for the next step and so on. In each step there may also be a related warrant, qualifier, backing and rebuttal which explain how the writer of the proof goes from the data to the conclusion of that specified step, (Aberdein 2005). By using this developed scheme for analyzing arguments in mathematical proofs, a description of a proof can be obtained and how the parts of an arguments can fit together. Figure 3.3

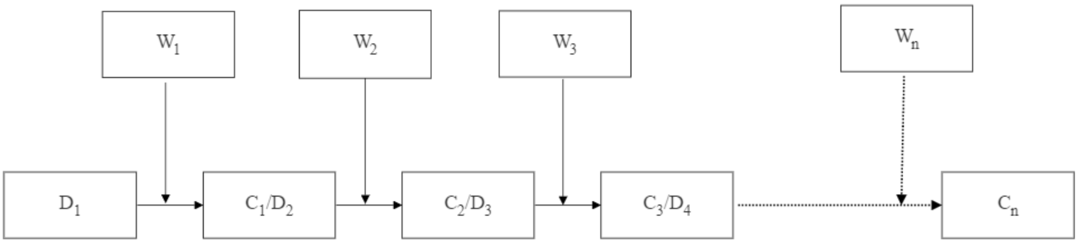
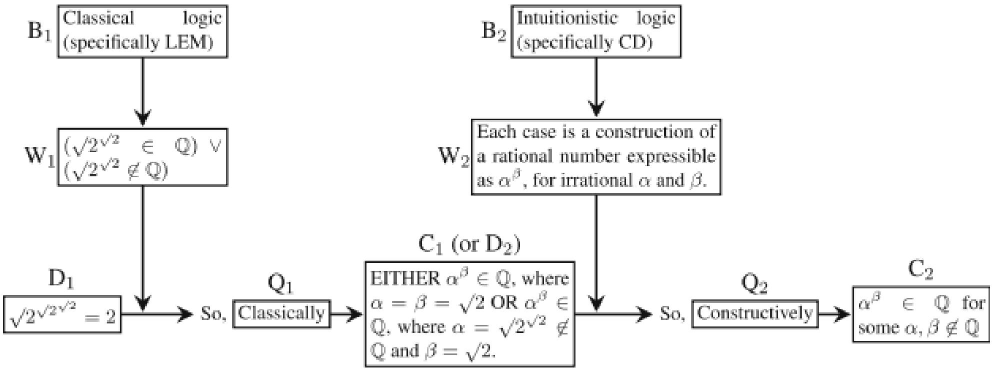


Figure 3.3: A diagram of mathematical argumentation according to Aberdein (2005)



B_1
Classical logic (specifically LEM)
B_2
Intuitionistic logic (specifically CD)
W_1
$(\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}) \vee (\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q})$
W_2
Each case is a construction of a rational number expressible as α^β, for irrational α and β.
C_1 (or D_2)
D_1
$\sqrt{2}^{\sqrt{2}^{\sqrt{2}}} = 2$
So,
Q_1
Classically
EITHER $\alpha^\beta \in \mathbb{Q}$, where $\alpha = \beta = \sqrt{2}$ OR $\alpha^\beta \in \mathbb{Q}$, where $\alpha = \sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$ and $\beta = \sqrt{2}$.
So,
Q_2
Constructively
C_2
$\alpha^\beta \in \mathbb{Q}$ for some $\alpha, \beta \notin \mathbb{Q}$

Figure 3.4: An example from Aberdein (2005) to model a Classical proof that there are irrational numbers α and β such that α^β is rational adapted from Aberdein (2005)

shows a diagram of the model suggested by Aberdein (2005).

Figure 3.4 illustrated the developed scheme of Aberdein (2005) to model a classical proof that there are irrational numbers α and β such that α^β is rational, where the rebuttal components have been omitted for simplicity (Aberdein 2005). Clearly, this decomposition of the proof into its separate steps shows how each step is dependent on the others: the first step employs the non-constructive law of excluded middle (LEM), whereas the second step relies on the constructively acceptable inference rule of constructive dilemma (CD). (Aberdein 2005) suggested that this fine-grained application of Toulmin's developed scheme can make the guilty steps explicit, unlike a simple scheme in which the qualifier for the whole proof would merely indicate that the result is classically, but not constructively, valid.

Inglis, Mejia-Ramos, and Simpson (2007) used Toulmin's scheme to analyse arguments from highly talented postgraduate mathematicians. Inglis argued that Toulmin's reduced scheme is inadequate for accurately modelling the full range of arguments constructed by postgraduate mathematics students. He found that frequent use of non-deductive warrants to derive non-absolute conclusions and highlights that these forms of components (i.e, qualifiers) are important for them in the process of solving the problem. Therefore, by using the reduced scheme, it would be impossible to model this type of argumentation. He concludes that modal qualifiers play an important and previously unrecognised role in modeling mathematical argumentation.

In mathematics, proofs are likely to follow a fixed form of argument. However, the form of arguments in the classroom play an important role in sociomathematical norms, and teachers and pupils may have to deal with shifting norms where some arguments may be allowed to be generic or heuristic while others must be formal (Simpson 2015). It is perhaps the role of teachers and the wider mathematical educational system to clarify acceptable forms of argument, such as justifications for new inferences based on old ones (Simpson 2015).

According to Simpson (2015), the expanded Toulmin's scheme provided insight into what examiners may be expecting from their students in terms of the level of explicit warrant. The highest level of warrant considered essential, as well as providing at least some warrant for main steps, but other justifications for proof steps are not as highly valued. Calculation steps which are directly in the focus of the module (e.g., those resting on commutativity) must be explicit, but those which are not important (e.g., those resting on division by a non-zero polynomial) need not. One could easily imagine a course in which a model solution would not make explicit the commutativity of multiplication, but would require students to be explicit about steps involving the division by a polynomial (and were clear that the polynomial was not zero) (Simpson 2015).

Simpson (2015) concluded that in order for students to succeed on even the smallest part of their mathematics assessment, students need not only to understand the course concepts, but also to be fluent on how to integrate those concepts with multiple basic proof steps.

Rasmussen and Stephen (2007) described a procedure for documenting the structure of elements of argumentation when analysing students' statements:

> In general, documenting the structure and function of students' augmentations is facilitated by the following rules of thumb. Claims are the easiest type of contribution to identify in an argumentation and consist of either an answer to a problem or a mathematical statement for which the student may need to provide further clarification. Data are less easy to document but usually involve the method or mathematical relationships that lead to the conclusions. Most times, warrants remain implied by the speaker and are elaborations that connect or show the implications of the data to the conclusion. Finally, a backing is identified typically by answering the question: "Why should I accept your argument (the core) as being sound mathematically?" Backings, therefore, function to give validity to the argumentation. (Rasmussen and Stephen 2007)

The Toulmin scheme of argumentation can provide an insight into the logical relationships within a proof and it has been used widely in the mathematics education literature. This model will be used throughout this thesis as part of a coding scheme to describe the logical steps in mathematical responses that generated by students. The difference between the present research and previous uses of the Toulmin model is that the research reported in this thesis uses the model to break down the components of a mathematical arguments in order to discuss the effect of writing mathematical arguments in different formats using online assessment.

3.0.2 Hodds' Coding Scheme

Hodds (2014) investigated the content and quality of self-explanations generated by students after reading a mathematical proof. The work of Hodds (2014) provided a qualitative method of gaining insight into students' possible self-explanations for a proof. In order to judge and analyse students' self-explanations, Hodds (2014) developed a coding scheme, in which justifications were classified into either explanation or non-explanation categories. The coding scheme was first proposed by Renkl (1997) to analyse the thinking aloud protocols while studying worked-out examples. Later, the scheme altered slightly by Ainsworth and Burcham (2007) to explore the roles of self-explanation and text coherence for novices learning relatively complex material in biology. The only significant differences between Renkl (1997) and Ainsworth and Burcham (2007) are that the category of anticipative reasoning used by (Renkl 1997) was not felt appropriate and hence removed, whilst false self-explanation was added. Later, Hodds (2014) adapted the version of Ainsworth and Burcham (2007) and slightly altered to be applicable on coding verbal protocols for mathematical proof rather than biology. Hodds (2014) assigned each category as either an explanation or non-explanation as follows:

Explanation Categories

1. Principle-Based Explanation: this category was scored if a participant gave any explanation that was derived from definitions, theorems not explicitly written in the proof. For example, if a participant said: "...this is because by the definition of..."

2. Goal Driven Explanation: a positive explanation was coded as goal-driven if a participant gave a statement that associated with the proof structure. For example, if a participant said: “OK, we’re doing this because we are going to use it later on in the proof.”

3. Noticing Explanation: this category was scored when a participant provided any explanation that linked to a previous idea used in the proof. For example, if a participant said: “...this is because in line 5 we introduced...”

Non-Explanation Categories

1. False Explanation: this category was scored when a participant provided an incorrect explanation.

2. Paraphrasing: when a participant repeated the line or part of the line using similar or the same words as used in the line.

3. Positive Monitoring: This category was scored if a participant stated “I understand this”, “OK, this makes sense” or words to that effect.

4. Negative Monitoring: This category was scored if a participant said “I don’t understand this”, “How is this true?” or words to that effect.

The work of Hodds (2014) provided a qualitative method of gaining insight into students’ possible justification for a proof. There was one main difference between using the justification categories in this thesis and the study of Hodds (2014). The major difference in this thesis is instead of using the categories to code verbal justification produced by students in self-explanation training, I use the categories to code written justification produced by students to discuss the potential format effect of writing mathematical arguments using online assessment.

3.0.3 Back’s structured derivations

Since (Toulmin 1958) provided us with a framework that can be used to analyse one simple argument in general. The work of (R. J. Back 2010; R. J. Back, Mannila, and Wallin 2010) provided a specific format when modeling mathematical derivations in proofs which was missing in (Toulmin 1958).

The structured derivation is a further development of the *calculational proof* style developed by Dijkstra and his colleagues (R. J. Back 2016). The goal of Dijekstra and his colleague was to carry out mathematical proofs and derivations in the same way as in traditional calculations, like when simplifying expressions, solving equations or finding values of functions. The calculational proof introduced the idea of explicit justifications on separate lines. The style of calculational proof has been adapted widley in articles and textbooks for programming in general. Dijkstra’s calculational proof has been the main inspiration for the development of structured derivation’s format.

Structured derivations are an alternative proof format, introduced by (R. Back, Grundy, and Von Wright 1997), as a method to present proofs in programming logic. Later, they adapted the method to give what they claim is a practical approach for presenting proofs and derivations in a simple readable and well-structured format in high school mathematics. In particular, structured derivations include more formal presentation, nested sub-derivations and inherited assumptions. Structured derivation is a more formal proof format for presenting mathematical arguments. The format is based on a nested and hierarchical view of derivations, where a main derivation can be divided into a number of more detailed sub-derivations. According to R. J. Back, Mannila, and Wallin (2010), a calculation is essentially a relation chain of this form:

$$term_0 \ rel_1 \ term_1 \ rel_2 \ term_2 \ ... \ term_{k-1} \ rel_k \ term_k$$

where $term_0$, $...term_k$ are terms and rel_0, $...rel_k$ are relations between terms. The general structured derivation for a calculation has the following syntax with $(k \geq 1)$:

derivation::=

$$\begin{array}{ll} ø & term_0 \\ rel_1 & \{justification\}_1 \\ & term_1 \\ & \vdots \\ & term_{k-1} \\ rel_k & \{justification\}_k \\ & term_k \\ \square & \end{array}$$

The $justification$ explains why the relationship (i.e. rel_i) holds between the terms (i.e. $term_i$). The $justification$ is given in curly brackets. In (R. J. Back 2016), there are two different cases of justifying a calculation step. In the first case, the justification states which mathematical rule is used for a calculation step:

$(x+1)(x+y)$
$\equiv$ {the distribution law for polynomials}
$(x+1)x+(x+1)y$

The second case, the justification states which operation is applied for a calculation step:

$(x+1)(x+y)$
$\equiv$ {distribute the first term across the second term}
$(x+1)x+(x+1)y$

Both cases in justifying a step are useful but they both have different features. For the first case, the justification can be seen as an observation for why the equality holds deductive. For the second case, the justification clarified why the transformation is acceptable (R. J. Back 2016). This is an instruction to perform some calculation.

According to R. J. Back, Mannila, and Wallin (2010), structured derivations provide several new features compared to the traditional way of presenting mathematical proofs: the use of logical notation, sub-derivations and justifications. The following example was adapted from R. J. Back, Mannila, and Wallin (2010) which demonstrates both the format for calculations and the use of explicit logical notation in mathematical derivations. The problem is to solve the following equation in the domain of real numbers:

$$x^3 - x^2 + x - 1 = 0 \tag{3.1}$$

So transitivity now gives us that

$$\begin{array}{ll} & x^3 - x^2 + x - 1 = 0 \\ \equiv & \\ & x = 1 \end{array}$$

The eMath project 2011 - 2016 was to pilot the structured derivation approach to mathematics education in high schools, and build a digital platform for mathematics education in general. The Virum project 2014 - 2017 was to develop a digital platform for virtual education. The research in these projects was carried out at Abo Akademi University and University of Turku. eMathStudio is a learning platform for assessments and interactions in mathematics. eMathStudio provides a digital math notebook allows students to explain why something is allowed and what the intentions behind the calculation steps are. eMathstudio also provides eMath checker which used to check the correctness of every step of your own

$x^3 - x^2 + x - 1 = 0$

$\equiv$ {grouping}

$(x^3 - x^2) + (x - 1) = 0$

$\equiv$ {factorization}

$x^2(x - 1) + (x - 1) = 0$

$\equiv$ {factorization}

$(x - 1)(x^2 + 1) = 0$

$\equiv$ {zero product rule}

$x - 1 = 0 \vee x^2 + 1 = 0$

$\equiv$ {add -1 to both sides in right disjunct and simplify}

$x - 1 = 0 \vee x^2 = -1$

$\equiv$ {a square is never negative}

$x - 1 = 0 \vee F$

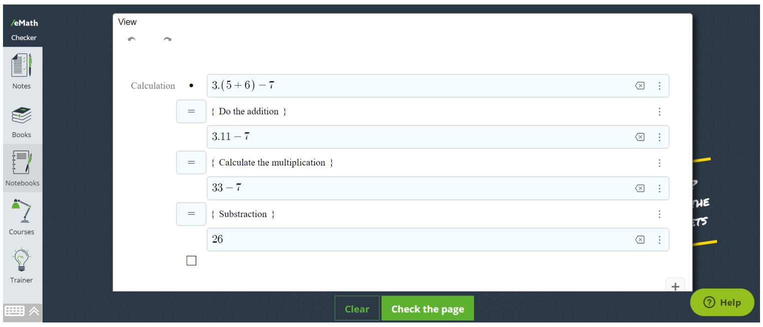
$\equiv$ {disjunction rule: $p \vee F \equiv p$}

$x = 1$

□

Figure 3.5: Solution for Equation. 3.1 using structured derivation's format

calculation. There is therefore enough space between each calculation step to explain the steps verbally. A scrrenshot of eMATH checker is illustrated in 3.6. The checker formulates each step in the derivation as a mathematical theorem, and sends the theorem to an automatic theorem prover that runs in the cloud. The derivation is correct, if each step in the derivation is proved correct. The checker warns for each derivation step that it was not able to prove correct. Mathematical symbols are entered by selecting them from a palette (see Figure 3.7), or by writing simple shorthand keywords.



Calculation • $3.(5+6)-7$

$=$ { Do the addition }

$3.11-7$

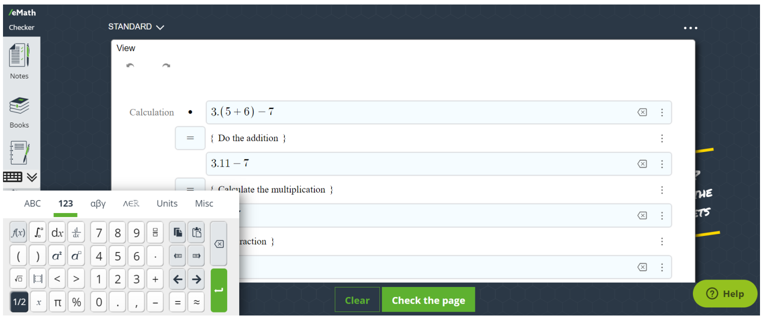
$=$ { Calculate the multiplication }

$33-7$

$=$ { Substraction }

26

Figure 3.6: eMath checker in eMathStudio



Calculation • $3.(5+6)-7$

$=$ { Do the addition }

$3.11-7$

$=$ { Calculate the multiplication }

raction }

Figure 3.7: Selecting math symbols from a palette in eMath checker

A large number of pilot studies were also conducted on the use of structured derivations in high schools and introductory mathematics courses in universities in Finland, Sweden, and Estonia. The results indicated that both students and teachers appreciated the method. The students found the method as different but not difficult. The clarity of derivations and proofs is increased, making the derivations easier to read, check and correct, for both students and teachers (R. J. Back, Mannila, and Wallin 2010). The main drawbacks of structured derivations as mentioned by students were related to length and time (R. J. Back, Mannila, and Wallin 2010). As the explicit justifications naturally increase the length of a proof and also take some time to be written. However, this is considered an advantage. Since justifying each step will let students think carefully about the solution. Structured derivations also have the potential to increase students' self-perceived level of understanding. R. J. Back, Mannila, and Wallin (2010) suggested that the use of sub-derivations makes the format suitable for new types of assignments and self-study material. Moreover, a familiar format can enhance students' confidence, giving students belief in their own skills to solve a problem. This can be especially important when considering the typical 'fear' for proof found among learners. When proofs and simple calculational derivations are written using the structured derivations, students may feel less intimidated by proofs. Structured derivations introduce many new features compared to the traditional way in presenting mathematics: a greater use of logical notation, sub-derivations and justifications. As each step is justified, the final product contains a documentation of the thinking that the student or the teacher was engaged in while writing the derivation. Hence, the format of structured-derivations and the justifications facilitate presentation of proofs and derivations in class.

Using structured derivations, the standardized format gives students a fixed model for how proofs and solutions can be written. The format also has potential to make the presentation of mathematics more consistent in the classroom and in text. This format will be used throughout this thesis as part of a coding scheme to describe the calculation steps in mathematical responses that are generated by students to discuss the effect of writing mathematical arguments in different formats when using online assessment.

3.1 A proof unit

It has been discussed in Section 3.0.1 that a typical mathematical arguments consist of many steps and contingent parts, which cannot be adequately described by Toulmin's model. This has already been acknowledged by (Aberdein 2005), who suggests that a proof will usually have several sub-conclusions that clarify each step on the path to the main conclusion. Aberdein's steps can overlap, i.e. often the conclusion of one step is the data for the next step and so on. However, Aberdein's scheme is essentially linear. As a result, I have been influenced by the work of (R. J. Back 2010; R. J. Back, Mannila, and Wallin 2010) who argue that writers should be more specific when writing mathematical derivations and proofs.

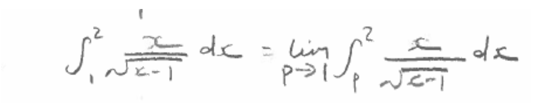
To analyse the structure of the argument, I firstly introduced the idea of a "proof unit." My view of proof is based on a nested and hierarchical view of derivations, where a main derivation can be divided into a number of more detailed internal sub-derivations, in a recursive manner. Those sub-derivations are treated as the units of the proof. A *proof unit* is defined as the the smallest section to which Toulmin's model can be applied. That is, each unit (U) will normally contain data (D), conclusion (C) and may have explicit warrant (W). Units can consist of algebraic expressions or written words or both. Indeed, I consider algebraic expressions to be an integral part of a complete mathematical sentence.

Framed in terms of R. J. Back, Mannila, and Wallin (2010) nested arguments, the unit of a derivation, $Unit_i$, is defined as:

$$\begin{array}{ll} & term_{i-1} \\ rel_i & \{justification\}_i \\ & term_i \end{array}$$

$term_{i-1}$ is considered as a data for $Unit_i$, and $term_i$ as a conclusion for $Unit_i$. The $justification_i$ would be classed as a warrant for $Unit_i$.

The idea of a proof unit is illustrated by Figure 3.8 taken from one of students' responses who participated in Study 1:



$$\int_1^2 \frac{x}{\sqrt{x-1}}\,dx = \lim_{p\to 1}\int_p^2 \frac{x}{\sqrt{x-1}}\,dx$$

Figure 3.8: An example for a Unit in the coding scheme

Based on Toulmin's model, the left hand side is the data and the right hand side is the conclusion with an implicit principle warrant based on the definition of the improper integral.

Since mathematical arguments might contain an implicit or explicit warrant that explains the relationship between data and conclusions, it was meaningful to consider the coding scheme of (Hodds 2014) to classified the written warrant. Only four of seven categories used by Hodds (2014) were adapted in the unit coding scheme. The four categories were principle-based explanation, goal-driven explanation (an explanation that inferred a goal to a particular structure or sentence ((Ainsworth and Burcham 2007)), noticing explanation (noticing any connections between previous and current lines), and paraphrasing. Note that the "false explanation", "negative monitoring" and "positive monitoring" categories refer to spoken justifications and so are not relevant to the written justifications.

In conversation, a single unit might be unpacked. For example, finding the partial fraction form of $\frac{1}{x^2-1}$ might be written as a single unit:

$$\underbrace{\frac{1}{x^2-1} = \frac{A}{x-1} + \frac{B}{x+1}}_{Unit_a}$$

Since $x^2 - 1$ is the difference between the two squares, $x^2 - 1 = (x-1)(x+1)$ which is the product of two linear factors. When we have distinct linear factors, the form of partial fraction is $\frac{A}{x-1} + \frac{B}{x+1}$.

So if $Unit_a$ is written with explicit justification using the definition of unit then it will be as follows:

∅ $\frac{1}{x^2-1}$

$\equiv$ {Since $x^2 - 1$ is the difference between the two squares, $x^2 - 1 = (x-1)(x+1)$ which is the product of two linear factors. When we have distinct linear factors, the form of partial fraction is $\frac{A}{x-1} + \frac{B}{x+1}$}

$\frac{A}{x-1} + \frac{B}{x+1}$

□

The level of details in a justification, and the number of steps may also vary, depend on the target audience (R. J. Back 2016). If the reader is an experienced mathematician then brief steps and justification will be enough. However, if the purpose of the derivation is to explain how to use a certain rule or a new concept, then explicit justifications and more detailed steps would be helpful. As that might help interested readers to check every step of the proof directly without trivial mistakes or without doing the complected calculations on paper or mentally (R. J. Back 2016).

3.2 The unit coding scheme

The unit coding scheme divides up a larger argument into smaller self-contained "units". A *unit* is the the smallest section to which Toulmin's Scheme can be applied. Each unit has data (D), a conclusion (C) and a warrant (W) justifying why the conclusion follows from the data. The number of units will tell us how many steps students write to achieve their answer.

- Data (D): typically a statement is coded as data if it directly followed 'consider', 'if', 'let' or if it is mentioned as an obvious fact/hypothesis at the start of this unit without support.
- Conclusion (C): a statement is coded as a conclusion if it followed 'then' or 'therefore' or when it stated as a result of a calculation from previous data.
- Warrant (W): an explanation is coded as a warrant when it is used to connect data to conclusion in a way that explains how the data supported the conclusion. There are four categories for warrants:
  1. Principle Based Explanation (W_P): when participants provide any explanation based on definitions, theorems, rules not explicitly mentioned in the proof. For example, when a student wrote "This is because by the definition of ..."
  2. Goal-Driven Explanation (W_G): when a participant gave an explanation that related to the structure of the proof (how it is used to reach the goal of the unit or wider proof). For example, student wrote "We use .... to evaluate ..."
  3. Noticing Explanation (W_N): when a participant gave explanation that linked to a previous idea used in the proof. For example, "... this is because in line 5 we used..."
  4. Paraphrasing (W_R): E.g. repeating or paraphrasing a calculation in words that has just been done in algebra. For example, when a student wrote "...separating, simplifying ..." .

To use this coding scheme, identify the first data in the proof, and draw a circle around the data. The first data is usually information given in the question that students start with. Label it D1, and draw a circle around the first conclusion resulting from that data and call it C1 and so on.

The second conclusion is usually coded as the third data, (i.e C_2/D_3) as we did not reach the final answer yet. The final answer will be coded as C_n, where n is the number of the last conclusion in the proof. Then, we can say that there are n Units in the proof.

Example

Example 1. Explain why the following integral is improper and determine whether it converges or diverges.

$$\int\limits_{-\infty}^{0} xe^{x}\mathrm{d}x$$

Solution:

By definition of improper integral,

$$\int\limits_{-\infty}^{0} xe^{x}\mathrm{d}x = \lim_{t\to-\infty}\int\limits_{t}^{0} xe^{x}\mathrm{d}x$$

We integrate by parts with $u = x$, $\mathrm{d}v = e^{x}$, so that $\mathrm{d}u = \mathrm{d}x$, $v = e^{x}$

$$\int\limits_{t}^{0} xe^{x}\mathrm{d}x = (-te^{t} - 1 + e^{t})$$

We know that $e^{t} \to 0$ as $t \to -\infty$, and by l'Hospital's Rule,

$$\begin{aligned}\lim_{t\to-\infty} te^{t} &= \lim_{t\to-\infty}\frac{1}{-e^{-t}}\\ &= 0\end{aligned}$$

Therefore,

$$\begin{aligned}\int\limits_{-\infty}^{0} xe^{x}\mathrm{d}x &= \lim_{t\to-\infty}(-0 - 1 + 0)\\ &= -1\end{aligned}$$

Thus the given improper integral is convergent. □

The example after applying the unit coding scheme

U_1:

(W_{P_1}) By definition of improper integral,

$$\underbrace{\boxed{\int\limits_{-\infty}^{0} xe^{x}\mathrm{d}x}}_{D_1} = \underbrace{\boxed{\lim_{t\to-\infty}\int\limits_{t}^{0} xe^{x}\mathrm{d}x}}_{C_1}$$

U_2:

(W_{P_2}) We integrate by parts with $u = x$, $\mathrm{d}v = e^x$, so that $\mathrm{d}u = \mathrm{d}x$, $v = e^x$

$$\underbrace{\boxed{\int_t^0 xe^x \mathrm{d}x}}_{D_2} = \underbrace{\boxed{(-te^t - 1 + e^t)}}_{C_2}$$

$U_3 + U_4$:
(W_{P_3}) We know that $e^t \to 0$ as $t \to -\infty$, and by l'Hospital's Rule,

$$\underbrace{\boxed{\lim_{t\to-\infty} te^t}}_{D_3} = \underbrace{\boxed{\lim_{t\to-\infty} \frac{1}{-e^{-t}}}}_{C_3/D_4}$$
$$= \underbrace{\boxed{0}}_{C_4}$$

$U_5 + U_6$:
Therefore,

$$\underbrace{\boxed{\int_{-\infty}^0 xe^x \mathrm{d}x}}_{D_5} = \underbrace{\boxed{\lim_{t\to-\infty} (-0 - 1 + 0)}}_{C_5/D_6}$$
$$= \underbrace{\boxed{-1}}_{C_6/D_7}$$

U_7:
(C_7) Thus the given improper integral is convergent. □

To code the example, I started by identifying the first data in the solution. The first data is usually information given in the question to start with. I highlighted the first data and called it D_1, Then, I highlighted the first conclusion resulting from that data and called it C_1 and so on. The statement "By definition of improper integral" was considered as a principle based warrant W_{P_1} to connect the first data D_1 with it's conclusion C_1. Therefore, D_1, C_1 and W_{P_1} were all parts of the first unit in the solution which called U_1.

For the second unit U_2, the statement "We integrate by parts with $u = x$, $\mathrm{d}v = e^x$, so that $\mathrm{d}u = \mathrm{d}x$, $v = e^x$ " was coded as a second principle based warrant W_{P_2} which used to connect the second data D_2 with it's conclusion C_2.

In the third unit U_3, the statement "We know that $e^t \to 0$ as $t \to -\infty$, and by l'Hospital's Rule" coded as a third principle based warrant W_{P_3} which used for connection D_3 to C_3.

The third conclusion C_3 was also considered as data for the fourth unit D_4, so it can be written as C_3/D_4. Similarly, the conclusion for the fifth unit is the data for the sixth unit (i.e, C_5/D_6) and so on.

The statement "Thus the given improper integral is convergent." was coded as a conclusion C_7 for the last unit U_7. In this example, there were seven units, since there were 7 conclusions. According to the coded example, there are seven units written, and three justification coded as principle based warrants.

In this thesis, the unit coding scheme is basically used as a tool for analysing a single argument in students' responses. Note that the unit coding scheme provides a qualitative method of gaining insight into how many units are typed and what kind of justification are written. Although students' arguments are varied and might be wrong. For example, some students might generate wrong warrants to connect their data to the conclusions, or they might give warrants that do not match the calculation steps.

The importance of the unit coding scheme is not only considering the number of units but also the quality and relevance of the statement within those units. In the following two studies, it was shown that it is specific types of explanations that are associated with subsequent marks and high number of units. The feature of the unit coding scheme is to consider the units as a measure of the depth or quality of students' proof construction which was not provided in other schemes. The unit coding scheme can be applied into many subjects and topic especially when using online assessment. The unit coding scheme can be used as an electronic tool to support online assessment systems.

This thesis introduces the unit coding scheme, a novel methodological contribution for assessing written work, particularly in online assessments. Studies conducted within this thesis demonstrate that the unit coding scheme effectively identifies specific types of explanations associated with higher marks and deeper understanding of the subject matter.

The unit coding scheme differentiates itself by directly measuring the depth and quality of students' proof construction through the concept of "units." This capability does not exist in other marking schemes. While these other schemes hold individual value, the unit coding scheme is used as a proxy for assessing proof quality in subsequent studies within this thesis. The scheme's flexible design allows it to be adapted to diverse disciplines, offering a significant advantage in online assessment environments.

In summary, this research presents the unit coding scheme as a significant methodological contribution, offering a nuanced approach to assessing written work in online environments by evaluating both the quantity and quality of student explanations and proof construction.

Chapter 4

Study 1: Investigating a potential format effect with two-column proofs

This chapter presents the first of three studies focused on a potential format effect when writing mathematics online. The study aims to explore the impact of the two-column format in writing simple mathematical arguments. In this study, I focus exclusively on the relationship between formats and types of justifications given by students. I undertook an experiment to compare students' responses between traditional arguments and arguments in a two-column format. Four experimental groups resulted, which allowed a comparison to be made between the effects of formats and type of justifications when writing proofs. That is to say, a structured method of presenting a mathematical proof or argument by using a tabular layout with two-columns.

Indeed, this experiment seeks to consciously constrain the format. Proof in this experiment uses calculations, rather that the more abstract proof which our university students will eventually learn, this sitting in a middle ground between proof at high school and advanced mathematical proof. A secondary outcome was the use and further development of the coding scheme which I provided in Section 3.2.

This study is based on data collected in collaboration with Prof. Chris Sangwin. Students' responses were collected from a university mathematics course run at the University of Edinburgh in 2018/19. As part of my doctoral research, I independently conducted all analysis presented in this chapter. The work involves a quantitative analysis of the available marks, and a qualitative analysis of students' justifications. This study is published on Alarfaj and Sangwin (2021).

4.1 Methods

4.1.1 The purpose of the study

The purpose of this study was to explore the potential format effect when writing mathematical arguments using two-column format. I considered the following two research questions.

1. Is there a format effect between traditional proof and the two-column format in writing mathematical arguments?

2. What kinds of justifications do the students actually write?

In this study, I chose handwriting to gather the data because some understanding of what students do in practice will be helpful to the future design of online assessments.

It was expected to confirm the well-established findings that novices benefit from guided instruction (Kalyuga, Rikers, and Paas 2012). When instructional guidance is provided to learners who already have a sufficient knowledge for dealing with presented information, an unnecessary extraneous cognitive load could be imposed either on novices or experts, resulting in an expertise reversal effect. See Section 2.2.1 on the expertise reversal effect.

4.1.2 Procedure

Participants were 80 first year undergraduate mathematics students at The University of Edinburgh who were enrolled in a calculus course, *Calculus and its Applications* (CAP) in 2018/19. See Appendix C for more details.

Participants were randomly assigned into four groups (T1, C1, T2 and C2), where T is used for "Traditional" and C for "Two-Column". The four groups arise because, for equity reasons, I want to make sure each student writes in both formats (Traditional/Two-Column) and that I have materials which contain the two proofs in each format. For that reason I used two questions namely Q1. and Q2. The four groups were designed as follow:

T1 : Answering Q1 in traditional way, then answering Q2 in two-column format.

C1 : Answering Q1 in two-column, then answering Q2 in traditional way.

T2 : Answering Q2 in traditional way, then answering Q1 in two-column format.

C2 : Answering Q2 in two-column, then answering Q1 in traditional way.

All students answered the two questions using the two formats for fairness. However, for the analysis I only considered the answer of the first question for each student to avoid any potential carry over effect from writing in the first format to the second format.

The experiment was conducted as part of one normal scheduled workshop. Participation in the workshop was a compulsory component of the course, but inclusion in the study was optional. A workshop tutorial is a face to face meeting between up to 12 students and their tutor. In a workshop, students will be expected to engage with unseen problems provided on the day, normally working in groups with a focus on discussion. In a tutorial, students will have opportunities to ask questions about ongoing coursework, assessed problems and receive and discuss feedback on previously submitted written work. In context, asking students to complete a task and then discuss it would be entirely normal.

To recruit students to the study, I sent an email to students and made a short announcement explaining that I would like their permission to use results from a forthcoming workshop as part of a study to improve the quality of online assessment in mathematics. Students had to opt in to or opt out of having their results included in the study. I randomly assigned each *tutorial group* to one of the four experimental groups. So that each tutor/table was working in the same way. Students were asked to work individually, and no textbooks or notes were allowed. Since participation in the study was voluntary some groups had fewer students participate in the experiment.

4.1.3 Task construction

To conduct a study that would answer the research questions, two questions were chosen from the book (Stewart 2007) that was used in the course. The questions were chosen so that the mathematics should be familiar to most of the students. I don't want to confound a lack of mathematical understanding with the format effect. Furthermore, the questions needed to be of a standard that was typically found during the students' studies and one that they had not seen before. I worked with the course organiser to choose the two questions. The two questions are given below.

Q1. Explain why the following integral is improper and determine whether it converges or diverges.

$$\int_{1}^{2} \frac{x}{\sqrt{x-1}} \mathrm{d}x$$

Q2. Explain why the following integral is improper and determine whether it converges or diverges.

$$\int_{-\infty}^{0} xe^{x} \mathrm{d}x$$

Notice that both questions involve finding limits using L'Hopital's rule. They both involve very simple algebraic expressions in the numerator and denominator, where I anticipate no difficulty for the student group I have in mind correctly differentiating the expressions.

The materials were provided to participants as a paper worksheet. Two worksheets were created; two-column worksheet, and traditional worksheet. Both worksheets introduced the format by an example that illustrates how to apply the format when writing a proof. Then brief advice was given on how to use this format. The two-column worksheet used in this study is shown Figure 4.1.

In the same way, participants in the traditional group were given a worksheet that was of a similar length to the worksheet given to the two-column group. The worksheet provided a brief description and advice on how to present a mathematical argument in a paragraph. An example was also provided. The traditional worksheet used in this study is shown in Figure 4.2.

In both worksheets, students were then asked to answer a question that was of approximately the same length and difficulty, according to the group they were randomly assigned to. This ensured students in all groups spent approximately the same amount of time on task. There was no limit to the time spent on answering the questions, however there was a total time limit of one hour and a half. The traditional example does differ in that it requires two applications of L'Hopital's rule, rather than just one for the two-column example. This potential additional complexity is accepted because otherwise the traditional presentation is too short. The traditional example also does not have a finite limit, whereas the two-column example has limit 1. In order to simulate learning conditions familiar to the students, the participants in this study needed to see an example of the whole proof as constructed in the two format (two-column, traditional.)

4.1.4 Coding students' responses

For the purpose of this study, I have applied the unit coding scheme which was discussed in Section 3.2.

The written justifications made by participants were transcribed and coded using the four categories described in the unit coding scheme. The explanation category (i.e. principle-based, goal-driven and noticing) has meaning that is identical to those used in (Hodds, Alcock, and Inglis 2014). The non-explanation category (i.e. paraphrasing) is illustrated by the following examples shown in Figure 4.3 and Figure 4.4 taken from the study.

In the first example (see Figure 4.3), a student wrote "Evaluate integral" three times as justifications. So, based on the definition of paraphrasing in the coding scheme, the exact statement "Evaluate integral" has been repeated and used to describe the calculations. A student also used the same statement in the final answer. This was coded as paraphrasing since there is no explanation or additional information added by this statement.

The second example (see Figure 4.4), also presents statements classed as paraphrasing. The student wrote three statements "simplifying integral, performing integral, and Evaluating the integral." So the student merely uses the three statements to paraphrase in words what is already in the calculations that he/she has just done.

Presenting arguments using a two-column format
A mathematical argument in *two-column format* is written in columns, separating out the statements in the argument from the justifications of those statements.
Please read the following example carefully to help you understand how to use this format yourself.

Example. Use l'Hospital's Rule to evaluate the given limit

$$\lim_{x\to 1}\frac{\ln x}{x-1}$$

Solution (using *two-column format*):

No.	Statement	Justification
1	Consider $\lim\limits_{x\to 1}\dfrac{\ln x}{x-1}$	Problem statement
2	$\lim\limits_{x\to 1}\ln x = 0$ and $\lim\limits_{x\to 1} x-1=0$	Direct evaluation gives $\frac{0}{0}$ which is an indeterminate form
3	Now consider $\lim\limits_{x\to 1}\dfrac{\frac{\mathrm{d}}{\mathrm{d}x}\ln(x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x-1)}$	Attempt l'Hospital's Rule
4	$=\lim\limits_{x\to 1}\dfrac{1/x}{1}$	Evaluate derivatives
5	$=1$	Evaluate limit, which exists
6	So $\lim\limits_{x\to 1}\dfrac{\ln x}{x-1}=\lim\limits_{x\to 1}\dfrac{\frac{\mathrm{d}}{\mathrm{d}x}\ln(x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x-1)}$	Apply l'Hospital's Rule
7	$\lim\limits_{x\to 1}\dfrac{\ln x}{x-1}=1$	5 & 6: Conclusion

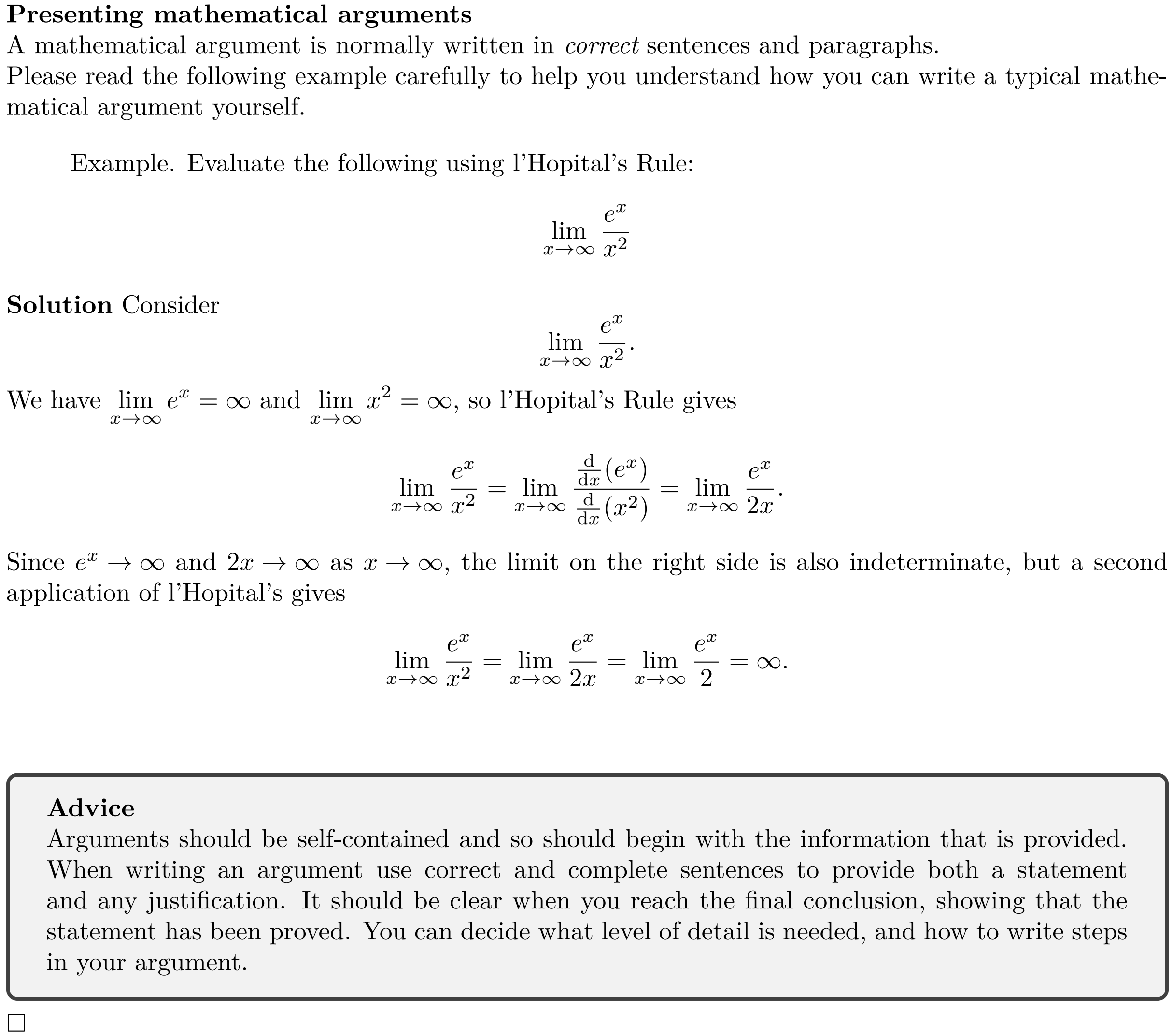
Advice To use the two-column format, please apply the following steps:

1. Start with the given information
2. Number each step
3. Write justifications in the second column referring to step numbers as needed

You can decide what level of detail is needed, and when to combine small steps into a single step.

□

Figure 4.1: Two-column worksheet



Presenting mathematical arguments
A mathematical argument is normally written in *correct* sentences and paragraphs.
Please read the following example carefully to help you understand how you can write a typical mathematical argument yourself.

Example. Evaluate the following using l'Hopital's Rule:

$$\lim_{x\to\infty}\frac{e^x}{x^2}$$

Solution Consider

$$\lim_{x\to\infty}\frac{e^x}{x^2}.$$

We have $\lim\limits_{x\to\infty} e^x=\infty$ and $\lim\limits_{x\to\infty} x^2=\infty$, so l'Hopital's Rule gives

$$\lim_{x\to\infty}\frac{e^x}{x^2}=\lim_{x\to\infty}\frac{\frac{\mathrm{d}}{\mathrm{d}x}(e^x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x^2)}=\lim_{x\to\infty}\frac{e^x}{2x}.$$

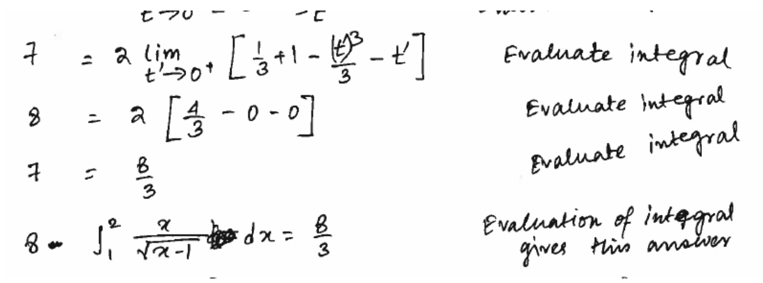
Since $e^x\to\infty$ and $2x\to\infty$ as $x\to\infty$, the limit on the right side is also indeterminate, but a second application of l'Hopital's gives

$$\lim_{x\to\infty}\frac{e^x}{x^2}=\lim_{x\to\infty}\frac{e^x}{2x}=\lim_{x\to\infty}\frac{e^x}{2}=\infty.$$

Advice
Arguments should be self-contained and so should begin with the information that is provided. When writing an argument use correct and complete sentences to provide both a statement and any justification. It should be clear when you reach the final conclusion, showing that the statement has been proved. You can decide what level of detail is needed, and how to write steps in your argument.

□

Figure 4.2: Traditional worksheet



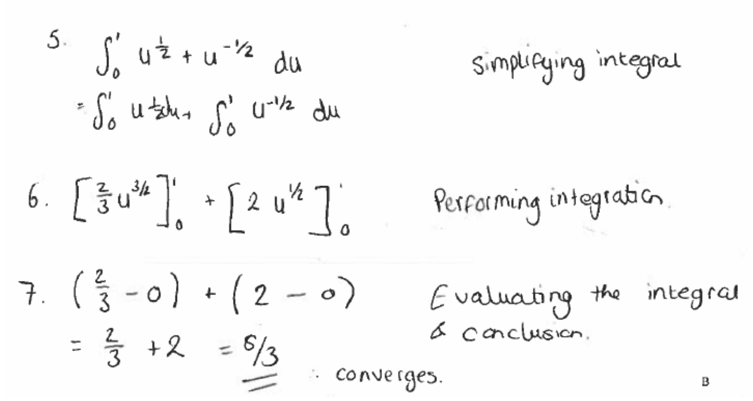
7 $= 2 \lim_{t' \to 0^+} \left[\frac{1}{3} + 1 - \frac{(t')^3}{3} - t' \right]$ Evaluate integral

8 $= 2 \left[\frac{4}{3} - 0 - 0 \right]$ Evaluate Integral

7 $= \frac{8}{3}$ Evaluate integral

8 $\int_1^2 \frac{x}{\sqrt{x-1}} \, dx = \frac{8}{3}$ Evaluation of integral gives this answer

Figure 4.3: Repeated statements



5. $\int_0^1 u^{\frac{1}{2}} + u^{-1/2} \, du$ Simplifying integral

$= \int_0^1 u^{\frac{1}{2}} du + \int_0^1 u^{-1/2} \, du$

6. $\left[\frac{2}{3} u^{3/2} \right]_0^1 + \left[2 u^{1/2} \right]_0^1$ Performing integration

7. $\left(\frac{2}{3} - 0 \right) + (2 - 0)$ Evaluating the integral & conclusion.

$= \frac{2}{3} + 2 = 8/3$ ∴ converges.

B

Figure 4.4: Paraphrased calculations in words

4.1.5 Model Solution

Table 4.1, demonstrates using the coding scheme on the model solution for the first question. There are two main conclusions for this question: explain why the integral is improper and if it is convergent. Notice that for this answer there are eight units. In the first unit U_1, the statement "By def'n of improper integral" was considered as the first principle-based warrant W_{P_1} which used to connect the first data D_1 with it's conclusion C_1. The statement "$f(x)$ has the vertical asymptote $x = 1$" was coded as a qualifier Q_1 for C_1. The second conclusion C_2 was also considered as data for the third unit D_3, so it can be written as C_2/D_3. Similarly, the conclusion for the third unit is the data for the forth unit (i.e, C_3/D_4) and so on. The statement "Using integration by substitution" was coded as a principle based explanation W_{P_2}. There were also two statements coded as paraphrasing W_R "Algebraic rearrangement" and "Evaluate integral" for U_3 and U_5 respectively. The statement "Taking limit using 3" was coded as a noticing explanation W_{N_6} for U_6. The total mark for this answer was 10.

4.2 Results

The data were analysed in three parts. Firstly, a marking scheme was created after discussion between me, Chris Sangwin and Toby Bailey. Initially, I choose 10 responses randomly and each one of us coded the responses individually to try to maintain objectivity and avoid bias with data analysis. If there were any discrepancies, they were discussed until an agreement was reached on the coding scheme. Secondly, the data were marked by hand in the normal way by myself and the justification were coded using the unit coding scheme (see Section 3.2). Finally, the analysis proceeded by using marks, number of units, and written justifications of each format.

No.	Statements	Justification
1	Consider $\int\limits_1^2 \frac{x}{\sqrt{x-1}}\mathrm{d}x$	Problem statement
2	U_1: $\underset{D_1}{\boxed{\int\limits_1^2 \frac{x}{\sqrt{x-1}}\mathrm{d}x}} = \underset{C_1}{\boxed{\lim_{t\to 1+}\int\limits_t^2 \frac{x}{\sqrt{x-1}}\mathrm{d}x}}$	(2 marks) (W_{P_1}) By def'n of improper integral (Q_1) $\frac{x}{\sqrt{x-1}}$ has the vertical asymptote $x = 1$, so the integral is improper
3	U_2, U_3, U_4 and U_5: $\underset{D_2}{\boxed{\int\limits_t^2 \frac{x}{\sqrt{x-1}}\mathrm{d}x}} = \underset{C_2/D_3}{\boxed{\int\limits_{t-1}^1 \frac{u+1}{\sqrt{u}}\mathrm{d}u}}$ $= \underset{C_3/D_4}{\boxed{\int\limits_{t-1}^1 \frac{u}{\sqrt{u}} + \frac{1}{\sqrt{u}}\mathrm{d}u}}$ $= \underset{C_4/D_5}{\boxed{\frac{8}{3} - \frac{2}{3}(t-1)^{3/2} - 2(t-1)^{1/2}}}$ $= \underset{C_5}{\boxed{\frac{8}{3}}}$	(4 marks) (W_{P_2}) Using integration by substitution (W_{R_3}) Algebraic rearrangement (W_{R_5}) Evaluate integral
4	U_6 and U_7: $\underset{D_6}{\boxed{\int\limits_1^2 \frac{x}{\sqrt{x-1}}\mathrm{d}x}} = \underset{C_6/D_7}{\boxed{\frac{8}{3} - 0 - 0}}$ $= \underset{C_7/D_8}{\boxed{\frac{8}{3}}}$	(3 marks) (W_{N_6}) Taking limits using 3
5	U_8: C_8 $\boxed{\text{The integral is convergent}}$	(1 mark) Conclusion

Table 4.1: A model solution for Q1

	Traditional "Tx"($n = 20$)		Two-column "Cx" ($n = 20$)	
	Mean	SD	Mean	SD
Marks for Q1	7.40	1.31	7.70	1.87
Marks for Q2	8.20	1.70	9.40	1.17

Table 4.2: Participants' marks by formats for Q1 and Q2

4.2.1 Effects of formats on students' marks

The influence of formats on students' marks was examined by an independent t-test. There was no significant difference in the marks for traditional "T1" ($M = 7.40$, $SD = 1.31$) and two-column "C1" ($M = 7.70$, $SD = 1.87$) groups who answered Q1; $t(38) = 0.588$, $p = 0.560$ (see Table 4.2).

However, for Q2 there was a significant difference in the marks for traditional "T2" ($M = 8.20$, $SD = 1.70$) and two-column "C2" ($M = 9.40$, $SD = 1.14$) groups who answered Q2; $t(38) = 2.62$, $p = 0.013$ (see Table 4.2).

These results suggest that in some situations there is a difference, so that significant format effects can exist.

4.2.2 Effects of format on the number of students' units

To count students' work in each format, students' derivations were divided into a number of sub-derivations. These sub-derivations were coded based on our definition of a proof unit using the unit coding scheme (see Section 3.2.) There was no significant difference in the total number of units between the two groups; traditional "T1" and two-column "C1" who answered Q1 (T1: $Mdn = 8$), (C1: $Medn = 9$); $U = 198.5$, $p = 0.978$. Similarly there was no significant difference in units between the two groups; traditional "T2" and two-column "C2" who answered Q2 (T2 : $Mdn = 8$), (C2: $Mdn = 9$); $U = 188$, $p = 0.743$.

A Spearman's correlation was run to assess the relationship between students' scores and the number of units written ($N = 80$). There was a positive correlation between students' scores and the number of units, which was statistically significant, $r_s = 0.229$, $p = 0.041$. A higher level of marks is associated with a higher number of units.

4.2.3 Effects of formats on the justifications written

To investigate whether there is a format effect between traditional and two-column formats when students write mathematical arguments, participants' written justifications were coded using the unit coding scheme given in Section 3.2. Principle-based, goal-driven, and noticing statements (classified as explanations) in the mathematical context of this study. Paraphrasing statements (those classified as non-explanations) have meanings directly analogous to those used in previous self-explanation studies (for example Hodds (2014)).

The number of comments of each type given by students in each group, shown in Table 4.3, were analysed using a Mann-Whitney U test. In some cases these data failed to meet the assumption of normality, so a non-parametric test was used. I found that participants in the two-column group "C1" who answered Q1 gave significantly more explanations; that is, they gave more comments categorized as principle-based, $U = 126.5$, $p = 0.041$; goal-driven, $U = 70.5$, $p < 0.001$; or noticing explanations, $U = 140$, $p = 0.009$. Indeed, they gave a median of 4 explanations of these types, whereas those in the traditional group "T1" gave a median of 2, $U = 35.0$, $p < 0.001$. Similarly, those participants in the two-column group "C2" who answered Q2 gave significantly more explanations; they gave more comments categorized as principle-based, $U = 123$, $p = 0.029$; goal-driven, $U = 63$, $p < 0.001$; or noticing explanations, $U = 134$, $p = 0.008$. They gave a median of 4 explanations of these types, whereas those in the traditional group "T2" gave a median of 1, $U = 48.5$, $p < 0.001$. These results suggest that, when using the two-column format, students generated significantly more comments classed as explanation.

Statements	Question	Traditional “Tx” ($n = 20$)		Two-column “Cx” ($n = 20$)	
		Median	Sum	Median	Sum
Principle-based	1	1	22	2	39
Goal-driven		1	20	3	49
Noticing		0	0	1	6
Paraphrasing		0	0	1	31
Principle-based	2	0	15	3	29
Goal-driven		0	9	2	40
Noticing		0	1	1	6
Paraphrasing		0	1	1	32

Table 4.3: Number of explanation and non-explanation statements by formats for Q1 and Q2

Participants in the two-column group “C1” also gave significantly higher non-explanations classed as paraphrasing, $U = 30$, $p < 0.001$. Similarly, participants in the two-column group “C2” produced significantly higher non-explanation paraphrasing, $U = 66.0$, $p < 0.001$. These results suggested that participants in the two-column group produced significantly more paraphrasing.

While the use of comments classed as backing was rather infrequent, there was also a significant difference in the comments classed as backing between the two formats. There was a significant difference between the two groups; traditional “T1” and two-column “C1” who answered Q1 (T1: $Mdn = 0$), (C1: $Mdn = 1$); $U = 138$, $p = 0.028$. Similarly there was a significant difference between the two groups; traditional “T2” and two-column “C2” who answered Q2 (T2: $Mdn = 0$), (C2: $Mdn = 1$); $U = 122.5$, $p = 0.029$. These results suggest that participants in the two-column groups produced significantly more backing to support their warrants.

The Spearman’s correlation was also performed separately for each type of explanations. The only type of explanations that relates positively to students’ marks was the number of principle-based explanations $r_s = 0.375$, $p < 0.001$. This result confirms previous findings by (Ainsworth and Burcham 2007).

The results also show that the number of backings positively correlated with the number of principle-based warrants $r_s = 0.285$, $p = 0.010$.

Overall, these results indicate that using two-column format increased the number of explanation given by participants. However, as a side effect of increasing the amount of justifications, students in the two-column generated more non-explanation statements which coded as paraphrasing.

4.3 Discussion

For the purpose of this study I applied the unit coding scheme to answer the research questions which investigates any format effects between traditional presentation and the two-column format when writing mathematical arguments. A secondary outcome is our use and further development of the unit coding scheme which has provided us with a useful tool to understand students’ writing more generally. The unit coding scheme was adapted from (Hodds, Alcock, and Inglis 2014) and (Toulmin 1958), combined with ideas of structured derivations from (R. J. Back 2010), and combined them to describe students’ mathematical arguments.

I now contrast the results of analysing students’ work on the two formats with references to the aspects of the coding scheme. The findings show that participants in the two-column group produced significantly more explanations (principle, goal, noticing) than those in the traditional group.

I suggest this is due to the format structure in which students are expected to justify each line in the statements’ column with a corresponding comment in the justification column. However, as a side effect of increasing the amount of justifications, students in the two-column group generated more non-explanation statements which coded as paraphrasing.

I compared students’ marks on writing traditional and two-column proofs designed and used for a workshop tutorial which was a compulsory component of the course. To ensure fairness and reduce the impact of pre-existing differences between students, I randomly assigned participants to groups in this experiment. This technique helps control for selection bias and strengthens the internal validity of the study. By design, students were asked to answer questions in both formats, but only the responses to the first format were included in the study to avoid any potential carryover effect from the first format to the second format.

The results suggest that it is specific types of explanation that were associated with subsequent marks. The number of principle-based explanations positively correlated with marks, which is in line with (Ainsworth and Burcham 2007). However, the number of goal-driven explanations did not correlate with marks, which is also in line with (Ainsworth and Burcham 2007).

Moreover, it was observed that there were significantly higher numbers of principle-based and noticing explanations produced by the two-column group than the traditional group. According to (Hodds, Alcock, and Inglis 2014), increased numbers of noticing statements between lines and the principles used to generate the proof would indicate that a student is being encouraged to better use of existing understanding of logical reasoning.

The results also show that there was a significant difference between the number of goal-driven explanations given by the two formats. A possible reason for this is that the structure of the two-column format might lead students to write more justifications than is expected with traditional proofs. So, while a student was thinking to write an explanation for each unit they would often write explanations related to the structure of the proof to move from one line to the other. In the two-column format, students provided more backing to support their warrants.

The results also show that the number of backings positively correlated with the number of principle based explanations. One possible reason for this is that when a student writes a statement and justifies it using definitions or theorems, they usually support them to provide further evidence. According to (Inglis, Mejia-Ramos, and Simpson 2007), backing is usually used to provide more evidence to support the warrants.

In addition, in the traditional groups, some students only stated the final answer as a numerical value without writing whether the integral converges or diverges although that was a part of the task and that was coded as an incomplete final unit.

Students in the two-column groups generated more paraphrased and repeated justifications due to the format effect. A possible reason is that while some students were thinking about what to write in the justification column, they would just repeat the calculation they had just done in words in the right column to fill the gaps.

There are cognitive aspects linked to using constraints, which I believe are potentially beneficial. The two-column proof format separates out the individual steps from the justification for the legitimacy of individual steps. Using the format to consciously separate the steps from their justification is highly likely to reduce cognitive load, particularly for novices, by providing a structured and explicit format to work within. In the two-column proof, the constraints in the interface could help to reduce the extraneous cognitive load by providing a structured and explicit format to work within. The format has the potential to help us better understand the nature of mathematical arguments, and how students write these arguments, by comparing constrained (two-column) arguments from unconstrained free form arguments. Two-column format could lead to an efficient construction of cognitive load, because they focus the attention of learners on specific steps. For enhancing cognitive skills acquisition, I believe that it is useful to use Two-column format for novices by providing a structured and explicit format to work within.

Hence when measuring the effectiveness of using such proofs with instructional designs, it is expected to see the expertise reversal effect (Kalyuga, Rikers, and Paas 2012) which refers to the reversal of the effectiveness of instructional techniques on learners with differing levels of prior knowledge. In particular, the two-column format might well significantly help early proof attempts, but hinder experts for whom the scaffolding is a distraction.

While the study yielded a clear result, caution must be taken when interpreting the generality of the finding. The following limitations are highlighted. In this study, research was conducted in one university in the UK. As a result, students from different universities and countries will have different mathematical backgrounds. The study used a modest sample of students ($N = 80$) from a single course at a single university. While I cannot claim that the sample is completely representative of the broader population of students undertaking mathematics modules at universities around the world, there is a good reason to think that the results will be applicable in some other situations. A blend of pure and applied mathematics is taught in lectures and small group tutorials. Consequently, caution should be exercised about stating which students the effects observed in the studies apply to, but one should be reasonably confident that the format effect would influence the way in which mathematical arguments are written. The format might have an effect on students' marks, the written justification or the number of units. There is a lot of common ground amongst universities and so, despite this limitation, similar format effects are expected in other contexts. Individual student circumstances may need to be considered in order to understand how the written format might affect students' written work. Students' mathematical backgrounds and prior teaching experiences might differently influence how the format effect exist on students' written arguments. In this study, I chose handwriting to gather the data because some understanding of what students do in practice will be helpful to the future design of online assessments. In addition, I do not wish to confound the possible format effect with an effect due to either lack of familiarity (e.g. LaTeX is difficult for students to learn), or to resistance to change. However, I cannot be certain whether the effect is confounded with working on paper as opposed to the use of technology. The same methods can be applied to different materials, using different samples of the student population, and implementing the two-column format using online systems.

4.4 Conclusion

The study took the suggestion from the literature that students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). Many of these online assessment or learning systems are implementing two-column input mechanisms (e.g STACK, SOWISO, MathXpert). The study investigated whether there is a format effect between traditional and two-column formats when students write mathematical arguments. I undertook an experiment to compare students' writing between unconstrained traditional arguments and arguments in a two-column format. Students from a first year calculus course were invited to take part in the study. Four experimental groups resulted which allowed a comparison to be made between the effects of formats and type of justifications when writing proofs. The analysis of the data obtained revealed there is a difference in students' marks, so that significant format effects can exist. The results also suggest that it is specific types of explanations that are associated with subsequent marks which is in line with (Ainsworth and Burcham 2007). It was found that students in the two-column group produced significantly more higher quality explanations than did the traditional group. One drawback of the two-column format is that participants tend to justify all the steps which resulted in producing some paraphrased and repeated explanations. However, the paraphrasing can be considered as accepted truths as defined by (Stylianides, Sandefur, and Watson 2016). Some accepted truths may be considered trivial or basic knowledge for a particular audience and thus may be omitted from a proof. However, the novices found what I called "paraphrasing" is important to be explicitly written.

One of the methodological contributions of this thesis is to illustrate how to use the unit coding scheme as a guide to analyse the participants' responses. In this study, the unit coding scheme was applied to analyse students responses. The unit coding scheme was adapted from (Hodds, Alcock, and Inglis 2014) and (Toulmin 1958), combined with ideas of structured derivations from (R. J. Back 2010), and combined them to describe students' arguments (See Section 3.2). The Toulmin scheme of argumentation can provide an insight into the logical relationships within a proof and it has been used widely in the mathematics education literature. The difference between the unit coding scheme presented in this thesis and previous uses of the Toulmin model is that the research reported in this thesis uses the unit coding scheme to break down the components of a mathematical arguments in order to discuss the effect of writing mathematical arguments in different formats using online assessment. Note that the unit coding scheme provides a qualitative method of gaining insight into how many units are typed and what kind of justification are written. The importance of the unit coding scheme is not only considering the number of units but also the quality and relevance of the statement within those units. In this thesis, the unit coding scheme is basically used as a tool for analysing a single argument in students' responses. Although students' arguments are varied and might be wrong. For example, some students might generate wrong warrants to connect their data to the conclusions, or they might give warrants that do not match the calculation steps. In general, the number of units considered a measure of the depth or quality of students' proof construction and that the other marking schemes used are used to balance limitations of this. The unit coding scheme can be applied into many subjects and fields when analysing students arguments particularly in for the online assessment purposes.

The research was conducted in one university in the UK. As a result, students from different universities and countries will have different mathematical backgrounds. While I cannot claim that the sample is completely representative of the broader population of students undertaking mathematics modules at universities around the world, there is good reason to think that the results will be applicable in some other situations. Consequently, caution should be exercised about stating which students the effects observed in the studies apply to, but one should be reasonably confident that the format effect would influence the way in which mathematical arguments are written. The format might have an effect on students' marks, the written justification or the number of units. There is a lot of common ground amongst universities and so, despite this limitation, similar format effects are expected in other contexts.

Individual student circumstances may need to be considered in order to understand how the written format might affect students' written work. Students' mathematical backgrounds and prior teaching experiences might differently influence how the format effect exist on students' written arguments.

Chapter 5

Study 2: Typing vs. Photograph

This chapter presents the second of three studies focused on a potential format effect when writing mathematics online. The results of Study 1 (see Chapter 4) into the format effects when writing mathematics arguments are a promising indicator. Students are increasingly moving away from paper submission of assignments to working online, a trend accelerated in 2020/21 by the global pandemic. Online submission includes both automated online assessment and online submission of written work for human making. A natural question is therefore this: Is there a difference in *performance* and *justifications* between uploading handwritten and typing in writing mathematical responses? The following study reported in this chapter addresses this question. A secondary outcome was the use and further development of the coding scheme which I provided in Section 3.2.

I conducted an experiment in which participants responded to an online task containing equivalent typing and uploading handwritten items, and students' reactions immediately after the task were obtained.

Factors explored included the overall score awarded, types of justification, and the number of units.

5.1 Method

5.1.1 The purpose of the study

This study is undertaken to investigate the difference and the format effects of typing vs handwriting mathematical responses. The main goal of this study is to understand potential format effects when students input their responses into a machine for the purpose of human assessment of complete solutions. I also want to understand how students can more effectively input their responses for the purposes of assessment online by a human.

I undertook an experiment to compare students' mathematical responses to routine coursework tasks, i.e. short mathematical questions, in two formats: i) uploading handwritten mathematical response, and ii) typing responses directly.

Quantitative and qualitative key aspects were considered by answering the following research questions.

1. Is there a difference in *performance* between uploading handwritten and typing in writing mathematical responses?
2. How many units has a student used in each format? The number of units is a quantitative measure of the amount of written work submitted. See Section 3.2 for more details about the unit coding scheme.
3. Is there a difference in types of justification between the two formats?
4. Is there a correlation between the number of units and performance?

I.e. when students write more details do they receive high scores?

5. What are students' attitudes toward the formats when writing mathematical responses?

It was expected that students might get slightly higher marks for uploading their handwritten responses as typing mathematics is difficult. Participants in the uploading group might also provide more justifications.

5.1.2 Procedure

There were 86 volunteer students recruited from the course *Introduction to Linear Algebra* (ILA) in the first semester of 2020/21. Details are in Appendix C. Ultimately 62 students answered the two questions, either typing or handwriting based on their assigned group. The reminder did not answer the questions, but submitted their views regarding the preferable format.

Students were asked to complete two routine coursework tasks, varying the format between writing and typing.

Participants were randomly assigned into two groups (Group 1: TY for "Typing" and, Group 2: UP for " Uploading") as follow:

TY : Answering Question.1 (Q1) by typing the response online using the editor provided in Moodle, then to handwrite the response of Question.2 (Q2) and upload it (Type then upload).

UP : Answering Question.1 (Q1) by handwriting the response and upload it, then to type the response of Question.2 (Q2) using the editor provided in Moodle. (Upload then type).

The two groups arise for equity reasons: I want to make sure each student writes in both formats (typing/uploading) and that I have to make sure I have materials which contain the two questions in each format. For that reason two questions were used; Q1 and Q2. All students answered the two tasks using the two formats for fairness. However, for the analysis I only considered the answer of Q1 for each students to avoid any potential carry over effect from writing/typing in the first format to the second format.

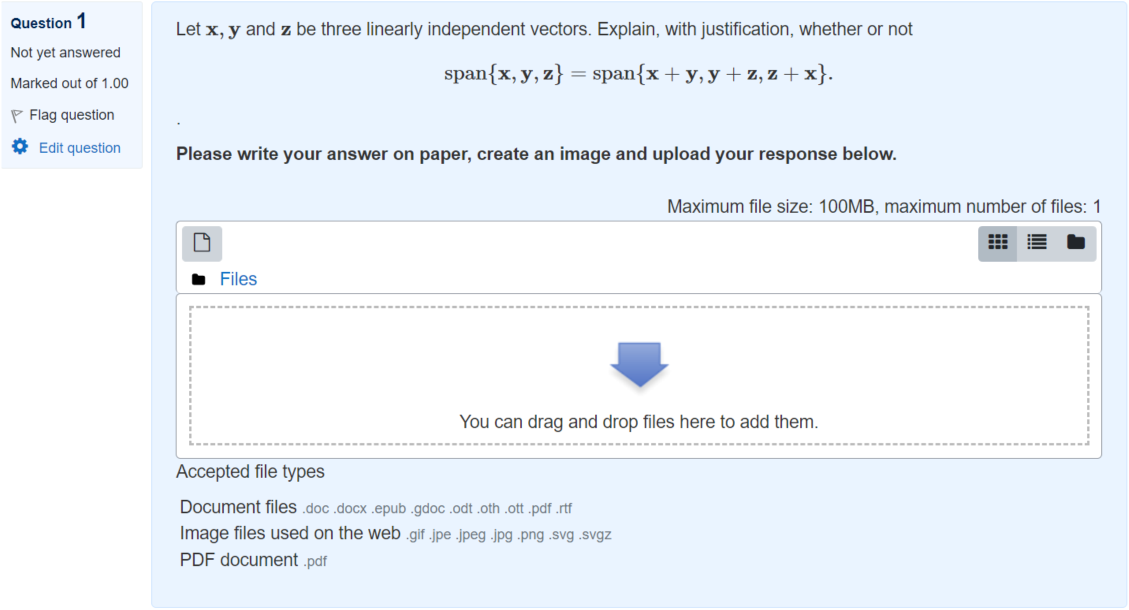
The instrument also contained a short questionnaire (see Appendix 5.2.3) with a space for open comments regarding the process. The students' responses were analysed qualitatively in order to elicit in an open-ended way participants' perceptions on the convenience, usefulness and acceptability of each format. To analyse the data, I only considered the first stage of the design where I compare Q1 between the two groups (TY vs. UP)

5.1.3 Task construction

For the purpose of the study, I choose two relatively short problems on the same topic "Subspaces and Spanning" from the course ILA taught at the University of Edinburgh year 1 in the first semester of 2020/21. The questions, together with model solutions are shown in Figures 5.3 and 5.4.

The two problems illustrate typical mathematics at university level. Both problems are typical in style and of appropriate level of difficulty for the group participating in the study. Both problems require use of a formal definition and also involve the use of some formal notation. E.g. they require minimal use of displayed equations and two dimensional layout such as large matrices which are difficult to typeset. Specifically, these tasks contain the following notation features, typical of more advanced mathematics.

- Use of special fonts. E.g. $\mathbb{R}$ refers to the set of real numbers. The typeface is referred to as "blackboard bold" and is particular to mathematical writing.
- Bold typeface fonts are often used in published work to distinguish between scalars, e.g. a, and vectors, e.g. $\mathbf{x}$ (Poole 2011). In written work, a vector could be distinguished by underlining it, e.g. $\underline{x}$ or with an arrow over the letter, e.g. $\vec{x}$.



Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\mathrm{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \mathrm{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

.

Please write your answer on paper, create an image and upload your response below.

Maximum file size: 100MB, maximum number of files: 1

You can drag and drop files here to add them.

Accepted file types

Document files .doc .docx .epub .gdoc .odt .oth .ott .pdf .rtf
Image files used on the web .gif .jpe .jpeg .jpg .png .svg .svgz
PDF document .pdf

Figure 5.1: Screen shot for a Moodle's essay question interface where students can upload a picture of handwritten work

- Mathematics makes extensive use of both superscripts and subscripts. For example, the n-dimensional real vector space is written $\mathbb{R}^n$. In some situations subscripts can be avoided. For example in Q1 the scalars are taken to be a, b, c (avoiding subscripts) but in Q2, a_1, a_2, a_3 (with subscripts) have been chosen. (Poole 2011) for instance often subscripts.

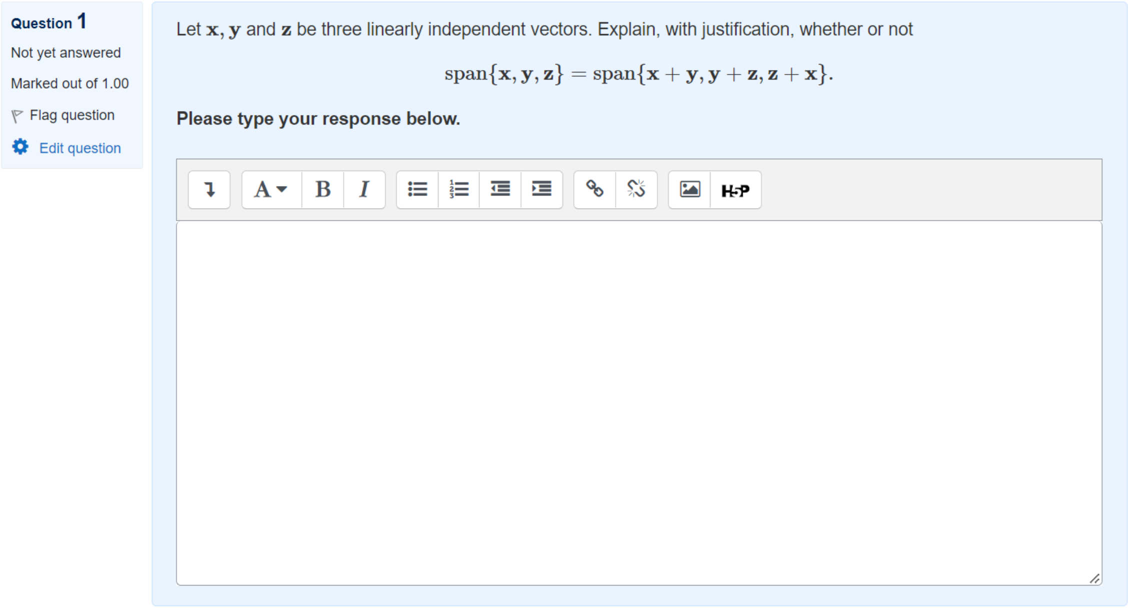
- Traditional mathematical writing makes extensive use of relative position on the page, rather than having consecutive words follow in a linear fashion. This includes displaying single equations centered on a line, and large two dimensional structures such as the matrices used in Q2. Equations are lined up, and work is sometimes displayed in columns, e.g. as in the two-column study in Chapter 4.

- Mathematical writing makes extensive use of special symbols. Mathematics symbols are often synonyms, e.g. $=$ is a synonym for "is equals to" and $\subseteq$ is a synonym for "is a subset of". Most mathematical expressions form part of a sentence and can be read as such. Some aesthetic judgement is needed to choose the linguistic or symbolic form, but modern fashion often favours symbolic alternatives.

While the two tasks do not have the distinctive features of calculus notation these simple proofs in linear algebra certainly have a typical range of typographical features which makes mathematics distinctive as a subject. Also note I choose no graphs or diagrams.

The tasks were implemented directly in Moodle's essay question type. Figure 5.1 illustrates a screen shoot for a Moodle's essay question interface where a student can upload a picture of handwritten work. Figure 5.2 illustrates a screen of an example of Moodle's essay question interface where students can type directly.

To recruit students to the study I sent a short announcement explaining the need for some volunteers to participate in a study. The announcement is provided in Appendix B.2.1. The questions used in the study were administered using the Moodle virtual learning environment as is standard practice at the university. The data collection of students' responses to these items was implemented directly in Moodle's essay question type. Since the submission and marking procedure used in this study is very similar to normal assessment practice in this module, I believe the research has an important element of authenticity.



Question 1

Not yet answered

Marked out of 1.00

Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\operatorname{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

Please type your response below.

Figure 5.2: Screen shot for Moodle's essay question interface where students can type

Q1. Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\operatorname{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

If $U \;= \operatorname{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ in $\mathbb{R}^n$ show that $U = \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$.

Model solution of Q1 :
Assume $W := \mathbf{v} \in \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$ then

$$\begin{aligned} \mathbf{v} &= a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x}) \\ &= (a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z} \end{aligned}$$

so that if $\mathbf{v} \in W$ then $\mathbf{v} \in U$, i.e. $W \subseteq U$.
Assume $\mathbf{v} \in U$ then

$$\mathbf{v} = a\mathbf{x} + b\mathbf{y} + c\mathbf{z}$$
$$= \frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W$$

i.e. $U \subseteq W$.

Figure 5.3: Q1 and its model solution

Q2 Let $\mathbf{u}, \mathbf{v}$ and $\mathbf{w}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\operatorname{span}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \operatorname{span}(\mathbf{w}, 2\mathbf{v} + \mathbf{u}, \mathbf{u} + 2\mathbf{v} - \mathbf{w}).$$

Model solution of Q2 :
Let $\mathbf{x} \in \operatorname{span}(\mathbf{w}, 2\mathbf{v} + \mathbf{u}, \mathbf{u} + 2\mathbf{v} - \mathbf{w})$. Then

$$\mathbf{x} = a_1\mathbf{w} + a_2(2\mathbf{v} + \mathbf{u}) + a_3(\mathbf{u} + 2\mathbf{v} - \mathbf{w})$$

$$= (a_2 + a_3)\mathbf{u} + (2a_2 + 2a_3)\mathbf{v} + (a_1 - a_3)\mathbf{w}.$$

So $\mathbf{x} \in \operatorname{span}(\mathbf{u}, \mathbf{v}, \mathbf{w})$.
Let $\mathbf{x} \in \operatorname{span}(\mathbf{u}, \mathbf{v}, \mathbf{w})$, so that $\mathbf{x} = a_1\mathbf{u} + a_2\mathbf{v} + a_3\mathbf{w}$.
To show $\mathbf{x} \in \operatorname{span}(\mathbf{w}, 2\mathbf{v} + \mathbf{u}, \mathbf{u} + 2\mathbf{v} - \mathbf{w})$ we need to find b_1, b_2 and b_3 so that

$$\mathbf{x} = b_1\mathbf{w} + b_2(2\mathbf{v} + \mathbf{u}) + b_3(\mathbf{u} + 2\mathbf{v} - \mathbf{w})$$

$$= (b_2 + b_3)\mathbf{u} + (2b_2 + 2b_3)\mathbf{v} + (b_1 - b_3)\mathbf{w}.$$

I.e. we need to solve

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 2 & 2 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

The reduced row echelon form is

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

so a solution does not exist. Therefore $\operatorname{span}(\mathbf{u}, \mathbf{v}, \mathbf{w}) \neq \operatorname{span}(\mathbf{w}, 2\mathbf{v} + \mathbf{u}, \mathbf{u} + 2\mathbf{v} - \mathbf{w})$.

Figure 5.4: Q2 and its model solution

5.1.4 Coding students' responses

To minimize bias, students' responses were given to two mathematics PhD students at the University of Edinburgh to mark independently. The two markers received an email with instructions on how to code students' responses. The data consisted of 28 handwritten responses and 34 typed responses. The handwritten and typed scripts were printed and split randomly between the two markers for fairness, so both can deal with the two formats equally. Any discrepancies were discussed with me, until an agreement was reached in line with standard practice when undertaking qualitative research.

The instructions for the two markers are provided in Appendix B.2.3.

The coding scheme used to code the responses for this study was consisted of two parts:

1. Traditional mark scheme for the awarding marks (see Section 5.1.4).

2. The unit coding scheme to code students' mathematical arguments (see Section 3.2).

Recall, a proof is nested and hierarchical. A main derivation can be divided into a number of more detailed internal sub-derivations, in a recursive manner, I call *units* (see Section 3.1). Each unit (U) will normally contain data (D), conclusion (C) and may have explicit warrant (W) (justification). As discussed, a unit can consist of algebraic expressions or written words or both. Indeed, algebraic expressions are considered to be an integral part of a complete mathematical sentence. So for coding participants' response, the task for the two markers were to firstly identify the awarded mark using the traditional mark scheme. Second, to identify the total number of units of each proof, where each unit contains data (D), conclusion (C) and may have an explicit warrant (W).

Traditional coding scheme

The criteria for awarding marks were adapted from (Sangwin and Köcher 2016), who used the criteria from the International Baccalaureate examination board. The coding scheme includes instructions on how to use the mark scheme, including the awarding of marks in the following categories:

1. M Marks awarded for attempting to use a correct Method.

2. (M) Marks awarded for Method; may be implied by correct subsequent working.

3. A Marks awarded for an Answer or for Accuracy: often dependent on preceding M marks.

4. (A) Marks awarded for an Answer or for Accuracy; may be implied by correct subsequent working.

5. R Marks awarded for clear Reasoning.

6. N Marks awarded for correct answers if no working shown.

Figure 5.5 contains a "model answer" to Q.1 written in a traditional way, together with specific marks.

Q1. Prove $\text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\} = \text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$.

Response	Marks
Let $U := \text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$. Assume $\mathbf{v} \in W := \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$	M1: Attempt to show $A = B$ by showing $A \subseteq B \wedge B \subseteq A$.
then there exist $a, b, c \in \mathbb{R}$ such that	R1A1
$\mathbf{v} = a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x})$ $= (a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z}$	
so that if $\mathbf{v} \in W$ then $\mathbf{v} \in U$, i.e. $W \subseteq U$.	R1A1: Reasoning by taking an arbitrary member of W, and accuracy in using the definition of W).
Assume $\mathbf{v} \in U$ there exist $a, b, c \in \mathbb{R}$ such that	M1: Attempt to show $U \subseteq W$.
$\mathbf{v} = a\mathbf{x} + b\mathbf{y} + c\mathbf{z}$ $= \frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W$	R1A1: Reasoning by taking an arbitrary member of U, and accuracy in using the definition of U.
i.e. $U \subseteq W$.	R1
Total	[9 marks]

Figure 5.5: Q.1 with a traditional marking scheme

Markers first assign a mark using the traditional scheme. They then identify the total number of units in each proof, analyzing their structure based on the unit coding scheme. This process allows the researchers to analyze both the awarded marks (traditional scheme) and the structure of students' arguments (unit coding scheme) to gain a deeper understanding of their mathematical thinking. This coding scheme is appropriate for the study for several reasons. First, it employs independent double marking and blinded marking to minimize bias. Second, it uses a combined approach with both a traditional coding scheme and a unit coding scheme, allowing for a comprehensive assessment of both student outcomes and reasoning processes. Finally, this approach aligns well with the study's likely goal of understanding mathematical argumentation and investigate the difference in performance and justification between uploading handwritten and typing in writing mathematical responses.

5.2 Results

The data were analysed in three parts. Firstly, a marking scheme was created after discussion with the course organiser. If there were any discrepancies, they were discussed until an agreement was reached. Secondly, the data were marked by hand by the two markers using the traditional mark scheme. Finally, the analysis proceeded by using the awarding marks and the total number of units of each format. In the study, all students answered the two questions using the two formats for fairness. However, for the analysis I only considered the answer of the first question for each student to avoid any potential carry over effect from writing in the first format to the second format.

5.2.1 Effects of formats on students' marks

The influence of formats on students' marks was examined by an independent sample t-test. The independent samples t-test is used to compare the mean scores between two independent groups. There was a significant difference in the marks for uploading 'UP' (M = 6.48, SD = 2.86) and typing 'TY' (M = 4.76, SD = 2.57) groups who answered Q1; $t(59) = 2.46$, $p = 0.017$ (Table 5.1).

These results suggest that in this situation there is a difference, so that significant format effects can exist.

Response	Codes
Assume $\underset{D_1}{\boxed{W := \mathbf{v} \in \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}}}$	$D1$: Attempt to show $A = B$ by showing $A \subseteq B \wedge B \subseteq A$.
then $\underset{C1/D2}{\boxed{\mathbf{v} = a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x})}}$	$C1/D2$: The conclusion for the first unit is the data for the second unit.
$= \underset{C2/D3}{\boxed{(a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z}}}$	$C2/D3$: The conclusion for the second unit is the data for the third unit.
so that if $\underset{C3}{\boxed{\mathbf{v} \in W \text{ then } \mathbf{v} \in U, \text{ i.e. } W \subseteq U}}$	$C3$: Reasoning by taking an arbitrary member of W, and accuracy in using the definition of W.
Assume $\underset{D4}{\boxed{\mathbf{v} \in U}}$	D4: Attempt to show $U \subseteq W$.
then $\underset{C4/D5}{\boxed{\mathbf{v} = a\mathbf{x} + b\mathbf{y} + c\mathbf{z}}}$	$C4/D5$: The conclusion for the fourth unit is the data for the fifth unit.
$= \underset{C5/D6}{\boxed{\frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W}}$	$C5/D6$: The conclusion for the fifth unit is the data for the sixth unit.
i.e. $\underset{C6}{\boxed{U \subseteq W}}$	$C6$: the conclusion of the proof
Total: [6 Units]	The total number of units is 6 since there are 6 conclusions in the proof.

Figure 5.6: A coded model solution for Q.1

Statements	Typing “TY” ($n = 34$)		Uploading “UP” ($n = 27$)	
	Median	Sum	Median	Sum
Principle-based	3	28	2	20
Goal-driven	0	16	1	18
Noticing	3	14	1	5
Paraphrasing	0	6	0	4

Table 5.3: Number of explanation and non-explanation statements by formats for Q1.

	Uploading “UP”($n = 28$)		Typing “TY” ($n = 34$)	
	Mean	SD	Mean	SD
Marks (out of 9)	6.48	2.86	4.76	2.57

Table 5.1: Participants’ marks by formats

5.2.2 Effects of format on the number of students’ units

To quantify the amount of students’ work in each format, students’ derivations were divided into a number of sub-derivations. These sub-derivations were coded based the unit coding scheme (see Section 3.2). These data were also analysed using a Mann-Whitney U test, since these data (i.e. Units and justification) failed to meet the assumption of normality, so a non-parametric test was used.

These results suggest that participants in the uploading group ‘UP’ generated significantly more derivations coded as units, $U = 238, p = 0.001$. Indeed, they gave a median of 9 units, whereas those in the typing group ‘TY’ gave a median of 4 (Table 5.2). These results suggest that participants in the uploading group produced significantly more units.

	Uploading “UP” ($n = 28$)		Typing “TY” ($n = 34$)	
	Median	Sum	Median	Sum
Units	9	201	4	143

Table 5.2: Number of units by formats

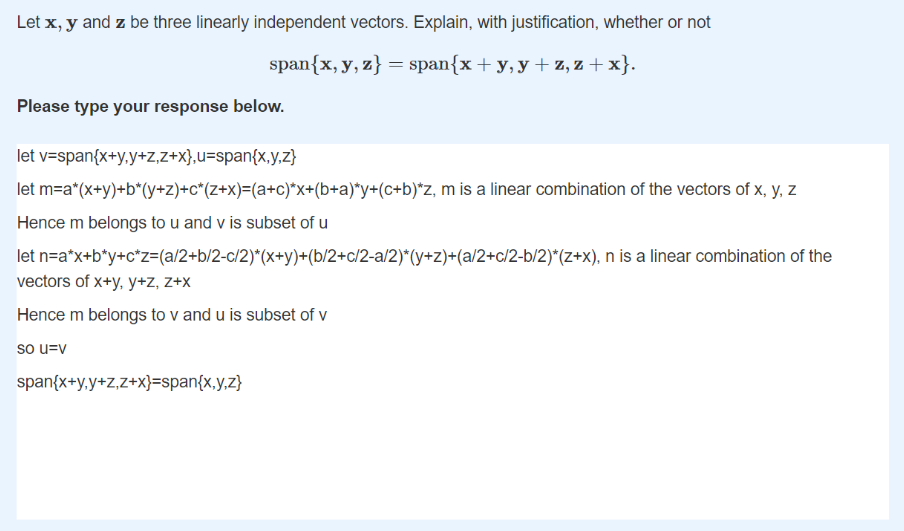
5.2.3 Effects of format on the justifications written

To investigate whether there is a format effect between typing and uploading handwritten responses, participants’ written justifications were coded using the scheme given in Chapter 3, and used here in ways directly analogous to those used in previous self-explanation studies. Recall that the unit scheme given in Chapter 3 and used in this study includes explanations, which are principle-based, goal-driven, and noticing statements and non-explanations such as paraphrasing statements. The number of comments of each type given by students in each group, is shown in Table 5.3.

There was no significant difference in the total number of explanations between the two groups; typing ‘TY’ and Uploading ‘UP’ who answered Q1; $U = 448$, $p = 0.87$.

Participants in the typing group gave more explanation classified as principle based and noticing. However, participants in the uploading group gave higher explanations classed as goal-driven but this was not statistically significant; $U = 394$, $p = 0.29$.

A Spearman’s correlation was run to assess the relationship between students’ scores and the number of units written ($N = 62$). There was a positive correlation between students’ scores and the number of units, which was statistically significant, $r_s = 0.815$, $p < 0.001$. A higher level of marks is associated with a higher number of units. This result confirms previous findings by (Ainsworth and Burcham 2007), and (Alarfaj and Sangwin 2021). The Spearman’s correlation was also performed separately for each type of explanations. The only type of explanation that relates positively to students’ marks was the number of goal-driven explanations $r_s = 0.285$, $p < 0.026$. In addition, the number of goal-driven explanations was also correlated positively with the total number of units $r_s = 0.294$, $p < 0.022$.



Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

Please type your response below.

let v=span{x+y,y+z,z+x},u=span{x,y,z}

let m=a*(x+y)+b*(y+z)+c*(z+x)=(a+c)*x+(b+a)*y+(c+b)*z, m is a linear combination of the vectors of x, y, z

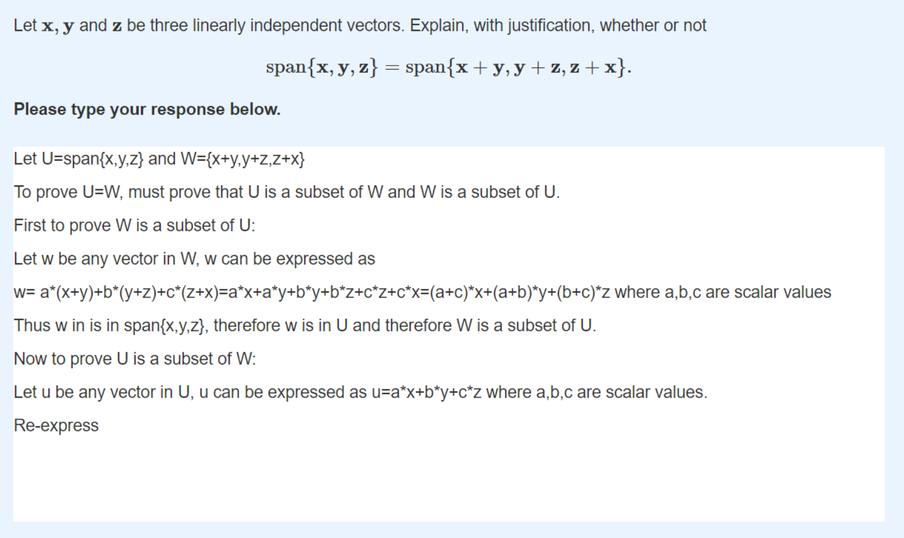
Hence m belongs to u and v is subset of u

let n=a*x+b*y+c*z=(a/2+b/2-c/2)*(x+y)+(b/2+c/2-a/2)*(y+z)+(a/2+c/2-b/2)*(z+x), n is a linear combination of the vectors of x+y, y+z, z+x

Hence m belongs to v and u is subset of v

so u=v

span{x+y,y+z,z+x}=span{x,y,z}



Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

Please type your response below.

Let U=span{x,y,z} and W={x+y,y+z,z+x}

To prove U=W, must prove that U is a subset of W and W is a subset of U.

First to prove W is a subset of U:

Let w be any vector in W, w can be expressed as

w= a*(x+y)+b*(y+z)+c*(z+x)=a*x+a*y+b*y+b*z+c*z+c*x=(a+c)*x+(a+b)*y+(b+c)*z where a,b,c are scalar values

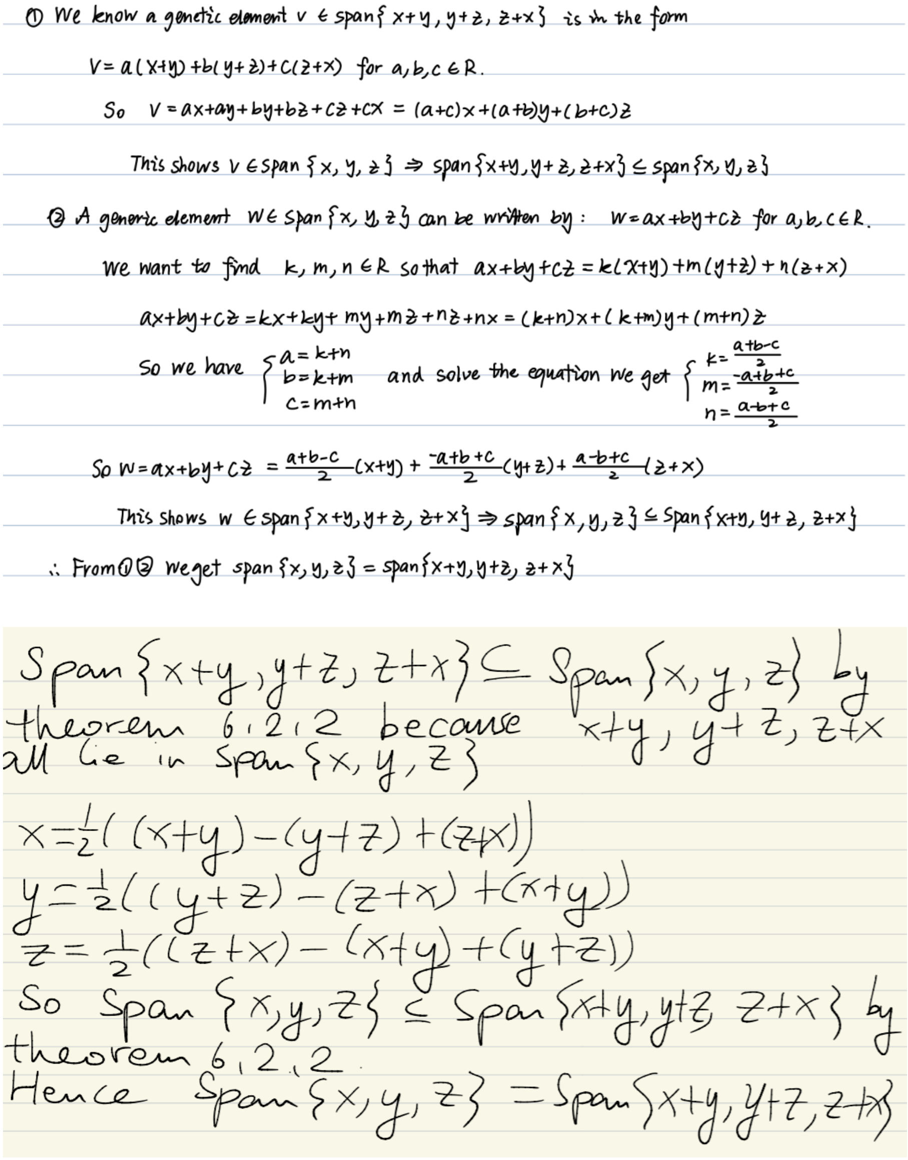
Thus w in is in span{x,y,z}, therefore w is in U and therefore W is a subset of U.

Now to prove U is a subset of W:

Let u be any vector in U, u can be expressed as u=a*x+b*y+c*z where a,b,c are scalar values.

Re-express

Figure 5.7: Two examples for typed responses taken from the study



① We know a genetic element $v \in span\{x+y, y+z, z+x\}$ is in the form

$V = a(x+y)+b(y+z)+c(z+x)$ for $a,b,c \in R$.

So $V = ax+ay+by+bz+cz+cx = (a+c)x+(a+b)y+(b+c)z$

This shows $v \in span\{x, y, z\} \Rightarrow span\{x+y, y+z, z+x\} \subseteq span\{x, y, z\}$

② A generic element $W \in span\{x, y, z\}$ can be written by: $W = ax+by+cz$ for $a,b,c \in R$.

We want to find $k, m, n \in R$ so that $ax+by+cz = k(x+y)+m(y+z)+n(z+x)$

$ax+by+cz = kx+ky+my+mz+nz+nx = (k+n)x+(k+m)y+(m+n)z$

So we have $\begin{cases} a=k+n \\ b=k+m \\ c=m+n \end{cases}$ and solve the equation we get $\begin{cases} k=\frac{a+b-c}{2} \\ m=\frac{-a+b+c}{2} \\ n=\frac{a-b+c}{2} \end{cases}$

So $W = ax+by+cz = \frac{a+b-c}{2}(x+y) + \frac{-a+b+c}{2}(y+z) + \frac{a-b+c}{2}(z+x)$

This shows $w \in span\{x+y, y+z, z+x\} \Rightarrow span\{x, y, z\} \subseteq Span\{x+y, y+z, z+x\}$

∴ From ①② we get $span\{x, y, z\} = span\{x+y, y+z, z+x\}$

$Span\{x+y, y+z, z+x\} \subseteq Span\{x, y, z\}$ by theorem 6,2,2 because $x+y$, $y+z$, $z+x$ all lie in $Span\{x, y, z\}$

$x = \frac{1}{2}((x+y)-(y+z)+(z+x))$

$y = \frac{1}{2}((y+z)-(z+x)+(x+y))$

$z = \frac{1}{2}((z+x)-(x+y)+(y+z))$

So $Span\{x, y, z\} \subseteq Span\{x+y, y+z, z+x\}$ by theorem 6,2,2

Hence $Span\{x, y, z\} = Span\{x+y, y+z, z+x\}$

Figure 5.8: Two examples of uploaded responses taken from the study

		Total
Benefit	Easier to write mathematics symbols	41 (48%)
	Familiarity	17 (20%)
	Express what I'm thinking about	16 (19%)
	Other : taking notes, sketches, spotting mistakes, and reading the work	5 (6%)
Drawback	Time consuming	8 (9.3%)

Table 5.4: Percentages of participants' views on the benefits and drawback of uploading handwritten responses

Participants' view

Students were asked immediately after completing the two questions to choose one of the following statement and to justify their answer.

- I prefer typing my mathematical response directly online rather than handwriting, scanning, and uploading.
- I prefer to handwrite, scan, and upload my mathematical response rather than type it directly online.

The vast majority of students (78 of 86; 92%) preferred to handwrite, scan and upload their mathematical responses. While only 8 out of 86 (9.3%) participants preferred to type their responses directly online.

As the question was open-ended, the collected data were of qualitative nature. In order to analyse qualitative data, the responses need to be categorised by identifying common themes. Content analysis was chosen for this purpose. The basic idea of content analysis is to take textual material and analyse, reduce and summarize it using emergent themes (Ingleby 2012). These themes can then be quantified and hence content analysis is suitable for transforming rich data into a form which can be statistically analysed. Most students mentioned several drawbacks and/or benefits, and these comments were split accordingly based on the benefits and drawback of each format. I coded the responses according to the themes that they contained. In order to try to maintain objectivity and avoid bias with qualitative data analysis, I reviewed the identified themes and coding with Chris Sangwin, and resolved disagreements by consensus. Percentages were calculated and presented in total. The analysis revealed (four) main categories describing the benefits of uploading handwritten responses. Categories that were represented in less than 10% of student answers were combined into a single 'other' category. Table 5.4 summarised students' views on the benefit of handwriting.

The most common benefits of uploading handwriting responses was dealing with mathematical symbols. This benefit was predicted by this study, and has been confirmed. Just under half of the study's participants (41 of 86; 48%) preferred handwriting especially when writing mathematical symbols and notations.

> It is easier to handwrite mathematical symbols as using the keyboard for certain unique symbols makes us slow and confused.

> It is easier to write matrices and row operations and makes what I am trying to say easier to express

Within this broad finding, students preferred to handwrite mathematical symbols for different reasons. Half of those who preferred handwriting suggested that they preferred handwriting because they found typing mathematical notations hard, and they are slow in typing.

> It is very hard to type the symbols with keyboard and also I am not fast in typing.

Two of those who preferred handwriting suggested it was easier to handwrite mathematical notation especially when under a time limit (which was not the case during this study.)

> I find it easier to write out the response as it is difficult to use mathematical notation when typing, especially under a time limit.

The second common benefit of handwriting that participants mentioned was that handwriting really helps to express what you think about (16 of 86 ; 19 %).

> It is a lot easier to get your ideas across by writing them by hand otherwise it takes too long to set out your answers correctly

> It's often a lot easier to write what you're thinking on paper, as you can easily have arrows pointing to parts of your solution (as an example)

> Handwriting the solution instead of typing it out gives me more freedom in formulating my reasoning.

In addition, participants reported that the familiarity of handwriting was a benefit, suggesting that the unfamiliarity of typing could be a potential drawback. The familiarity aspect was found to be related either to what students themselves are used to (17 of 86; 20%)

> It is easier to write the calculations and feels more familiar writing in pen.

> I am used to working on paper. Editing equations is quite tedious in Word.

Students also referred to their lack of experience in typing (10 of 86; 12%).

> It's harder type the symbols and matrices (especially under timed conditions). I am not good at electric technology.

> I am more comfortable with handwriting. I have never typed my mathematical response before.

> When typing responses in mathematics it slows you down as you need special characters etc. . I wouldn't know how to type a matrix or column vector for example. When I typed my answer I felt like I spent more time getting the subscripts to go right than I actually spent doing the maths. I also didn't know how to get the improper subset symbol which would have made my life easier.

> I still prefer handwritten, since I am really slow when typing math. I can not find how to insert a formula and low speed of typing definitely influences me. I am afraid I can not finish the test in 2 hours if every question needs me to type my proof or solution in many words.

One common drawback of typing, as mentioned by the participants, is that typing is time consuming (18 of 86 ; 21%).

> Typing would be very time consuming as I would have to spell out certain symbols such as 'lambda' or 'sqrt' etc and it can easily get quite messy.

Although speed was considered one of the main benefits of the uploading handwritten approach, some students also found it a drawback especially when students need to scan and upload their response online. For those who preferred to type their mathematical responses, they mentioned the fact that uploading a handwriting response is time consuming (8 of 86; 9.3%).

> It takes me a long time to take a picture, crop it, send it to my laptop and then upload. I also have very bad handwriting.

Four of theses suggested they agreed that the copy, or paste functions while typing may actually save times

> Copy repetitive lines and paste them is an effective time-saver. Although the mathematical symbols expressions need time to learn.

5.3 Discussion

In this study, two formats are considered to compare students' mathematical responses: i) uploading handwritten mathematical response, and ii) typing the response online using the editor provided in Moodle. The research questions, detailed in Section 5.1.1, consider quantitative measures such as whether there is a difference in performance, how much work students do in each format, and a quantitative measure of the types of units used by students. Qualitative questions consider students' attitudes toward the formats when writing mathematical responses.

I again applied the units coding scheme described in Section 3.2, and a secondary outcome is to gain experience of, and illustrate use of, how to use the coding scheme. I believe the coding scheme is a useful tool to understand the structure of students' writing more generally. The units coding scheme was adapted from Hodds (2014) and Toulmin (1958), and combined with ideas of structured derivations from R. J. Back (2010) to describe students' arguments. I now contrast the findings of the analysis of students' responses on the two formats with references to the aspects of the coding scheme.

1. Is there a difference in *performance* between uploading handwritten and typing in writing mathematical responses?

   First, the findings show that participants in the uploading group have scored significantly higher than those in the typing group. I suggest this is due to familiarity: students are more familiar with handwriting where they can derive and express their work freely with no constrains or limitation.

2. How much work a student done in each format?

   The results show that the uploading group generated significantly higher number of derivations, which I coded as units, than those in the typing group. A possible reason for this is that the constraining interface when typing using the editor might limit students' ability to express their reasoning freely using the mathematical symbols as well as some lack of familiarity in typing.

3. Is there a correlation between the number of units and performance?

   The results suggest that the number of units positively correlated with marks. The uploading groups scored higher and accordingly generated more derivations and units.

4. Is there a difference in the number of justification between the two formats?

   In general, there was no significant difference between the number of the justifications given by the two formats. The goal-driven warrants was noticed to be the only justification that was associated with the number of units and scores, and were generated more by the uploading group. However, participants in the typing group produced more principle-based and noticing warrants than those in the typing group. A possible reason is that students in the uploading group tend to provide more derivations and explanations that associated with the proof structure. On the other hand, participants in the typing group were provided more sentences classified as explanations than typing mathematics derivations so they use more principle based and noticing warrants in their responses.

5. What are students' attitudes toward the formats when writing mathematical responses?

   As expected, the majority of students preferred to upload and hand-write mathematical responses than to type. The most common features of uploading as mentioned by participants were dealing

> with mathematical symbols and notations and familiarity. The small number of respondents who preferred to type, especially in comparison with the number who preferred to hand-write and upload, means the results must be viewed with a degree of caution. The students who preferred to type are potentially more confident users of technology than their peers. Those perhaps already had sufficient familiarity with typing mathematical responses. Much of the concern expressed by students about moving to typing was due to their reactions to the time constraint which was inline with (Mogey, Cowey, et al. 2012). We want our students to produce and show their best work. Why then impose a time limit that causes them to worry, adjust their strategies, and potentially under-perform?

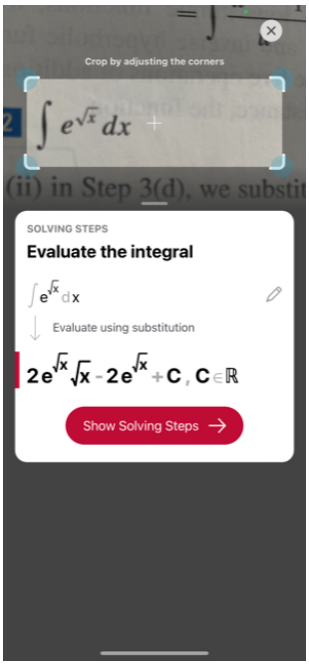
This study has demonstrated that students who wrote by hand have, in general, written substantially more than students who typed. However, a further issue here is general familiarity with technology for typing mathematical symbols. It was noticed that as an effect of typing, more than a third (41%) of the participants in the typing group spelled out the mathematical notations instead of using their symbols. The words “subset”, “contains”, “belongs”, and “equivalent” were used more frequently in the typed responses. Two screenshot of typical typed responses are shown in Figure 5.7.

As shown in Figure 5.7, the overall style of the responses is best described as as a paragraph. Students typed in lines without reasoning or using equivalent relationships. Notice that students also used to spell out or typed some words, for example, the word “supset”. Although this is not a substantial problem, it is an interface issue. This might be due to lack of familiarity with typing using the editor and also the time limit as indicated by students. Some students mentioned in the questionnaire that typing responses in mathematics slows them down as they need special characters which is a distraction from the “flow” of thinking about mathematics. Certainly, a short structured training session using LaTeX might be helpful at the beginning of each mathematics course. A link for how to use an online LaTeX editor could be provided before the first lecture of a mathematics course where students can practice typing. Students might also be encouraged to type a few equations at the first tutorial workshop or to type the first assignment of the course to be more familiar and confident with typing. According to (Mogey, Cowey, et al. 2012), any move to more widespread adoption of typing should be supported by practice, since without previous experience in a like situation many students will prefer to stick with what they used to. Less experienced typists might find it helpful to focus some of their practice time on getting a feel for roughly how much content they can type in a fixed time. However, the problems is much more than an issue regarding training and practicing. There is still no uniformly way students can type traditional mathematics beyond LaTeX , which is only presentation of written mathematics in any case, and LaTeX never sought to encode the actual meaning unambiguously.

One of the significant barriers faced by mathematics students is the heavy use of special symbolism, and the two-dimensional traditional layout of symbols on a page. While professional mathematicians typeset their work with systems like LaTeX this is difficult to learn and the typesetting process can interrupt the train of thought. Knuth’s goal in the design of TeX was to replicate movable type (i.e. the arrangement of boxes containing symbols on a printed page) rather than encoding mathematical meaning which could be automatically checked. This is different than CAS and automatic theorem proving which are all about mathematical meaning. At the time Knuth wrote TeX computers encoded symbols with a very limited character set, ASCII the *American Standard Code for Information Interchange*. Unfortunately, most computer keyboards only contain a very limited range of symbols as keystrokes, typically letters, numbers and a small range of others, such as brackets/parentheses and punctuation (essentially ASCII for English-speaking keyboards). These users would still require software tools to generate the Unicode symbols. It is surprising that even in 2021 there is no widespread system for typing mathematics comes after LaTeX nor is there agreement on how to enter short mathematical text in an online system.

In addition, online systems are varied in term of how to solve a mathematical problem. While there are systems that let students to type or upload their own handwritten solutions (e.g STACK, Gradescope, SOWISO). There are systems or software that can be used to read and solve mathematical expressions using smartphone camera in real time (e.g. Photomath, 2014). Figure 5.9 illustrates the idea of Photomath, where we can photo a question and the system can provide the solution.

In this study, two formats were considered for comparison; (1) scanned handwriting and uploading, and (2) using an editor as part of a web browser and typing directly online. The reasons for choosing the



Evaluate the integral

$\int e^{\sqrt{x}} dx$

Evaluate using substitution

$2e^{\sqrt{x}}\sqrt{x} - 2e^{\sqrt{x}} + C, C \in \mathbb{R}$

Figure 5.9: An illustration of Photomath

two formats is that students are used to making regular use of both, and so this study is really about typing vs. writing, rather than learning LaTeX . Many people find any change challenging. I do not wish to confound the possible format effect with an effect due to either lack of familiarity (e.g. LaTeX), or to resistance to change.

All students who participated can already type so the study design does not need training materials in LaTeX for participants. The students are also familiar with uploading their work online. It was expected that the handwritten responses would make extensive use of arrows, sketches and notes. For this reason the handwritten responses were kept in their original form without typing them for markers. The markers were not blind in marking the responses.

There are some important limitations of this study. This study has only considered a modest sample of students (N = 86) from one discipline area, where essay question type may be approached differently from other subjects. Nevertheless, I cannot claim that the sample is representative of the broader population of students undertaking mathematics modules at universities around the world.

Additionally, to avoid confounding potential format effects with those arising from unfamiliarity with tools like LaTeX or initial resistance to change, we opted not to explicitly control for these factors. Future studies could address this limitation by using more diverse materials and samples. This study did not explicitly control for prior knowledge, which could have acted as a potential confounding factor and influenced the observed relationship between the format effect and students' performance. Students with varying prior knowledge could have experienced different levels of difficulty with the tasks, potentially inflating results in either format. While we acknowledge this limitation, recognizing the need for future research to explicitly control for prior knowledge using methods like pretests or matched groups. It is important to acknowledge that this study is also potentially subject to self-selection bias. Since not all students volunteered to participate, those who participated might have more confident in their mathematical abilities, potentially leading to better performance regardless of format. While acknowledging this potential bias, planning future studies with more representative samples can help reduce its impact on the validity of the findings. Students could be involve in different ways, maybe even choosing them randomly, so the findings reflect the whole class better.

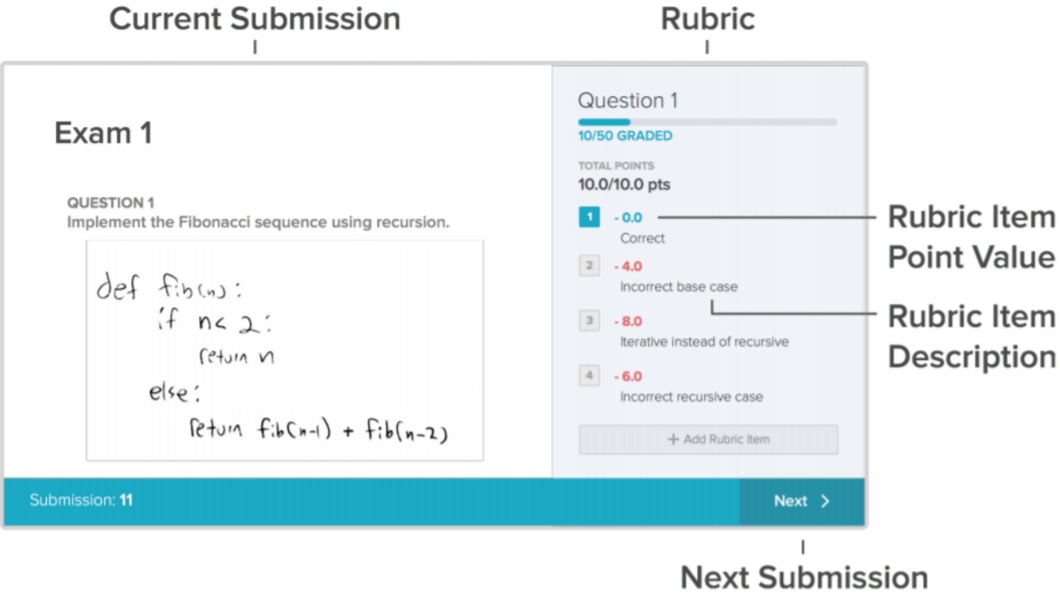
5.4 Conclusion

The aim of this study was to identify and explore any systematic differences that may be introduced between typing and uploading handwritten mathematical responses to simple coursework tasks. Participants responded to two relatively short problems on the same topic “Subspaces and Spanning” from the course Introduction to Linear Algebra taught in the first semester of 2020/21, containing equivalent typing and uploading handwritten items, and students’ reactions immediately after the tasks were obtained. These were then marked by two markers. Typing and uploading responses were distributed equally between the two markers so each marker dealt with both formats.

Factors explored included overall score awarded, justifications provided, and number of steps or “units”. The analysis of the data obtained from the study indicated that, overall scores were higher in the uploading handwritten group compared to the typing group.

In this work, a coding scheme 3 developed by the author and adapted from (Hodds, Alcock, and Inglis 2014) and (Toulmin 1958), was combined with ideas of structured derivations from (R. J. Back 2010), to describe students’ arguments. The coding scheme provided a useful tool to understand students’ writing more generally. It was found that students in the uploading handwritten group generated significantly more units than did the typing group. As a result, it was found that a higher level of marks was associated with a higher number of units. Students who wrote more tended to get slightly more marks.

While most college-aged students in today’s world know how to operate computers and type, many may not know the proper techniques and finger strokes for maximum typing speed and accuracy especially in typing mathematics. With the increase use of submitting online work, a short LaTeX training would be perhaps helpful for students at the beginning of each mathematics course. It is also possible for teachers to encourage students to use LaTeX for their first assignment. Therefore, students might become more confident and familiar with notations typing mathematics. Note that this study took place before the widespread change from COVID-19. According to Alarfaj, O’Hagan, and Sangwin (2022), electronic submission of written work (human marked) saw significant increases in use in 2020/21 compared with before the COVID-19 pandemic (See Chapter **??**). Now students are much more familiar with photographing and uploading responses. For example, Gradescope is a system for online assessment of handwritten homework assignments and exams. Students can scan and upload their homework or exams which will be graded manually. Figure 5.10 illustrates an example for the grading interface in Gradescape. The interface is divided in two parts. In the left part, a single student’s submission can be seen by the grader to the question they’re grading. On the right part, the rubric which is composed of multiple rubric items that each have point values and descriptions.



Current Submission

Rubric

Rubric Item Point Value

Rubric Item Description

Next Submission

Exam 1

QUESTION 1
Implement the Fibonacci sequence using recursion.

```
def fib(n):
    if n< 2:
        return n
    else:
        return fib(n-1) + fib(n-2)
```

Question 1

10/50 GRADED

TOTAL POINTS
10.0/10.0 pts

1 - 0.0
Correct

2 - 4.0
Incorrect base case

3 - 8.0
Iterative instead of recursive

4 - 6.0
Incorrect recursive case

Figure 5.10: Screenshot of Gradescope assessment system

Chapter 6

Mathematical induction and Separated Concerns

Learning to write proofs is a key intended outcome of most mathematics curricula in the later stages of school SQA Advanced Higher Mathematics includes proof by induction. This chapter starts by discussing some difficulties students have when learning mathematical induction identified by previous educational research. Four concerns in learning mathematical induction were identified. I introduced the phrase "Separated Concern" which I used to describe materials in which potential misconceptions are addressed directly.

6.1 Mathematical induction

The goal of a proof by induction is to show that each member of an infinite family of statements is true. Typically n is a natural number, the statement to be proved is written as $P(n)$ and so the goal is often abbreviated to

$$\forall n \in \mathbb{N}\ P(n).$$

A typical modern proof by induction proceeds in a step-by-step fashion as follows.

1. A clear and explicit statement of the "induction hypothesis", $P(n)$.

2. Prove a "base case" is true. Typically prove the single statement $P(1)$.

3. Prove the "induction step". Typically, if $P(n)$ is true then $P(n+1)$ is also true.

4. Conclude that steps (2) and (3) allow the conclusion that $P(n)$ is true for all natural numbers n by the principal of mathematical induction.

In essence, an induction argument starts from a true statement and then shows the truth is maintained as we proceed forward, step-by-step. The following is a typical example.

Example 1. *Prove that the sum of the first n odd natural numbers equals n^2.*

Proof. Let $P(n)$ be the statement "$\sum_{r=1}^{n}(2r-1) = n^2$".

The sum of the first odd number, 1, equals 1^2 so that $\sum_{r=1}^{1}(2r-1) = 1^2$ and $P(1)$ is true.

Assume that $P(n)$ is true. Then

$$\begin{aligned}\sum_{r=1}^{n+1}(2r-1) &= 2(n+1)-1+\sum_{r=1}^{n}(2r-1)\\ &= n^2+2n+1=(n+1)^2.\end{aligned}$$

Hence $P(n+1)$ is also true.

Since $P(1)$ and $P(n) \Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n \in \mathbb{N}$ by the principal of mathematical induction.

In preparation for this thesis, I reviewed the mathematics of induction and examples of questions from SQA examination. These were published as an article for school teacher in Alarfaj and Sangwin (2020).

6.2 Difficulties in learning mathematical induction from the educational research

The educational challenges of learning induction have been widely studied, perhaps because induction is often a student's first formal encounter with the concept of infinity. Young (1908) commented that *"The process of mathematical induction is exceptionally well fitted to introduce the beginner to the philosophic study of mathematical thinking, [...]"* Furthermore, it is often some students' first attempts to develop rigours proofs beyond algebraic reasoning by equivalence. Induction is complicated, and it is unusual in that as a form of proof it is explicitly taught. Through the years, researchers have reported several difficulties that students are facing with proofs by induction ((Ernest 1984), (Baker 1996), (Thompson 1996), and (Michaelson 2008)) and some have proposed solutions to mend the situation (e.g. Ernest (1984)), but little improvement has been achieved. Ernest (1984) identified the following misconceptions, or conceptual difficulties, in the learning of induction with some possible solutions that might help.

One of the typical misconceptions among students with mathematical induction is that you assume what you want to prove, and then prove it, and that *"it has a suspicious likeness to assuming what you have to prove"* (Newman and et.al. 1957). That observation illustrates that the student does not understand clearly the structure of a proof by induction, in particular does not understand the implication $P(n) \Rightarrow P(n+1)$ and how to connect the base case to the induction step. The first remedy, suggested by Ernest (1984), is that students should learn both the meaning and the methods of proof of implication statements. Second, reducing the Principle of Mathematical Induction by expressing it in a two-variable form. For example,

If $P(1)$, and if all for all k $P(k)$ implies $P(k+1)$, then $P(n+1)$ for all n.

In this form the variable k occurring in the inductive hypothesis is localised to the induction step. This makes it easier for the student to understand that the inductive hypothesis is only assumed for the purpose of the induction step and that the assumption is used during this step. Another advantage of the two variable form is that it gives a sensible statements such as "if the proposition is true when $n = k$, then it is true when $n = k + 1$."

A further misconception sometimes arises from the view that one of the elements of an inductive proof is not really important, especially the base step (Ernest 1984). That is to say, students do not always appreciate why each separate part of the proof is necessary (Ernest 1984).

There is a difficulty to distinguish between the heuristic inductive method and the formal proof method of mathematical induction. Confusion between heuristic induction and formal mathematical induction can cause a great doubt in a student's mind (Ernest 1984). To avoid the above difficulty, Ernest (1984) recommended to explain carefully the difference between the two methods for students, and that the former method be referred to as the method of generalization or any other name other than induction.

One of the conceptual difficulties arises from the fact that the logical form of the principle is complicated (Ernest 1984). For example, some students encounter difficulties interpreting complex statements. To avoid this problem, Ernest (1984) suggested that students need to practice using more complicated types of statements.

One of the misconceptions that occurs is that students think proof by induction is only a tool to use when summing finite series (Ernest 1984). One solution for this problem is to provide students with a more variety of problems and exercises than those involving the sums of finite series (Inglis and Mejía-Ramos 2009).

The final issue in learning mathematical induction is not considered as a misconception, but a lack of understanding of the basic motivation and necessity (Ernest 1984). Many students experiencing the method of mathematical induction wonder why this rather complicated and apparently arbitrary principle is adopted. Unlike many principles, mathematical induction is neither a generalisation of previous more elementary experience nor self evidently necessary.

Baker (1996) investigated the difficulties encountered by high school and college students who started to learn the proof of mathematical induction. He found that many students did not describe mathematical induction in term of its concepts. In particular, he claimed there were four missing components from previous studies. These were (i) the effect of students' prior experiences, (ii) relationships between conceptual, procedural and application knowledge, (iii) a lack of understanding of the relationship between resources and cognitive factors and (iv) evidence on everyday reasoning have not been applied to learning proof by induction.

Empirical evidence suggested that students tend to concentrate their cognitive attention on the procedure of mathematical induction rather than on concepts and applications. Thompson (1996) noted that students apply induction in a mechanical way which was in line with Baker (1996). Baker (1996) found that students focus on the procedure they have to follow, instead of the substance of the proof, and they are not able to describe the proof in terms of its concepts. According to Thompson (1996), sometimes "students appear to be approaching the proof process very algorithmically, having memorized a process instead of understanding its origin".

In a study of both high school and undergraduate university students, Baker (1996) claimed that an "insufficient formal mathematics background" and "a lack of mathematical content knowledge" are the main factors that contribute most significantly to students' inability to construct a proof by mathematical induction. He recognised that some students do not know how to use the sigma notation correctly and are unable to realise basic algebraic arguments, such as $2^{k+1} = 2 \times 2^k$. Aviatl and Libeskind (1978) also commented more generally on the issue that students experienced in the algebraic manipulations of the induction step. Likewise, Ernest (1984) noted that substituting $k+1$ for n in the induction step is one of the main issues for many students in mathematical induction. For example, consider the induction step of the proof by mathematical induction that

$$\sum_{r=1}^{n} r^2 = \frac{n(n+1)(2n+1)}{6}$$

After establishing the supposition that

$$\sum_{r=1}^{k} r^2 = \frac{k(k+1)(2k+1)}{6}$$

for any $k \geq 1$,

$$\sum_{r=1}^{k+1} r^2 = \frac{(k+1)((k+1)+1)(2(k+1)+1)}{6}$$

It is required to place the brackets around each $k+1$ in the first and third factors and also provided in the second factor for consistency and emphasis only. Without using the brackets, students will use the distributive property of algebra to multiply $k+1$ properly (Michaelson 2008). The experience leads to maturity and that's mean you do not need to expand the brackets.

Indeed, one of the criticisms of proof by induction is that as a process it can become formulaic, even mindless which is the same criticism of two-column proof. The structured form of a proof by induction is both a strength and a weakness of proof by induction. By having a formal written proof format, we can establish the correctness (or otherwise) of a proof. The decision to use a specified formal format removes one worry in deciding "is this a correct proof"? On the other hand, anything which is formulaic can easily become mindless and this risks a retrograde step in helping students develop a deep understanding of both the mathematics being proved, and an appreciation of the legitimacy of the form of proof.

Many of the findings from educational research are also found in comments from official SQA Course Reports. In 2019, *"Many candidates found great difficulty in communicating explanations and logical processes."* The issue of communication was also mentioned in 2015. *"Question 5: it was disappointing that the proof by induction did not perform as well as expected – especially the last two marks. There appeared to be a lack of understanding, formality and rigour in both the process and communication."* The report in 2018 said the following about induction.

> Many candidates omitted important elements of the logic, including a correct statement of the inductive hypothesis. In a number of cases, candidates wrote down what they were attempting to prove instead of attempting the inductive step. Most candidates found the algebraic manipulation challenging.

6.3 Concerns in learning mathematical induction

In this thesis, I use the word "concern" for a specific issue students have when learning induction. Four concerns are selected from the issues previously identified from research on learning mathematical induction. This section discussed the four concerns. An example is provided at the end of this chapter.

6.3.1 Concern 1 : Basic algebraic manipulation

The first concern is algebraic manipulation, especially unusual cases. According to Alarfaj and Sangwin (2020), some students have difficulties with basic algebraic manipulations, especially when new constructs are introduced such as factorial notation and dealing with binomial coefficients.

For example, consider trying to prove the statement that "$1.1! + 2.2! + 3.3! + \cdots + n.n! = (n+1)!$".

The induction step involves the following sort of manipulation $(n+2)! - (n+1)! = (n+1)(n+1)!$.

This seems particularly difficult for many groups of students. Quite often such manipulations are needed in the induction step to show that $P(n+1)$ holds.

Some students are also less confident using the factored form. For example, when proving a formula such as

$$\sum_{k=1}^{n} k(k+1)(k+2) = \frac{n(n+1)(n+2)(n+3)}{4}.$$

the temptation for many students seems to be to expand out all the brackets. For example, partway through the induction step they are faced with the following

$$\frac{n(n+1)(n+2)(n+3)}{4} + (n+1)(n+2)(n+3).$$

The most efficient route would be to factorise by taking $(n+1)(n+2)(n+3)$ as a common factor which immediately result in

$$\frac{(n+1)(n+2)(n+3)\,(n+4)}{4}.$$

However, this requires quite a lot of confidence in taking out common factors. Of course, if they expand instead here they are going to end up with something quite hideous like

$$\frac{n^4}{4} + \frac{5\,n^3}{2} + \frac{35\,n^2}{4} + \frac{25\,n}{2} + 6.$$

From here they somehow have to factor this fourth order polynomial containing fractions.

Michaelson (2008) classified algebraic manipulation as technical difficulties with mathematical induction. He concluded that students without strong mathematical backgrounds will have difficulties with mathematical induction, since they tend to make technical and mathematical errors that are problematic when they construct proof by induction. He concluded that the concept of mathematical induction can only be fully comprehended once the technical difficulties are removed. Otherwise, students will routinely follow a prescribed script for writing mathematical induction proofs without really understanding the process (Z. Michalewicz and M. Michalewicz 2008).

Students previous experience of algebra might also be restricted to solving. They may have little experience of rearranging to a form in a variable $k+1$.

6.3.2 Concern 2: Dealing with sigma notation

The symbol $\sum$ (capital sigma) is often used as shorthand notation of writing infinite numbers of terms in a sequence. Sigma notation $\sum$ is covered in school mathematics in upper secondary schools in many jurisdictions, for example the International Baccalaureate Diploma Program, see (Haese et al. 2019, p. 112-113). However, sigma notation is relatively new for year 1 undergraduate students, and experience suggests students struggle to understand and use sigma notation effectively. The following is a (non-exhaustive) list of issues we know, from our teaching experience, that might test students' understanding.

2.1 Write out the terms of this sum in full, i.e. without "sigma notation" for a short example. E.g. $\sum_{k=4}^{7} n$

$$\sum_{k=4}^{7} n \; = 4+5+6+7.$$

Conversely, spotting a pattern from a sum and providing the general formula within sigma notation.

2.2 Paying close attention to variables in superscript and subscripts. E.g.

$$\begin{aligned}\sum_{k=1}^{4} n^2 \; &= n^2+n^2+n^2+n^2 \\ &\neq 1^2+2^2+3^2+4^2\end{aligned}$$

in which the variable is k but the formula a_k to be summed only has the different variable n in it.

2.3 Manipulating expressions where all the "action" happens in superscript and subscripts:

$$\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k = a_{n+1}$$

$$\sum_{k=0}^{n} a_k - \sum_{k=1}^{n} a_k = a_0$$

As specific examples, ask students to calculate the following

$$\sum_{k=1}^{n+1} \frac{k^2}{k^r} - \sum_{k=1}^{n} \frac{k^2}{k^r}$$

$$\sum_{k=0}^{n} \frac{(k+1)!}{2^k} - \sum_{k=1}^{n} \frac{(k+1)!}{2^k}$$

2.4 Making a substitution of the range of summation, e.g. summing from 0 to $n-1$ instead of from 1 to n. For example, $\sum_{k=1}^{n}(2n-1) = \sum_{k=0}^{n-1}(2n+1)$.

6.3.3 Concern 3: State $P(n+1)$

I found that students sometimes show fragile understanding of what the statement $P(n)$ actually is. E.g. when proving that for $n \in \mathbb{N}$

$$\sum_{k=1}^{n}(2k-1) = n^2$$

some students think that $P(n+1)$ is $(n+1)^2$. That is to say, they confuse the right hand side of the equation (in this case the value of the sum) with the whole statement. Do students know the whole statement $P(n+1)$ is the goal as the conclusion of the induction step? Do they know what is mean to assume $P(n)$ as a hypothesis? One way to test students' understanding is to also then to write the induction hypothesis in full for a specific example, e.g. $n = 3$ without sigma notation.

It's worth noting that Concerns 2 and 3 don't directly about mathematical induction rather than a specific issue with mathematical induction. Reviewing existing literature in mathematical induction suggests that students sometimes struggle with understanding sigma notation or applying algebraic manipulations rather than the core concept of induction. This aligns with the purpose of separating concerns: by isolating them, we can identify obstacles that hinder students from focusing solely on mastering mathematical induction itself.

6.3.4 Concern 4: Structure of the proof and identifying the parts of the proof

Can students identify the overall structure and parts of the proof?

4.1 Identify base case, and appreciate the importance of the base case.

An example of a mathematical mistake in the base case was provided by Aviatl and Libeskind (1978). Since $n^2 < 2^n$ is true for $n = 1$, but not for $n = 2$, 3 and 4, it is possible for the student to establish the base case for all $n \geq 1$ and then complete the inductive step correctly without recognising that the base case should have been established for all $n \geq 5$.

4.2 Recognizing and writing the induction hypothesis. E.g. Which line states the induction hypothesis? Where is the inductive hypothesis used in the inductive step?

4.3 Identify the start and the end of the inductive step.

6.4 Separated Concerns

In the previous section 6.3, four concerns in learning mathematical induction were identified based on prior research and teaching experience. In this section, I introduced the phrase "Separated Concern" which I used to describe materials in which potential misconceptions are addressed directly.

The idea of Separated Concerns is influenced by faded worked examples. As stated in the literature review, faded worked examples reduce working memory load since search is reduced or eliminated, and attention (i.e., working memory resources) are directed towards learning the essential relation between problem-solving moves (Kirschner, Sweller, and Clark 2006). The phrase "faded worked examples" refers to a progressive sequence of worked examples in which the scaffolding provided by the solutions steps within each worked example are systematically removed, requiring students to take progressive responsibility for completing the problem (Renkl, Atkinson, and Gross 2004). The research cited in the literature review does confirm that faded worked examples lead to superior performance and transfer relative to conventional problems and that it is useful to use carefully designed faded worked examples before starting to solve problems independently.

An alternative strategy is to identify issues which students typically find difficult and create questions which "Separate Concerns" and address each concern directly. Ernest (1984) found that some students encounter difficulties interpreting complex statements and to avoid this problem, students need to practice using more complicated types of statements. Such practice, outside a proof, is an example of what I mean by "Separated Concerns". For example, it has been found that one of the common misconceptions in learning mathematical induction was dealing with sigma notation (Baker 1996). So, by Separating Concerns, a question is designed involving the use of sigma notation with the expectation that a student will become generally more familiar and confident before they are asked to use this notation in induction proof. A difference between faded worked examples and Separated Concerns relates to the direct relevance. Whats new in Separated Concerns is that a teacher might choose to use a task which is related to, but not directly applicable when solving, the more complex problem. However, faded worked examples only deal with removing correct steps from a complete argument. Therefore, faded worked examples cannot directly address what happens when things go wrong. For example, by giving examples of statements which are false in general, but for which it is easy to prove $P(n) \Rightarrow P(n+1)$, you can illustrate the importance of the base case. Take any true statement of the form $\sum_{k=1}^{n} a_k = p(n)$ and add 1 to the right hand side. Therefore Separating Concerns, like faded worked examples, is a form of scaffolding. This does not automatically result in atomised, behaviourist teaching, since the longer-term goal of the task sequence is independent successful completion of the whole task by students. Fading worked examples cannot address structural issues affecting the whole argument, such as which type of proof to select. Essentially, the task of teaching students how to write mathematical proof is simply too large to apply fading to effectively. Fading certainly has a valuable place, but Separating Concerns is also needed.

Separating Concerns is a specific issue which explicitly identified and addressed in advance of using it in a more substantial application. This is different from testing the students' knowledge of a topic for its own sake.

Separating Concerns might create a systemic environment in which misconceptions may be easier to spot before learning more complex materials, which is the same criticism of using any constraint format (e.g., faded worked examples.)

By Separating Concerns, students can receive more specific feedback on the related concerns that the question target. Feedback provided to learners during problem solving is a common form of instructional support. (Krause, Stark, and Mandl 2009) found that low prior knowledge students learning statistics benefits more from receiving explicit feedback in problem-solving practice. However, no effect of feedback was found for students with higher prior knowledge.

Based on cognitive load theory, there are cognitive aspects linked to the use of Separating Concerns, which I believe are potentially beneficial. Since learning is strongly effected by the potential of working memory, it has been argued that working memory should be filled by task-relevant operations, especially in learning complex material (Van Gerven et al. 2002). The effect of Separated Concerns occurred when students can transfer and apply their understanding to solve a similar problem in more complicated tasks. By Separating Concerns, students can achieve success on the individual components with explicit and conscious knowledge of where these fit into the more complicated tasks. For example, I separate out tasks involving $\sum$ (sigma) notation so that student will become generally more familiar, confident and competent before they are asked to use this notation in an induction proof.

Previous research suggested that its important not to generate high extraneous load, especially when its connected with high intrinsic load, since no cognitive capacity may remain for germane load. The extraneous cognitive load could be reduced by designing effective materials (Sweller, Merriënboer, and Paas 2019). In Separating Concerns, the extraneous cognitive load could be reduced when designing a task that addresses each concern directly. Separating Concerns could lead to an efficient construction of cognitive load, because they focus the attention of learners on specific concerns. For enhancing cognitive skills acquisition, I believe that it is useful to use Separated Concerns before starting to solve more complicated problems.

Back's (2010) structured derivations could be a valuable tool for teachers to identify and separate students' concerns during proof construction. Writing proofs using this framework explicitly distinguishes between warrants (justifications) and computational steps, potentially helping teachers in separating out specific concerns faced by students.

In this thesis, four concerns were discussed in Section 6.3 when learning proof, in particular a proof by induction. The following example illustrates how to write an induction question based on Separated Concerns.

6.4.1 Example

Example 2. *Let $P(n)$ be the statement*

$$\sum_{k=1}^{n} (2k-1)^2 = \frac{n \cdot (2n-1) \cdot (2n+1)}{3}$$

.

1. *Write the statement $P(n+1)$.*
2. *Calculate*
$$\sum_{k=1}^{n+1} (2k-1)^2 - \sum_{k=1}^{n} (2k-1)^2$$
writing your answer in simplified form.
3. *Calculate*
$$\frac{(n+1) \cdot (2(n+1)-1) \cdot (2(n+1)+1)}{3} - \frac{n \cdot (2n-1) \cdot (2n+1)}{3} \tag{6.1}$$

There are three items in this example, where each item addresses one of the concerns in learning mathematical induction that is discussed earlier in this chapter. Hence, the scaffolding provided here is described as "Separated Concerns". In particular, Part 1. of this question addresses concern C3, i.e. understanding what the induction statement $P(n+1)$ actually *is*. Part 2. addresses concern C2.3, which designed to address dealing with sigma notation and how students manipulate an expression where all the action happen in the superscript and subscripts. Part 3. addresses C1: basic algebraic manipulation.

Part 1. is designed to address writing $P(n+1)$, i.e. we expect an answer

$$\sum_{k=1}^{n+1} (2k-1)^2 = \frac{(n+1) \cdot (2(n+1)-1) \cdot (2(n+1)+1)}{3} \tag{6.2}$$

In the context of induction, it is reasonable to assume Equation 6.2 is true, but this isn't really relevant.

Part 2. is designed to address dealing with sigma notation and how students manipulate an expression where all the action happen in the superscript and subscripts. Students were asked to find the difference between the two sums. We expected that students can notice the immediate pattern which is $a_k = (2k-1)^2$ and hence,

$$\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k = a_{n+1} = (2(n+1)-1)^2. \tag{6.3}$$

Part 3. is designed to address some basic algebraic manipulation. Students were asked to calculate the difference between the two fractions. The goal of Part 3. was to help students to recognize that the first fraction of Equation 6.1 is simply the R.H.S of $P(n+1)$, and the second fraction of Equation 6.1 is the R.H.S of $P(n)$ and encourage them to think before doing the calculations long-hand. Adding algebra is mechanical. Noticing things in context is important. The experience leads to maturity and that's mean you do not need to expand the brackets.

Assume Equation 6.2 is true in Part 1. and from noticing the pattern in Part 2. students are expected to combine Part 1. and Part 2. and find the difference in Part 3. without expanding the two fractions.

$$\begin{aligned} \underbrace{\frac{(n+1)\cdot(2(n+1)-1)\cdot(2(n+1)+1)}{3}}_{\text{R.H.S of } P(n+1)} - \underbrace{\frac{n\cdot(2n-1)\cdot(2n+1)}{3}}_{\text{R.H.S of } P(n)} = & \\ &= \sum_{k=1}^{n+1}(2k-1)^2 - \sum_{k=1}^{n}(2k-1)^2 \\ &= (2(n+1)-1)^2 \\ &= (2n+1)^2 \end{aligned} \tag{6.4}$$

Note, the second cohort 2021/22 were required to engage with the lecture quiz and pass with 80%. This change in design was an important part of the process.

Even if students really do the algebraic calculations, this observation should provide a check of the results.

One purpose of this question, in context, was to prepare student to later write a full induction proof of this particular theorem:

$$\sum_{k=1}^{n}(2k-1)^2 = \frac{n(2n-1)\cdot(2n+1)}{3}$$

In the next Chapter 7, this example was used to illustrate how to update STACK potential response tree based on students' concerns in learning induction. The underlying goal of Separating Concern is to develop online assessments by transforming existing (largely paper-based) problem sets into online assessments. In this thesis, I focused on the third misconceptions C3, i.e. understanding what the induction statement $P(n+1)$ actually is. The remaining concerns are broader, and probably better addressed once students have reached a basic competence in writing, and reading, simple proofs by induction.

Chapter 7

Study 3: Updating STACK Potential Response Tree (PRT) based on Separated Concerns

This chapter presents the third of three empirical studies focused on a potential format effect when using online assessments. In this study, I focus exclusively on students' responses, and misconceptions in learning induction proofs. Mathematical induction is used as a vehicle to illustrate the idea of Separated Concerns, discussed in Section 6.4.

The main goal of this chapter is to firstly illustrate how engagement with the lecture quiz related to success on the weekly assessed quiz, and the course total. Second, to explore the common mistakes made by students when using online materials (i.e., STACK questions) to prepare for mathematical induction? Secondary to the above research goals, this study illustrates how to use research to update the algorithms which assess students' answers, known as "STACK potential response trees", in questions written to support learning mathematical induction based on Separated Concerns. Separated Concerns is a phrase used to describe materials in which potential misconceptions are addressed directly. Assessing a full mathematical proof is currently well beyond the capabilities of computer systems, but one possible approach to assessing problem solving has been to break up larger tasks into smaller individual questions to which online assessment can then be applied. Four concerns in learning mathematical induction from prior educational research were discussed in Chapter 6. In this study, STACK questions were designed to test if students exhibit each concern when learning mathematical induction. The intended main advantage of this approach is that we better prepare students to attend class by using online materials. One of the contributions of this study is to improve our general understanding of how to design and use STACK potential response trees.

This study contributes to the field of proof instruction in two key ways. Firstly, it demonstrates the effectiveness of online learning materials in preparing students to learn induction proofs. Secondly, the research explores how the STACK platform's basic question usage report can be used to analyze student responses and identify common misconceptions. This analysis can then be used to refine STACK's potential response trees, leading to more targeted and effective feedback that directly addresses student needs

This study is based on data collected in collaboration with Prof. Chris Sangwin. Students' responses were collected from two quizzes from a university mathematics course run at the University of Edinburgh in 2020/21 and 2021/22. As part of my doctoral research, I independently conducted all analysis presented in this chapter. The work involves a quantitative analysis of the available marks, and a qualitative analysis of students' answers looking for mistakes and misconceptions we know exist from prior research. Part of this research is published on Alarfaj and Sangwin (2022).

7.1 Method

7.1.1 The purpose of the study

Initially, a number of concerns and misconceptions in learning mathematical induction were identified based on prior research and teaching experiences as discussed in Chapter 6. For the purpose of this study, STACK questions were designed based on the identified concerns in *Proof and Problem Solving* (PPS) course (See Appendix C). These STACK questions were used in a lecture quiz to prepare students for the assessed quiz. Two research questions were considered:

1. Broadly, in what way is engagement with the lecture quiz related to (i) success on the weekly assessed quiz, and (ii) the course total? This is a quantitative analysis of the available marks.
2. What are the common mistakes made by students when using online materials (i.e., Separated Concerns in writing STACK questions) to prepare for mathematical induction? This is a qualitative analysis of students' answers looking for mistakes and misconceptions we identified from prior research.

Secondary to the above research goals, this chapter illustrates how to use research to update STACK potential response trees in questions written to support learning mathematical induction based on Separated Concerns. This addressed a more general goal of how to design effective online materials and improve STACK potential response tress.

I am not conducting a random control trial, and so I cannot claim causation. Rather, I am conducting an observations study. Correlation between engagement with the lecture quiz and success on the course is expected, and it seems reasonable to suggest that engagement with the lecture quiz is responsible, in part, for students' success but I cannot conclude that from this study.

7.1.2 Procedure

In this study, students answered two online quizzes which were parts of the online assessment component of *Proofs and Problem Solving* (PPS) course. The quizzes formed part of material in the second week which comprised Mathematical Induction and Inequalities. Each element of the course was given a unique code, e.g. the lecture quiz in week 2 was coding as L02.

The first quiz (L02) was a formative "lecture quiz." The lecture quiz includes questions which address the concerns summarised in Section 6.2.

The second quiz (Q02) was an "assessed quiz" that students took at the end of the weekly learning cycle. Questions of the assessed quiz were also designed based on the Separated Concerns that identified in the lecture quiz.

In the first cohort 2020/21, taking the lecture quiz was optional. Students could access the assessed quiz without taking the lecture quiz. In the second cohort 2021/22, students were required to complete the lecture quiz, ultimately scoring at least 80%, in order to access the assessed quiz.

Students' responses from those who consented to have their data included, and for this week (only), were then copied to a separate online server for analysis. The consent form is available in Appendix B.3.3

The "lecture quiz" deserves some specific comment. In previous years PPS would have had three traditional 50 min lectures each week. COVID-19 made such lectures impossible. Updating PPS has been influenced by previous work developing fully online courses undertaken in Edinburgh: (Kinnear 2019; Kinnear 2018; Kinnear, Wood, and Gratwick 2021). Rather than pre-recording single lectures to replace hour-long lectures, the "lecture quiz" combines shorter video clips with directly relevant interactive practice exercises. The exercises were automatically marked with STACK or other online assessments, e.g. multiple choice. The design concept is to put the book inside the quiz, with the quiz being the primary vehicle through which students engage with the course, see (Sangwin and Kinnear 2022). The previous online course, described in (Kinnear, Wood, and Gratwick 2021), consisted of four such quizzes per week as the total formative teaching experience for students, with additional online assessments. During semester 1 of 2020/21, *Introduction to Linear Algebra* (ILA), had two “lecture quizes” per week, with one live lecture event. Students expressed a strong preference for live/synchronous teaching, and so (partly) as a response to students’ expectations and requests, PPS for 2020/21 was balanced back towards two live lectures and one lecture quiz per week. The experience with other courses leads me to believe there are many advantages to “lecture quizzes” of this design, even as part of a traditional on-campus courses. This feature is very likely to remain after on-campus teaching resumes post-COVID as an important part of the mix of experiences.

Since proof by induction is part of many advanced courses in school level, many of the students on PPS will have been taught proof by induction before, e.g. see (Alarfaj and Sangwin 2020). I specified four concerns related to mathematical induction, and gave examples of questions which address them. These concerns were based on the misconceptions raised by (Ernest 1984) and others, discussed in Chapter 6.

7.1.3 Task construction

I designed a model -either a lecture or assessed quizzes- consisting of four concerns targeted by the questions. Each question was based on a single concern or misconception. Developing materials for this study was done by working with the course organiser to develop the academic materials. The materials contain a balance between the goals of the course and this research study.

Section 7.1.4 discusses the questions in the two quizzes. The questions were designed with online assessment in mind, and the STACK online assessment system in particular. For completeness, I discuss the implementation details of one STACK question in detail in Section 7.1.5.

7.1.4 Questions which address each concern

This section provides a summary of the materials, showing the two quizzes and which questions test specific concerns. Tables 7.1 and 7.2 list the questions in L02 and Q02 respectively, record which concerns (if any) that question addresses, and describe the style of each question.

Style of questions:

- Faded worked examples: a sequence of questions in which students take progressive responsibility with the goal of independence of solving problems of a very specific class. See Section 2.3.1.
- Proof comprehension: students are asked *about* a given argument. In particular they are not asked to prove or solve themselves, see (Mejia-Ramos, Lew, et al. 2017).
- Separated Concerns: a specific issue is explicitly identified and addressed in advance of using it in a more substantial application (e.g. as part of a proof). This is different from testing the students’ knowledge of a topic for its own sake. Theory of Separated Concerns is proposed in Section 6.4.

Note, week 2 of PPS also included some material on manipulating mathematical inequalities. While interesting, and with concerns of its own, the materials for this study are restricted to mathematical induction.

7.1.5 Developing STACK questions

This section explains briefly how STACK questions work, and describes the data which the system collects to explain the process of qualitative analysis of students’ responses to STACK questions. The process of improving questions includes modifying the response trees to make more subtle judgements in the light of data from actual students’ use.

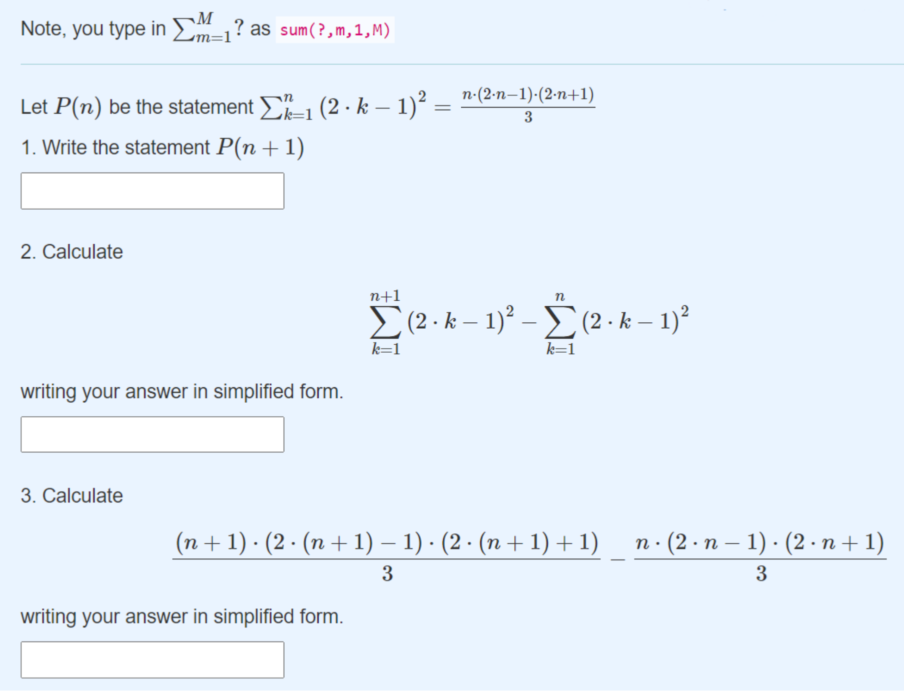
For the purpose of illustration, I use Example 6.4.1 (See page 88) to design a STACK question based on the idea of Separated Concern. This question is used as one of the assessed questions in the Assessed

Question	Topic	Concern	Style
L02.08.1	Base step	C4.1	Faded worked examples
L02.08.2	Base step	C4.1	Faded worked examples
L02.08.3	$\sum_{k=1}^{n+1} a_k = \sum_{k=1}^{n} a_k + a_{n+1}$	C2.3	Faded worked examples
L02.09.1	Identify induction hypothesis	C4.2	Proof comprehension
L02.09.2	Write out $P(3)$ in full	C4.2	Proof comprehension
L02.09.3	Start and end of induction step	C4.3	Proof comprehension
L02.09.4	Use of induction hypothesis	C4.2	Proof comprehension
L02.09.5	Algebraic simplification	N/A	Proof comprehension
L02.10.1	Identify base case	C4.1	Proof comprehension
L02.10.2	Recognising and writing the induction hypothesis	C4.2	Proof comprehension
L02.10.3	Write out $P(2)$ in full	C4.2	Proof comprehension
L02.10.4	Start and end of induction step	C4.3	Proof comprehension
L02.11.1	State $P(n+1)$	C3	Faded worked examples
L02.11.2	$\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k = a_{n+1}$	C2.3	Faded worked examples
L02.11.3	Dealing with algebraic manipulation	C1	Faded worked examples
L09.12	Properly Constituted proof by induction	C4	Separated Concerns

Table 7.1: Questions from Lecture Quiz 2 (L02), the concerns they seek to address, and the style of any scaffolding used

Question	Topic	Concern	Style
Q02.08.1	State $P(n+1)$	C3	Separated Concerns
Q02.08.2	$\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k = a_{n+1}$	C2.3	Separated Concerns
Q02.08.3	Simplify RHS of $P(n+1) - P(n)$	C1	Separated Concerns
Q02.09.3	$\sum_{k=1}^{n+1} a_k = \sum_{k=1}^{n} a_k + a_{n+1}$	C2.3	Faded worked examples
Q02.10.1	Identify induction hypothesis	C4.2	Proof comprehension
Q02.10.2	Start and end of induction step	C4.3	Proof comprehension
Q02.10.3	Use of induction hypothesis	C4.2	Proof comprehension
Q02.10.4	Write out $P(7)$ in full	C4.2	Proof comprehension
Q02.10.5	Algebraic simplification	N/A	Proof comprehension
Q02.11	Base case is not the best possible	C4.1	Separated Concerns
Q02.12.1	Importance of the base case	C4.1	Separated Concerns
Q02.12.2	Recognise use of the induction hypothesis	C4.2	Separated Concerns
Q02.12.3	$\sum_{k=1}^{n+1} a_k - \sum_{k=1}^{n} a_k = a_{n+1}$	C2.3	Separated Concerns

Table 7.2: Questions from Assessed Quiz 2 (Q02), the concerns they seek to address, and the style of any scaffolding used



Note, you type in $\sum_{m=1}^{M}?$ as `sum(?,m,1,M)`

Let $P(n)$ be the statement $\sum_{k=1}^{n}(2\cdot k-1)^2=\frac{n\cdot(2\cdot n-1)\cdot(2\cdot n+1)}{3}$

1. Write the statement $P(n+1)$

2. Calculate

$$\sum_{k=1}^{n+1}(2\cdot k-1)^2-\sum_{k=1}^{n}(2\cdot k-1)^2$$

writing your answer in simplified form.

3. Calculate

$$\frac{(n+1)\cdot(2\cdot(n+1)-1)\cdot(2\cdot(n+1)+1)}{3}-\frac{n\cdot(2\cdot n-1)\cdot(2\cdot n+1)}{3}$$

writing your answer in simplified form.

Figure 7.1: Q02.08 in STACK

Quiz (Q02) for mathematical induction. The question is called Q02.08 in the assessed quiz (i.e., question 8 of the second assessed quiz of PPS.)

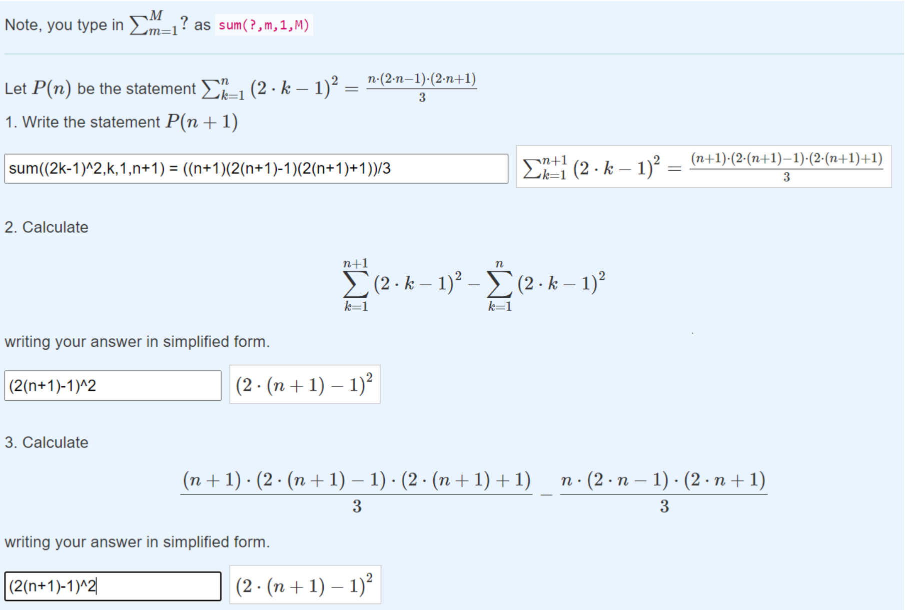
A screenshot of STACK equivalent question of 6.4.1 is shown in Figure 7.1. Notice that the question contains three boxes into which the students should enter their answer.

The scaffolding provided is described as Separated Concerns because there are three parts in this question, where each part addresses one of the concerns. Recall, there are three items in this example, where each item addresses one of the concerns in learning mathematical induction. In particular, item (1) addresses understanding what the induction statement $P(n+1)$ is. Item (2) is designed to address dealing with sigma notation and how students manipulate an expression where all the action happen in the superscript and subscripts. Item (3) is designed to address some basic algebraic manipulation. It asks students to calculate the difference between the two fractions. The goal of item (3) is to help students to recognize that the first fraction is simply the RHS of $P(n+1)$ and the second fraction is the RHS of $P(n)$ and so to encourage students to think before doing the calculations long-hand. This example discussed in Section 6.4.1. Figure 7.2 illustrates the same web page, but with the student's answer typed in. A core part of the design of STACK is that students should type in a mathematical expressions as their answer.

To begin I describe the process of interacting with question (Q02.8) in STACK from the student's perspective. The second assessed quiz of the PPS course contained question (Q02.8), and I shall assume that the student has navigated to the correct web page. In this situation the quiz consists of a fixed list of questions (12 questions), and the student may move freely between them. In other circumstances the questions might be selected randomly from a pool, or during adaptive testing the next question is determined from the outcomes to previous attempts by building a user profile. Here the course organiser has chosen to give students the option to repeatedly attempt the quiz and also make multiple attempts at individual questions within this.

The page contains the context in which the question is set, i.e. the subject, the quiz name (Q02: Assessed quiz), and the question.

For illustration, I consider the first part of the question where students are asked to write the statement $P(n+1)$ which is concern C3 that I described in Section 6.2. Once the student has valid expressions then the system can decide whether the answer is "correct". Validity and correctness in STACK are discussed



Note, you type in $\sum_{m=1}^{M} ?$ as `sum(?,m,1,M)`

Let $P(n)$ be the statement $\sum_{k=1}^{n} (2 \cdot k - 1)^2 = \frac{n \cdot (2 \cdot n - 1) \cdot (2 \cdot n + 1)}{3}$

1. Write the statement $P(n+1)$

```
sum((2k-1)^2,k,1,n+1) = ((n+1)(2(n+1)-1)(2(n+1)+1))/3
```

$\sum_{k=1}^{n+1} (2 \cdot k - 1)^2 = \frac{(n+1) \cdot (2 \cdot (n+1) - 1) \cdot (2 \cdot (n+1) + 1)}{3}$

2. Calculate

$$\sum_{k=1}^{n+1} (2 \cdot k - 1)^2 - \sum_{k=1}^{n} (2 \cdot k - 1)^2$$

writing your answer in simplified form.

```
(2(n+1)-1)^2
```

$(2 \cdot (n+1) - 1)^2$

3. Calculate

$$\frac{(n+1) \cdot (2 \cdot (n+1) - 1) \cdot (2 \cdot (n+1) + 1)}{3} - \frac{n \cdot (2 \cdot n - 1) \cdot (2 \cdot n + 1)}{3}$$

writing your answer in simplified form.

```
(2(n+1)-1)^2
```

$(2 \cdot (n+1) - 1)^2$

Figure 7.2: Q02.08 in STACK with the student's answer

in Section 2.4.1.

Notice that while simple algebraic expressions are not particularly problematic to type in, the sum requires knowledge of specific syntax. In particular, the student has to type in `sum((2k-1)^2,k,1,n+1)` to represent $\sum_{k=1}^{n+1} (2 \cdot k - 1)^2$. Because such linear syntax differs between systems (see e.g. Sangwin and Ramsden (2007)), and because students are known to find this so problematic, specific instructions have been provided at the top of the question, as can be seen at the top of Figures 7.1 and 7.2.

At the outset, based on my prior experience as a teacher, I implemented a marking algorithm based on establishing that the student typed in $P(n+1)$ as a correct equation. In particular, the following two tests applied.

1. Is a student's *left hand side* (LHS) equivalent with the teacher's LHS up to commutativity and associativity of the elementary operations?
2. Is a student's *right hand side* (RHS) algebraically equivalent with the teacher's RHS?

Equivalence up to commutativity and associativity of the elementary operations is rather a subtle test, which is considerably stronger than full algebraic equivalence. In particular, I wanted to accept only answers which contain the sum operator $\sum$ but would condone both

$$\sum_{k=1}^{n+1} (2 \cdot k - 1)^2, \text{ and } \sum_{k=1}^{1+n} (2 \cdot k - 1)^2$$

But I would not want to accept other, algebraically equivalent, forms such as

$$\sum_{k=0}^{n} (2 \cdot k + 1)^2, \text{ or } \sum_{k=1}^{n+1} 4k^2 - 4k + 1$$

It is very unlikely that a student would type this, but nevertheless, I certainly don't want to accept the right hand side of $P(n+1)$ which is algebraically equivalent to the left hand side as a correct form of the left hand side.

For the right hand side, any algebraically equivalent expression are acceptable. The most direct answer

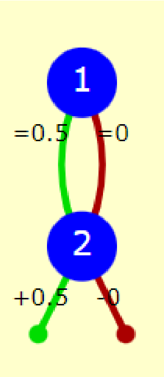


Figure 7.3: Potential response tree for Q02.08.1 in STACK

would be to replace n with $n+1$ leading to

$$\frac{(n+1)(2(n+1)-1)(2(n+1)+1)}{3}.$$

However, it is possible the student would do some minimal "tidying up", e.g. writing the second term $2(n+1)-1$ as $2n+1$.

Recall that students must provide an answer which consists of a mathematical expression, and when valid, the system then seeks to establish properties of this expression. In STACK, the *Potential Response Tree* (PRT) is the algorithm which establishes the mathematical properties of the student's answer and assigns outcomes (The potential response tree is discussed in Section 2.4.2.)

Let's call the potential response tree for Q08.01 `prt1`. Node 1 is to verify the (LHS) and Node 2 to verify (RHS) of students' input.

1. First, to test the whether the LHS is equivalent with the teacher's LHS up to commutativity and associativity of the elementary operations:

   (a) If `true` (prt1-1-T): the LHS is correct (0.5 mark)

   (b) If `false` (prt1-1-F): the LHS is wrong, feedback will be provided.

2. Second, to test the whether the RHS is equivalent with the teacher's RHS up to algebraic equivalence:

   (a) If `true` (prt1-2-T): the RHS is correct (0.5 mark)

   (b) If `false` (prt1-2-F): the RHS is wrong, feedback will be provided.

An illustration for the potential response tree for Q08.1 is shown in Figure 7.3. Both the RHS and LHS are checked independently, so both nodes are always executed in this question. This is an example where the potential response tree really is a graph, rather than a "tree".

7.2 Analysis of Students' Responses

In this section, I described how the data was used and analysed statistically. The purpose of this detailed account is to illustrate the more general process of data collection and review. The data are collected from the PPS course during 2020/21. Only students who consented to participate in the study were included for analysis.

The data was students' scores and responses of the designed STACK questions for the two quizzes; the lecture quiz L02 and the assessed quiz Q02 of week 2 of PPS. The data was both quantitative and qualitative in nature. Students' scores on each question and course total were quantitative, while students' responses were in qualitative nature. To answer the two research questions identified in Section 7.1.1, students were divided into two groups:

G1: students who took L02.

G2: students who did not take L02, or who got a score of 0 in L02.

To answer the first research question; i.e what way is engagement lecture quiz related to (i) success on the weekly assessed quiz, and (ii) the course total? I compare students' scores between the two groups. This is a quantitative analysis of the available scores.

To answer the second research question, i.e what are the common mistakes when using online materials (i.e, Separated Concerns in writing STACK questions) to prepare for mathematical induction? I compare students' responses of specific assessed questions of the assessed quiz Q02 between the two groups. To conduct the analysis for the second research question, I firstly reviewed the raw data for all the attempts from the "basic question use report" for the proposed question. After reviewing the raw data and identifying the common misconceptions, the hypothesis is: students who took the lecture quiz L02 did not make this misconception.

7.3 Results for Q02.08.1

Let us consider, in detail, how the first cohort 2020/21 of students responded to question Q02.08.1 in the assessed quiz Q02.

The "basic question use report" is provided as a simple text-based output summary of students' attempts at a particular STACK question. Part of this report is shown in Figure 7.4. I start the analysis by first looking at the basic question use report and see whether the student's final answer satisfies two tests in the potential response tree shown in Figure 7.3.

```
## prt1 (426)
31  (  7.28%); !
44  ( 10.33%); # = 0 | ATEqualComAss (AlgEquiv-false). | prt1-1-F | prt1-2-F
1   (  0.23%); # = 0 | ATEqualComAss (AlgEquiv-true). | prt1-1-F | ATAlgEquiv_TA_not_equation. | prt1-2-F
47  ( 11.03%); # = 0 | ATEqualComAss (AlgEquiv-true). | prt1-1-F | prt1-2-F
1   (  0.23%); # = 0 | ATEqualComAss ATAlgEquiv_SA_not_expression. | prt1-1-F | prt1-2-F
3   (  0.70%); # = 0 | ATEqualComAss ATAlgEquiv_TA_not_equation. | prt1-1-F | prt1-2-F
26  (  6.10%); # = 0.5 | ATEqualComAss (AlgEquiv-false). | prt1-1-F | prt1-2-T
8   (  1.88%); # = 0.5 | ATEqualComAss (AlgEquiv-true). | prt1-1-F | prt1-2-T
62  ( 14.55%); # = 0.5 | prt1-1-T | prt1-2-F
203 ( 47.65%); # = 1 | prt1-1-T | prt1-2-T
```

Figure 7.4: Raw basic question use report for the original Q02.08.1 in STACK

Now, consider the last line of the basic question use report illustrated in Figure 7.4

```
203 ( 47.65%); # = 1 | prt1-1-T | prt1-2-T
```

To read the last line, specifically, such as `203 (47.65%); # = 1`, is read as " 203 out of 426 (47.65%) got the correct answer and received 100 %." Note the 426 comes from `## prt1 (426)` in the first line.

`prt1-1-T` is read as "potential response tree 1, node 1, returned `true`." Recall, this is the node which determines if the LHS of the equation is correct.

`prt1-2-T` is read as "potential response tree 1, node 2, returned `true`." Recall, this is the node which determines if the RHS of the equation is correct.

The line shows that 203 out of 425 (47.65%) typed both LHS and RHS of the equation correctly and received 1 point.

Turning now to node 1, recall that the test of equivalence up to commutativity and associativity is used for the LHS. This test also records whether the two expressions were also algebraically equivalent in the answer node. Hence the note `ATEqualComAss (AlgEquiv-true). | prt1-1-F | prt1-2-T` is read that the first test established algebraic equivalence of the LHS of the student's and teachers answer, but since `prt1-1-F` they were not equivalent up to commutativity and associativity. However the `prt1-2-T` means the RHS of the student's answer was equivalent with the RHS of the teacher's answer.

This question was not trivially easy for this group of student since about half of the cohort did not answer it fully correctly. Some 31 students (7.28%) left an invalid or empty response to this part of the question, indicated by the `!` symbol. In fact, only three responses were syntactically invalid, demonstrating the value of the immediate feedback: most students correct syntax errors when prompted.

The note `ATEqualComAss ATAlgEquiv_SA_not_expression` indicates that the Algebraic Equivalence test decided that the student's answer (SA) was not an expression. That is to say, it is the wrong "type" of object. Digging further to look at the one answer causing this note, the student typed in

```
[sum((2*k-1)^2,k,1,n)]+(2*n+1)^2
```

Literally the student has a list containing the sum, and to which they have added the expression $(2n+1)^2$. It looks like the student has the left hand side of $P(n)$ to which they have added the next term in the sum. The square brackets create a list, and so this is the reason for the type mismatch. Square brackets in writing are ambiguously used for grouping terms but a computer is strict. One of the problems of allowing (requiring) students to type expressions is that they can type anything. Sometimes they can type syntactically valid but mathematically meaningless expressions.

Therefore, when the analysis is reported, the percentages of students who got correct, invalid answers, and also the percentages of students who made the common mistakes based on the concerns that had been addressed can be indicated.

From the basic question report it can be shown that some of students think the $P(n+1)$ is only the RHS and not an equation. Strangely, and somewhat unexpectedly, 43 students (10.80%) gave `sum((2*k-1)^2,k,1,n+1)` which is the LHS. I did not anticipate students typing only the left hand side of the equation as $P(n+1)$.

The original potential response tree has done a reasonable job in establishing the properties we want. However, now that I have a significant data set in which I can see what students actually do, the response tree can be improved accordingly. Therefore in the next section I describe the updated tree, together with the revised outcome created by re-grading the students' responses using this updated tree.

7.3.1 Modified potential response tree for Q02.08.1

The updated potential response tree for Q02.08.1 is shown in Figure 7.5. This now has five nodes and is significantly more complex. Node 1 establishes whether or not the student's answer is an equation. On the basis of this test, two different situations are branched. If the student has typed an equation then we proceed exactly as before:

1. Node 2: check the LHS of the student's answer with the LHS of the teacher's up to commutativity and associativity.

2. Node 3: check the RHS of the student's answer with the RHS of the teacher's up to algebraic equivalence.

The above test is performed, and I have chosen to award half marks for each test. If the student has not typed an equation then we proceed in a similar, but subtly different, way.

1. Node 4: check the student's answer with the LHS of the teacher's up to commutativity and associativity, and if so STOP.

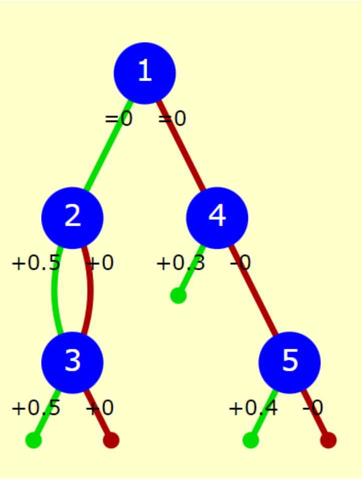


Figure 7.5: Revised potential response tree for Q02.08.1 in STACK

```
## prt1 (426)
31  (  7.28%); !
38  (  8.92%); # = 0   | prt1-1-F | ATEqualComAss (AlgEquiv-false). | prt1-4-F | prt1-5-F
1   (  0.23%); # = 0   | prt1-1-F | ATEqualComAss ATAlgEquiv_SA_not_expression. | prt1-4-F | prt1-5-F
3   (  0.70%); # = 0   | prt1-1-F | ATEqualComAss ATAlgEquiv_TA_not_equation. | prt1-4-F | prt1-5-F
6   (  1.41%); # = 0   | prt1-1-T | ATEqualComAss (AlgEquiv-false). | prt1-2-F | prt1-3-F
1   (  0.23%); # = 0   | prt1-1-T | ATEqualComAss (AlgEquiv-true). | prt1-2-F | ATAlgEquiv_TA_not_equation. | prt1-3-F
1   (  0.23%); # = 0   | prt1-1-T | ATEqualComAss (AlgEquiv-true). | prt1-2-F | prt1-3-F
49  ( 11.50%); # = 0.3 | prt1-1-F | prt1-4-T
46  ( 10.80%); # = 0.4 | prt1-1-F | ATEqualComAss (AlgEquiv-true). | prt1-4-F | prt1-5-T
26  (  6.10%); # = 0.5 | prt1-1-T | ATEqualComAss (AlgEquiv-false). | prt1-2-F | prt1-3-T
8   (  1.88%); # = 0.5 | prt1-1-T | ATEqualComAss (AlgEquiv-true). | prt1-2-F | prt1-3-T
13  (  3.05%); # = 0.5 | prt1-1-T | prt1-2-T | prt1-3-F
203 ( 47.65%); # = 1   | prt1-1-T | prt1-2-T | prt1-3-T
```

Figure 7.6: Raw basic question use report for the updated Q02.08.1 in STACK

2. Node 5: check the student's answer with the RHS of the teacher's up to algebraic equivalence.

If the student's answer is equivalent to the LHS of the teacher's answer (Node 4) then it will also be algebraically equivalent to the RHS, which is why we stop when Node 4 returns `true`. We have chosen to award lower marks for each of these situations. Other teachers may not agree with the marks awarded here, but marks and the value of students' work is an individual matter for course organisers and I don't comment on value in terms of marks further within this research. Recall, this research was done with real students. Note that for illustrative purposes here different marks result for each of the four end-nodes of the tree. The only different situations giving rise to the same marks now are `prt1-2-T | prt1-3-F` and `prt1-2-F | prt1-3-T`, both of which give 0.5 marks. One advantage of updating the potential STACK response tree is that students can have partial marks.

Revised data for students' attempts at Q02.08.1 are shown in Figure 7.6.

In fact, a slightly more useful way to look at the data for this question is to split up the answer notes and consider each node individually. This data is shown in Figure 7.7, together with brief narrative added in parentheses by me later for the convenience of the reader.

Node 1 checks if the student has entered an equation. Entering an equation is tested, in the revised tree, by `prt1-1-T`, which was given by 258 (60.56%) of the responses. Surprisingly, 49 students appear to have typed in the LHS only, which is slightly more than the 46 who typed in the RHS only. Since both of these outcomes were generated for more than 10% of the cohort it was well worth adding nodes to the tree with corresponding outcomes (e.g. feedback). It is probably not worthwhile adding nodes to test for strange individual responses.

Perhaps more time on the original response tree could be spent in advance. However, writing response trees is a significant amount of work. Experience suggests that a lot of time can be expended writing

```
## prt1 (426)
31  (  7.28%); !
137 ( 32.16%); prt1-1-F (Did not type in an equation)
258 ( 60.56%); prt1-1-T
42  (  9.86%); prt1-2-F
216 ( 50.70%); prt1-2-T (Correct LHS)
21  (  4.93%); prt1-3-F
237 ( 55.63%); prt1-3-T (Correct RHS)
88  ( 20.66%); prt1-4-F
49  ( 11.50%); prt1-4-T (Misconception (i): answer equivalent to LHS only)
42  (  9.86%); prt1-5-F
46  ( 10.80%); prt1-5-T (Misconception (ii): answer equivalent to RHS only)
```

Figure 7.7: Split notes for the updated Q02.08.1 in STACK, with narrative

potential tests with feedback which are never actually used by any significant proportion of the students. Time is then wasted trying to second-guess what students might do. Even then, a review is necessary because students do things we might not actually anticipate. Typing the LHS only was not anticipated and the original tree would not have tested for this anyway. In general, I have found it much more sensible to operate a two year (at least) cycle of question development.

1. First year cohort (2020/21): get the question working, with the essential properties for correct/incorrect established reliably.
2. Second year cohort (2021/22): review students' data and update response trees accordingly, making sure the second (and subsequent) years benefit from better feedback (formative) or more nuanced partial credit (summative).

In assessed quizzes, or online exams, marks are not available immediately. So I could improve the PRT before students marks are released to students. In addition, it is helpful to have STACK establish evidence of a particular misconception and create a note from the potential response tree (see Figure 7.7). This has the added benefit that subsequent users of the question will potentially benefit from feedback based on establishing this misconception - tagging all the misconceptions this year is worthwhile effort from a teaching prospective, as well as contributing to this research.

In summary,

- Students could enter complex expressions such as `sum((2k-1)^2,k,1,n+1)` to represent $\sum_{k=1}^{n+1}(2\cdot k-1)^2$. The syntax does not appear to have been an insurmountable barrier for most.
- According to the basic question use report in Figure 7.7, about a third of students (137 out of 426; 32.16 %) did not type in an equation, and these students were almost equally split between groups typing in (i) the LHS, (ii) the RHS or (iii) something else. I was surprised how many students typed in the LHS of $P(n+1)$. Students unfortunately did not make better use of the formative quizzes, as about a third of the students made a serious category error by not typing in an equation.
- Only about half of the students (258 out of 426; 60.56%) got this question correct, so that it is a non-trivial exercise and quite appropriate for our formative quiz. See Figure 7.7 for more details.

I identified a concern from previous research which was discussed in Section 6.3, designed materials to test if our students exhibit this misconception by separating out this concern, and have provided evidence they did indeed do so.

7.3.2 Results for L02.11.1

The only other questions to address C3, which is writing $P(n+1)$, is L02.11.1. I used the updated PRT developed for the assessed quiz Question Q02.08.1 in the lecture quiz Question L02.11.1, since the question is identical but using a different statement for $P(n+1)$.

The question L02.11.1 asked students.

Let $P(n)$ be the statement $\sum_{k=1}^{n} k \cdot k! = (n+1)!$.

1. Write the statement $P(n+1)$.

This is the first part of a Separated Concerns question. Note this statement is never true, rather $\sum_{k=1}^{n} k \cdot k! = (n+1)! - 1$. In any case, the potential response tree shown in Figure 7.8 was applied to students' answers.

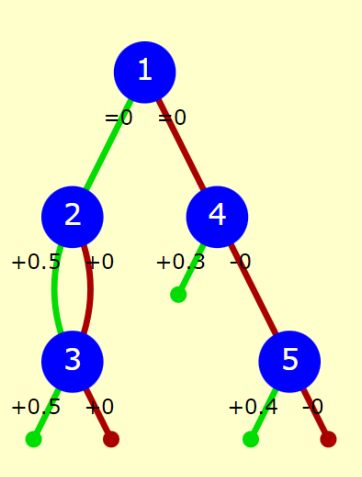


Figure 7.8: Revised potential response tree for L02.11.1 in STACK

A surprisingly large proportion (64%) of the 413 attempts at this question simply did not answer this part of the question at all. Of the remaining 36% of attempts 116 (28%) answered this question correctly. Students who attempted this question in the lecture quiz were perfectly able to complete it. For completeness, data is shown in Figure 7.9.

```
## prt1 (413)
266 ( 64.41%); !
3   (  0.73%); # = 0 | ATLogic_True. | prt1-1-T | ATEqualComAss (AlgEquiv-false). | prt1-2-F | prt1-3-F
10  (  2.42%); # = 0 | prt1-1-F | ATEqualComAss (AlgEquiv-false). | prt1-4-F | prt1-5-F
4   (  0.97%); # = 0.3 | prt1-1-F | prt1-4-T
1   (  0.24%); # = 0.4 | prt1-1-F | ATEqualComAss (AlgEquiv-false). | prt1-4-F | prt1-5-T
8   (  1.94%); # = 0.5 | ATLogic_True. | prt1-1-T | ATEqualComAss (AlgEquiv-false). | prt1-2-F | prt1-3-T
5   (  1.21%); # = 0.5 | ATLogic_True. | prt1-1-T | prt1-2-T | prt1-3-F
116 ( 28.09%); # = 1 | ATLogic_True. | prt1-1-T | prt1-2-T | prt1-3-T
```

Figure 7.9: Raw basic question use report for L02.11.1 in STACK

7.4 First Cohort (2020/21)

In the first cohort (2020/21), by design students could access the assessed quiz without taking the lecture quiz. The quantitative comparison of L02, the "lecture quiz" with Q02 the "assessed quiz" was conducted in three phases:

1. Splitting the data into 2 groups (i) students who did not take L02. (ii) students who did take L02 or who got a score of 0 (this indicates they didn't really "attempt" the quiz, but might have looked at the questions or embedded video). Only 4 students had more than one attempt, so it makes no sense here to have a group who look L02 more than once.

2. Comparing the two groups using an independent sample t-test to find out in what way is engagement lecture quiz related to (i) success on the weekly assessed quiz, and (ii) the course total.

Quiz	Total	Mean	Median	SD	Min	Max
L02	242 (58.0%)	6.70	7.30	3.27	0	10
Q02	413 (99.7%)	7.00	7.53	1.77	1.00	10

Table 7.3: Total students who took L02 and Q02 in 2020/21

Quiz	Total	Mean	Median	SD	Min	Max
L02	450 (95.3%)	7.70	8.67	2.62	0.14	10
Q02	471 (99.7%)	7.66	8.25	1.81	1.00	10

Table 7.4: Total students who took L02 and Q02 in 2021/22

3. Comparing the two groups using the Chi squared test to find out (i) is there any evidence that students who did take L02 are not making the known misconceptions? (ii) Is there any evidence that students who did not take L02, or who got 0, are making the known mistakes?

In the first cohort (2020/21), there were 470 students taking PPS course. Of these 414 participants freely gave their consent to have their responses to this week of the course analysed and reported as part of this research.

First, Table 7.3 summarised the data for both L02 and Q02. There were 242 out of 414 (58.0%) of PPS students who took L02, while the remaining 172 (41.5%) of the PPS students did not take L02. For the assessed quiz Q02, almost 99.7% of PPS students took the assessed quiz (Q02).

7.4.1 Results: how engagement in a lecture quiz related to (i) success on the weekly assessed quiz, and (ii) the course total?

In this section, I consider the results for the first research question which illustrates how engagement in a lecture quiz related to (i) success on the weekly assessed quiz, and (ii) the course total. An independent sample t-test was conducted and I found that those students who had taken the lecture quiz L02 had a statistically significantly higher total score in the assessed quiz Q02 ($N = 242$, $Mean = 7$, $Median =$ 7.53, and $SD = 1.77$) compared to those students who did not take the lecture quiz L02 ($N = 172$, $Mean = 6.68$, $Median = 7.04$, and $SD = 1.86$), $t(412) = 3.35$, $p < 0.001$.

An independent sample t-test was also conducted and I found that those students who had taken the lecture quiz L02 had statistically significantly higher course total at the end of the course ($N = 242$, $Mean = 75\%$, $Median = 77\%$, and $SD = 13.31$) compared to those who did not take the lecture quiz ($N = 172$, $Mean = 68\%$, $Median = 72\%$, and $SD = 16.75$), $t(412) = 4.89$, $p < 0.001$.

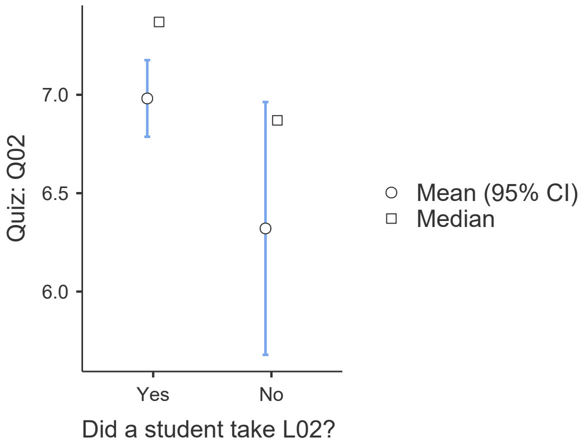


Figure 7.10: Descriptive plots of Q02 marks based on taking L02 in 2020

Taking L02	Final Grade		Total
	Pass	Fail	
No	141 (39.5 %)	27 (58.7 %)	168
Yes	216 (60.5 %)	19 (41.3 %)	235
Total	357 (100 %)	46 (100 %)	403

Table 7.6: Pass or fail the course for the two groups

Questions	Taking L02	N	Mean	Median	SD	SE
Q02.08	Yes	236	0.57	0.56	0.25	0.016
	No	161	0.51	0.56	0.26	0.021
Q02.09	Yes	238	0.75	0.83	0.15	0.01
	No	160	0.67	0.83	0.2	0.015
Q02.10	Yes	240	0.61	0.67	0.26	0.017
	No	162	0.52	0.50	0.25	0.019
Q02.11	Yes	238	0.79	0.83	0.174	0.011
	No	165	0.73	0.83	0.27	0.021
Q02.12	Yes	240	0.64	0.56	0.219	0.014
	No	159	0.62	0.56	0.24	0.019

Table 7.7: Scores for each question in the assessed quiz (Q02) by those who took the lecture Quiz (L02) and those who did not take L02.

	Taking L02 ($n = 371$)		Not taking L02 ($n = 50$)	
	Mean	SD	Mean	SD
Q02 marks	7	1.91	6	2.32

Table 7.5: Q02 marks based on taking L02 in 2020

The final grade (pass or fail) of the course was also considered for comparison between the two groups. I ran a Chi square test of independence between the two groups and the results summarised in Table 7.6. As shown, the proportion of students who took the lecture quiz and passed the course (216 out of 357; 60.5%) were higher than the proportion of those who did not take the lecture quiz and passed the course (141 out of 357; 39.5%.) The data suggests that there is a significant relationship between the two variables (taking the lecture quiz) and (final grade), $\chi^2(1, N = 403) = 6.18$, $p = .017$.

Quiz L02 had complete attempts from 371 students. This quiz could be taken as many times as students wished.

An independent sample t-test was conducted to see if there is a difference in answering each question in the assessed quiz between the two groups. The results of the independent sample t-test is summarised in Table 7.8.

The descriptive statistics for each question in Q02 are illustrated in Table 7.7.

As shown in Table 7.8, I found that those students who had taken L02 had a statistically significantly higher score in answering the assessed question Q02.Q8 which addressed sigma notation concern in writing the difference ($N = 236$, $Mean = 0.572$) compared to those who did not take L02 ($N = 161$, $Mean = 0.509$), $t(395) = 2.454$, $p = 0.015$.

There was also a significant difference in answering the assessed question Q02.Q10 which was a fading

Q02(Concern)	SD	df	P-value	Effect size
Q.8(C2.3: Writing the difference)	2.454	395	0.015	0.250
Q.9 (Proof by induction (general)	3.156	396	0.002	0.322
Q.10 (C2.1, C4.2, C4.3, C4.4, C4.5)	3.153	400	0.002	0.320
Q.11	2.608	401	0.009	0.264
Q.12 (C4.1: Identify base case)	0.854	397	0.394	0.087

Table 7.8: Independent sample t-test results comparing groups who did and did not take L02, showing results for individual questions in Q02.

	L02.11a	Q02.8.1
Not answered/Invalid	265 (67 %)	30 (8 %)
Correct	116 (39 %)	203 (55 %)
Incorrect "Did not type in an equation"	15 (6 %)	137 (37 %)

Table 7.9: Raw outcomes for questions L02.11a and Q02.8.1

question asking students to prove that $3^n < n!$ for $n \geq 9$.

Question Q02.10 addressed 4 concerns based on the concerns that were identified in Section 6.2: (C4.2: recognize induction hypothesis, C4.3 writing induction hypothesis, C4.4: where the hypothesis used in the induction step, C4.5: knowing the start and end of induction step). Students who took L02 scored higher ($N = 240$, $Mean = 0.606$) than those who did not take L02 ($N = 162$, $Mean = 0.52$) , $t(400) = 3.15$, $p = 0.002$.

In summary, the results show that engaging with the lecture quiz related positively to success on the weekly assessed quiz and the course total.

Given this statistical result, it appears worth digging in much greater detail as to exactly what answers students give. Do they really exhibit these misconceptions and can I find specific evidence?

7.4.2 Results: what are the common misconceptions?

In this section, I consider the results for the second research question i.e what are the common mistakes students make when using online materials? In particular, Separated Concerns in writing STACK questions to prepare for mathematical induction during the first cohort (2020/21). In this section, I focus on the third concern C3, i.e. understanding what the induction statement $P(n + 1)$ actually is. The remaining concerns are broader and probably better addressed once students have reached a basic competence in writing and reading simple proofs by induction.

The third concern was addressed in as question (L02.11a) in the lecture quiz (L02) and as (Q02.8.1) in the assessed quiz (Q02). In both questions, students were asked to write the statement $P(n + 1)$. Note, the immediate validity feedback in this question required an *equation* in 2020/21. The concern was also explicitly addressed in the lecture notes, the live lectures and in pre-recorded video clip sections.

Raw outcomes for questions L02.11a and Q02.8.1 are summarised in Table 7.9. It is clear that 67% of the students did not attempt to write $P(n + 1)$ during the practice lecture quiz. Of the remainder who did, the overwhelming majority got the answer completely correct, or equivalent to the correct answer in a way which made it clear they had typed in something equivalent to $P(n+1)$. For example, L02.10c expected students to write terms like $1^2 + 2^2$ rather than the simplified form $1 + 4$ for full correctness.

For the assessed quiz (Q02), this concern is raised in Q02.8.1, and here only 30 (8%) of the students failed to answer this part, or did not correct an invalid answer (for which feedback is immediately available during the assessed quiz). The problem in this question does not appear to be primarily a failure to be able to type in a complex expression, but the largest single mistake (and more prevalent than the correct answer) is a failure to recognise that $P(n)$ is a statement in the form of an *equation*. About 37% of students failed to type in an equation.

The results for Q02.8.1 are illustrated in the Tables 7.10, and 7.11. Based on students' reposes, the main misconception identified was that students did not write $P(n + 1)$ as an equation. Therefore, students either wrote the LHS or RHS.

A Chi-square test of independence was performed to test the hypothesis that students who took the lecture quiz L02 did not show evidence of misconception, which is writing $P(n + 1)$ as a non-equation when answering the assessed question Q02.8.1. The results show that there is a significant relationship between taking the lecture quiz L02 and the misconception in the assessed quiz Q02. Students who did not take the lecture quiz are more likely to show evidence of misconception in writing $P(n + 1)$ as a non-equation in answering the assessed quiz question Q02.8.1, $\chi^2(1, N = 207) = 11.1$, $p < .001$. A student seeking to write an induction proof without knowing that $P(n)$ is an equation would be in an

Taking L02	Q02.8.1 correct ans.	misconception; no eqn.	Total
No	16 (20%)	64 (80%)	80 (100%)
Yes	54 (42.5%)	73 (57.5%)	127 (100%)
Total	70 (34%)	137 (66%)	207 (100%)

Table 7.10: Chi square analysis of Q02.8.1 (no equation written)

Taking L02	Q02.8.1 correct ans.	LHS only	RHS only	Total
No	16 (29%)	18 (32%)	22 (39%)	56 (100%)
Yes	54 (51%)	28 (26%)	24 (23%)	106 (100%)
Total	70 (43%)	46 (28%)	46 (28%)	162 (100%)

Table 7.11: Chi square analysis of Q02.8.1 (LHS and RHS)

almost hopeless situation.

The misconceptions we had identified as C3 did occur, and was widespread. Sadly, for whatever reason, most students did not attempt the specific parts of the question which addressed this concern explicitly. A surprise to us was the number of students who wrote $\sum_{k=0}^{n-1} k \cdot r^k$, i.e. the LHS.

7.4.3 Discussion

The main goal of this research was to confirm the hypothesis that engaging with the lecture quizzes is associated with improved performance on the assessed quizzes and the course total. The second goal of this study was to address some concerns and common misconceptions when using STACK questions to prepare for mathematical induction.

I compared students' performance on questions with both faded worked examples and with Separated Concerns. To investigate these questions we specifically designed and tested questions as part of a university mathematics course. Proofs and Problem Solving (PPS) is a year 1, semester 2, module run at the University of Edinburgh. In this work, the questions were designed with online assessment in mind, and the STACK online assessment system in particular. The questions were designed based on four concerns identified by reviewing students responses from STACK basic question usage report in learning mathematical induction in PPS course. For example, some misconceptions were found in dealing with sigma notation, using the induction hypothesis in the inductive step and also in writing the statement $P(n+1)$. The first quiz was a formative "lecture quiz", used to prepare students for the lecture. The second quiz was an "assessed quiz" that students took at the end of the weekly learning cycle to test for evidence of this possible misconceptions. It was expected that if students solve these questions in the lecture quiz correctly, they will be able to solve the corresponding question on the assessed quiz without demonstrating evidence for these known misconceptions.

I compare between success on these items and a student's overall performance. I have noticed that only 242 out of 417 (58%) of the PPS students were involved in participating in the lecture quiz (L02) as participating in the lecture quiz was optional. Surprisingly, around 172 out of 414 (41.5%) of the PPS students did not take the lecture quiz (L02). Sadly, for whatever reason, those students did not attempt the specific parts of the questions which addressed the concerns explicitly. These students had statistically significant lower scores in the assessed quiz and also lower course total than those who took the lecture quiz. The results confirm the evidence that engaging with the lecture quizzes is associated with improved performance on the assessed quizzes. For the next cohort, we really want to see if really forcing students to do the lecture quiz will make a difference in the assessed quiz? In particular, improving overall performance and in helping students to avoid answering consistent with specific the misconceptions?

7.5 Second Cohort (2021/22)

7.5.1 Goal

In the first cohort (2020/21), we have noticed that around (172 out of 414; 41.5%) of the PPS students did not take the lecture quiz (L02). These students had statistically significantly lower scores in the assessed quiz and also had a lower course total than those who took the lecture quiz. For this reason, the goal is to investigate the following.

1. Does requiring students to take the lecture quizzes result in increasing the success rates in the assessed quiz and also the course total? (general goal)
2. Does requiring students to take the lecture quiz better prepare students for the assessed quiz? i.e those students who take the lecture quiz will not show evidence of the common misconceptions in induction that addressed in Section 6.2 (micro-goal)

7.5.2 Method

In order to encourage students to access the lecture quizzes, there are two options:

- Option 1: Students are required to complete the lecture quiz by scoring 80% in order to gain access to the assessed quiz.
- Option 2: make lecture quizzes count for some of the final course grade, for example 5%.

Both of these options are already used in various courses, and so to implement either within PPS for the second year of this study would not be unusual or out of place.

The first option is currently used in *Fundamentals of Algebra and Calculus*, which takes a mastery approach to learning, see (Kinnear, Wood, and Gratwick 2021). In this situation the lecture quiz, by design, can be taken as often as a student wishes so that ultimately they do achieve the 80% score. Hence, feedback and a full worked solution, will be available to students immediately after answering. In this situation there is nothing to force a student to engage: it would be very easy to "game" the system by randomly answering, reading the feedback and immediately entering the correct answer. I suspect that most students will actually engage in a meaningful way with the lecture quiz. This course is a post-compulsory course on a university mathematics degree. As a year 1 course on Proofs and Problem Solving students know that this is preparatory and will help them build understanding and skills (in proof writing). Part of the didactic contract (in the sense of (Brousseau 1997)) is that the lecture quiz is a formative experience through which they will learn. The advantage of this option is that students are strongly encouraged to read the feedback which addresses particular concerns. The disadvantage of this approach is that there will be less evidence on the prevalence of the misconception. It is possible to look at students' first attempts at the lecture quiz, and this will provide some data (provided the assumption that the prevalence of "gaming" is low is a correct assumption).

The didactic contract consists of the set of behaviors that the teacher expects from the student and the student from the teacher. The didactic contract is that the teacher is obliged to teach and the pupil to learn' (Brousseau 1997) or at least to pass the assessment. The teacher sets tasks, the learners carry them out; the contract is that by doing the tasks the learners will do enough to pass (Mason and Johnston-Wilder 2004).

The second option has been used in previous designs of another year one course *Introduction to Linear Algebra* (ILA). This option has the advantage that there will probably better data on frequencies of misconceptions. The disadvantage is that participation in the lecture quizzes will be lower. The 5% marks will act as some incentive, but probably to a less of an extent than the first option. For this work, I chose option 1, where students are required to complete the lecture quiz by scoring 80% in order to gain access to the assessed quiz.

Now, I expect most (essentially all) students are taking the lecture quiz. For the analysis, I compare students' scores using paired t-test to see students do well when they are forced to take the lecture quiz. I can also compare the mean scores and the data distribution between students in this semester and students in the previous semesters when the settings of the lecture quiz is different.

Note, overall only the best 8 out of 10 scores for the assessed quizzes actually contribute to the coursework component of the course. This mechanism accounts for inevitable minor illness, absence or technical problems without a complicate mitigation process. The University has a mitigation process for when students have more serious problems, but the 8/10 mechanism is a pragmatic and effective way to reduce the effect of missing one or two quizzes. For this reason, all students on the course would not expected to actually participate fully in the week 2 assessed quiz.

7.5.3 Design

The items used in both the lecture quiz and the assessed quiz are the same as used in 2020/21. The only difference in 2021/22 was to require students complete the lecture quiz by scoring at least 80% in order to access the assessed quiz.

7.6 Results

7.6.1 Quantitative comparison of the first cohort (2020/21) and the second cohort (2021/22)

There were 394 out of 474 PPS students who gave their consent to participate in the second cohort of the study that was conducted in the academic year 2021/22. In 2021/22, all students were required to complete the lecture quiz by scoring at least 80 % in order to access the assessed quiz as part of the weekly course design.

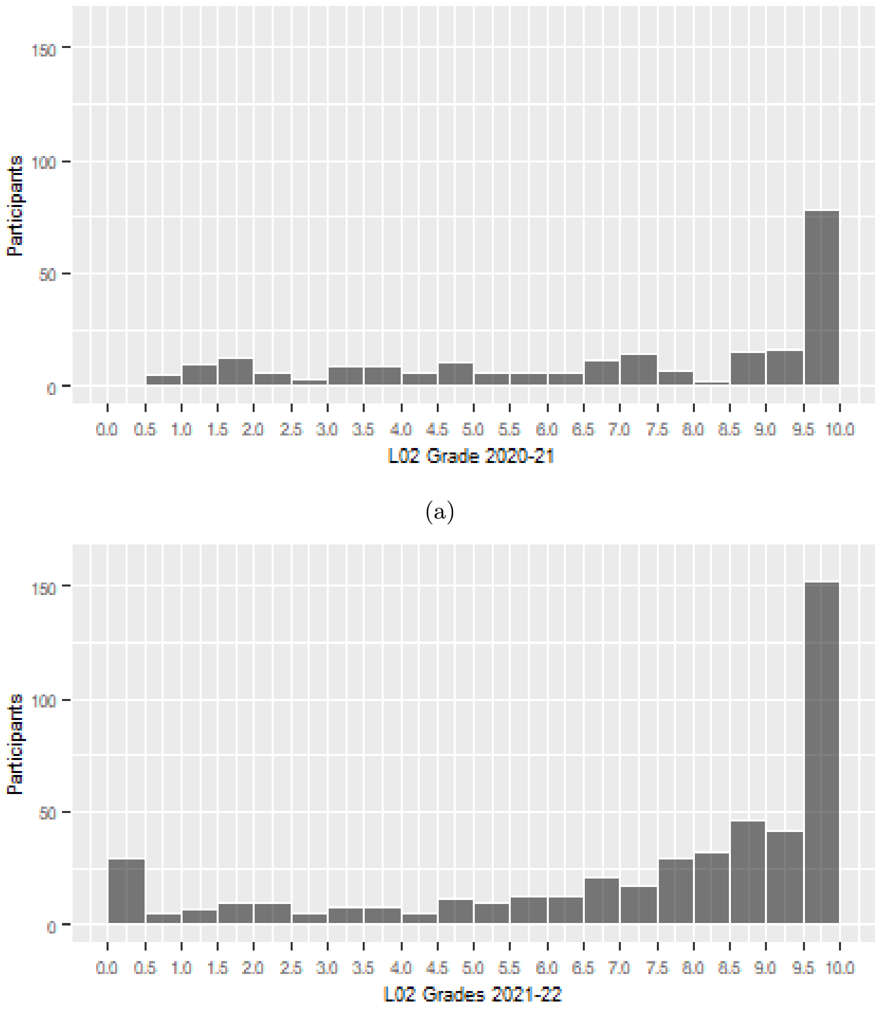
The quantitative comparison of students scores in the first cohort (2020/21), with the second cohort (2021/22) was conducted in three phases

1. Splitting the data into 2 groups (i) students who took Q02 in 2020/21 and (ii) students who took Q02 in 2021/22.
2. Comparing the two groups using independent sample t-tests to find out in what way is engagement with the lecture quiz related to success on the weekly assessed quiz.
3. Comparing the two groups using a Chi-square test to find out is there any evidence that students who complete the lecture quiz with scoring 80% in order to do the assessed quiz are not making the known mistakes?

Figure 7.11 shows the distributions of the lecture quiz L02 grades for the two cohorts. During the second cohort (2021/22), the distribution of L02 grades was highly skewed left.

The following Table 7.12 includes calculations of some basic (that is, descriptive) statistics from the data set. Examining these numbers, the mean grade of L02 during the second cohort (2021/22) was higher than the mean grade during the first cohort (2020/21.) This was expected as taking the lecture quiz L02 was optional in the first cohort. Since both charts are skewed to the left, the best measure of center is the median. The center of L02 scores in the first cohort was lower than the center of the second cohort by 7 grades. This might be because students in the second cohort were required to complete the lecture quiz with scoring 80% in order to do the assessed quiz, while taking the lecture quiz in the first cohort was optional.

Figure 7.12 shows the distributions of marks for the assessed quiz Q02 data for PPS students in the two cohorts. In the second cohort, most participants scored between 8.5 to 9 out of 10 in the assessed quiz



(a)

(b)

Figure 7.11: Histograms of L02 grades during the first cohort 2020/21 (a) and the second cohort 2021/22 (b)

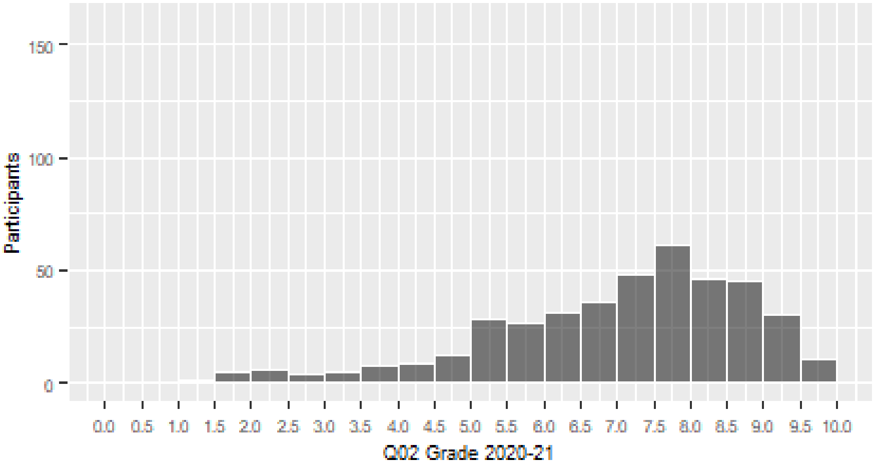
Q02. While in the first cohort, the data was at its peak between 7.5 to 8 out of 10. This shows us that participants in the second cohort scored higher in the assessed quiz Q02 than participants in the first cohort.

An independent sample t-test was conducted to compare the mean scores of the lecture quiz L02 between students of the first cohort (2020/21) and the second cohort (2021/22.) The descriptive statistics of the lectures quiz grade for the two groups summarised in Table 7.12 The mean score of L02 in the second cohort (2021/22) was significantly higher than the first cohort (2020/21) ($t = 14.1$, $p < 0.001$.)

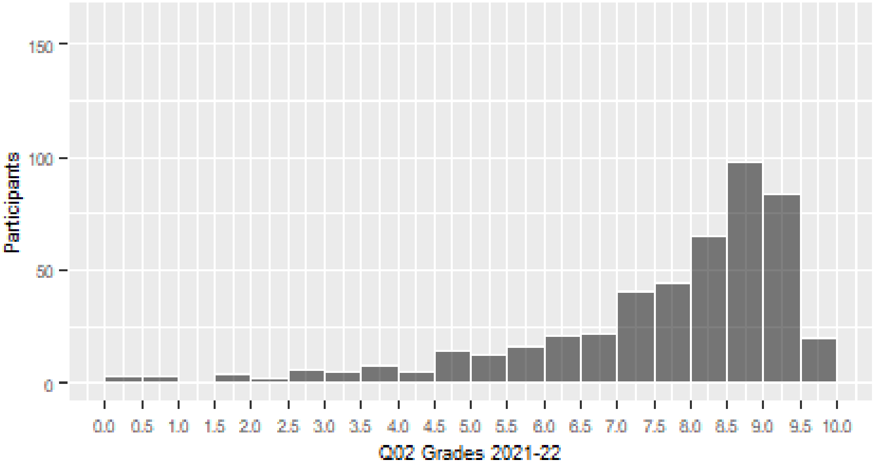
Year	N	Mean	Median	SD	SE
2020/21	242	4.0	2.0	4.13	0.203
2021/22	394	7.5	9.0	2.99	0.151

Table 7.12: L02 descriptive statistics for 2020/21 and 2021/22

An independent sample t-test was also conducted to compare the mean scores of the assessed quiz Q02 between students of 2020/21 and 2021/22. The mean score of Q02 in 2021/22 was significantly higher than 2020/21 ($t = 5.09$, $p < 0.001$.)

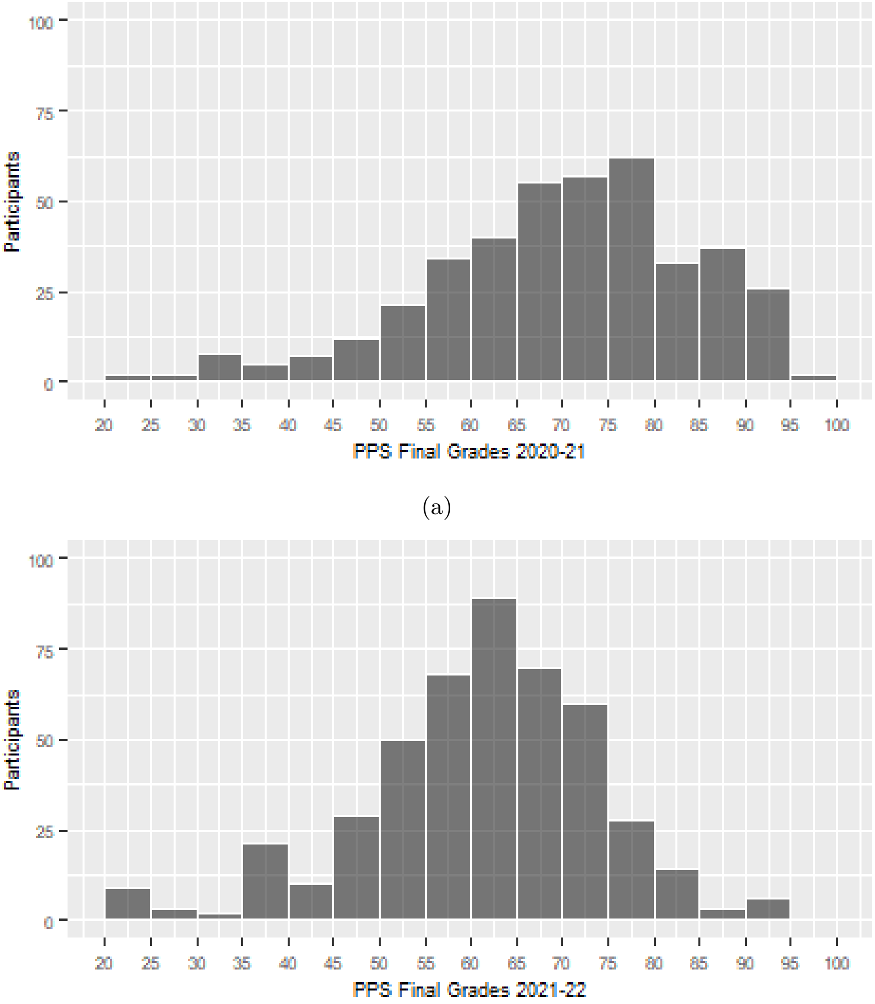


(a)



(b)

Figure 7.12: Histogram of Q02 grades during the first cohort 2020/21 (a) and the second cohort 2021/22 (b)



(a)

(b)

Figure 7.13: Histograms of final grades of the PPS course during the first cohort 2020/21 (a) and the second cohort 2021/22 (b)

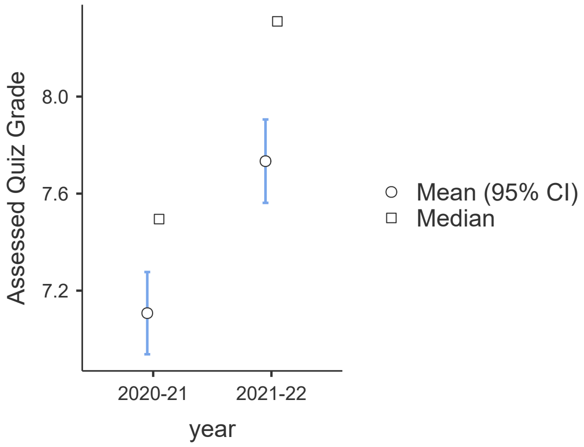


Figure 7.14: Descriptive plots of the assessed quiz for the first cohort 2020/21 and the second cohort 2021/22

The table 7.13 shows the descriptive statistics of Q02 for both 2020/21 and 2021/22.

Year	N	Mean	Median	SD	SE
2020/21	414	7.0	7.0	1.76	0.086
2021/22	394	8.0	8.3	1.74	0.087

Table 7.13: Q02 descriptive statistics for 2020/21 and 2021/22

Year	Q02.8.1: writing $P(n+1)$		Total
	correct ans.	misconception; no eqn.	
2020/21 (L02 optional)	202 (60%)	137 (40%)	339 (100%)
2021/22 (L02 required)	243 (75%)	79 (25%)	322 (100%)
Total	445 (67%)	216 (33%)	661 (100%)

Table 7.14: Chi square analysis of Q02.8.1 for students in the fisrt cohort (2020/21) and the second cohort (2021/22)

A chi-squared test of independence was performed to examine the relation between forcing students to take the lecture quiz and show evidence of the misconception (i.e. writing $P(n+1)$ as a non-equation) in the assessed question Q02.8.1. The relation between these variables was significant, $\chi^2(1, N = 661) = 18.9$, $p < 0.001$. When the lecture quiz was optional in 2020/21, students were more likely to show evidence of the misconceptions than students in 2021/22 when the lecture quiz was required. As shown in the contingency table 7.14, the percentage of the responses with the misconceptions (writing $P(n+1)$ as non-equation) has decreased from 2020/21 to 2021/22.

Year	Q02.8.1			Total
	correct ans.	LHS only	RHS only	
2020/21 (L02 optional)	202 (68%)	49 (16.5%)	46 (15.5%)	297 (100%)
2021/22 (L02 required)	243 (81%)	14 (5%)	41 (14%)	298 (100%)
Total	445 (74.8%)	63 (10.6%)	87 (14.6%)	595 (100%)

Table 7.15: Chi square analysis of Q02.8.1 (LHS and RHS) for the first cohort (2020/21) and the second cohort (2021/22)

7.6.2 Discussion

In this study, I focus exclusively on students' responses, and misconceptions in learning induction proofs. Mathematical induction is used as a vehicle to illustrate the idea of Separated Concerns, discussed in Section 6.4. By separating concerns, a specific issue is explicitly identified and addressed in advance of using it in a more substantial application. This is different from testing the students' knowledge of a topic for its own sake. The findings indicate that engaging with the practice STACK questions that were designed using the Separated Concerns format appears to reduce misconceptions that could arise in the assessed quiz later. Note, the second cohort 2021/22 were required to engage with the lecture quiz and pass with 80%. This change in design was an important part of the process.

The analysis of this work has been carried out using data for the PPS course for two terms 2020/21 and 2021/22. One main goal of this research is to understand students' responses and evaluate a question itself from an academic prospective and ultimately to improve the questions for future years. In 2020/21, only 242 out of 417 (58%) students were involved in participating in the lecture quiz L02, as participating in the lecture quiz was optional. It was found that students who engaged with the lecture quizzes performed better on the assessed quizzes. In 2021/22, the sittings of the lecture quiz were changed in order to encourage students to take the lecture quiz. Students had to do better in the lecture quiz to progress not just had to do it. Consequently, all students were required to complete the lecture quiz scoring at least 80% in order to access the assessed quiz. As a result, the mean score of the lecture quiz has increased in 2021/22 ($Mean = 7.5$) significantly compared to 2020/21 ($Mean = 3.90$).

Surprisingly, the mean score of the assessed quiz increased by 10% in 2021/22 than the previous year (2020/21: $Median = 7.0$, 2021/22: $Median = 8.3$) (see Table 7.13). Moreover, it was observed that when the lecture quiz was optional in 2020/21, students were more likely to show evidence of the misconceptions when answering the assessed quiz than students in 2021/22 when the lecture quiz was required. Accordingly, the results show that the misconceptions that were addressed in the assessed quiz has been decreased in 2021/22 compared to 2020/21.

Another outcome of this work is to use online materials to prepare students so later on students can write a full induction proof. In addition, one goal of this work is to illustrate how to use the basic question usage report, detailing students' responses, for analysis to improve STACK potential response trees. The basic question report in STACK gives all the attempts for any question which really helps to review what students type and to identify some common misconceptions. This study confirmed that the misconceptions identified as concerns occurred, and that they were widespread. The identified misconceptions were generated by more than 10% of the cohort, so it was well worth adding nodes with corresponding outcomes. Although we cannot accurately guess the most common misconceptions, STACK can help us identify common misconceptions and adjust the feedback scheme accordingly by reviewing students' basic usage reports. Consequently, the feedback will be more effective and precise in targeting students' misconceptions.

Based on Cognitive Load Theory, there are cognitive aspects linked to the use of Separating Concerns, which I believe are potentially beneficial. Since learning is strongly effected by the potential of working memory, it has been argued that working memory should be filled by task-relevant operations, especially in learning complex material (Van Gerven et al. 2002). The effect of Separated Concerns occurred when students can transfer and apply their understanding to solve a similar problem in more complicated tasks. By Separating Concerns, students can achieve success on the individual components with explicit and conscious knowledge of where these fit into the more complicated tasks. For example, I separated out tasks involving $\sum$ (sigma) notation in the lecture quiz so that student become generally more familiar, confident and competent before they are asked to apply this notation in an induction proof in the assessment quiz. Previous research in Cognitive Load Theory suggested that its important not to generate high extraneous load, especially when its connected with high intrinsic load, since no cognitive capacity may remain for germane load. The extraneous cognitive load could be reduced by designing effective materials (Sweller, Merriënboer, and Paas 2019). In Separating Concerns, the extraneous cognitive load could be reduced when designing a task that addresses each concern directly. Separating Concerns could lead to an efficient construction of cognitive load, because they focus the attention of learners on specific concerns. For enhancing cognitive skills acquisition, I believe that it is useful to use Separated Concerns before starting to solve more complicated problems.

While the study yielded a clear result, caution must be taken when interpreting the generality of the finding. The following limitations are highlighted. First, the findings apply exclusively and explicitly on applying Separated Concerns when designing STACK questions on learning mathematical induction. Second, the study used a modest sample of students ($N = 414$) from a single course at a single university. Nevertheless, I cannot claim that the sample is representative of the broader population of students undertaking mathematics modules at universities around the world.

In this study, I chose typing online to gather the data for the purpose of developing automatic online assessment when learning mathematics. However, I cannot be certain whether the effect is confounded with typing online as opposed to paper based.

In the first cohort, the proposed questions were designed but the settings of the lecture quiz was not utilized effectively. Accordingly, a large number of the first cohort did not attempt to solve the lecture quiz L02. Skipping some questions, or not answering the whole lecture quiz, resulted in unbalanced comparison among students who took the assessed quiz. For the purpose of this research, the settings of taking the lecture quiz for the next year has been changed from optional to required.

This study tasks placed within an existing course. This was not a pure researched study, e.g. separate quiz questionnaire. The main advantages of this approach was that students answer the tasks seriously. Moreover, the sample size was large, unlike if students are asked to volunteer or to participate in a separate quiz.

In the paper Alarfaj and Sangwin (2020), we reviewed the proof of mathematical induction. This includes summing series and trigonometry identities, and sequences. This review was based on analysis of 20 years of school exams, which revealed that summing series is a particularly one of the common concerns in induction proofs. While acknowledging the potential that our choice of mathematical induction could have biased the conclusions towards summing series, I argue that this choice is representative of how induction is commonly used in practice. Given its widespread application in various mathematical contexts, we believe our focus on induction for summing series remains representative of its real-world usage.

As this study was a part of an exciting course. The decision of taking the lecture quiz was taking by the course organiser and thus it was found that 70% of students in the first cohort did not attempt to answer the lecture quiz as taking the lecture quiz was optional. Consequently, in the second cohort, the setting for taking the lecture quiz changed into required for the purpose of this research. Analysing students responses in the first cohort has opened the door to study more about the misconceptions that students might have in dealing with mathematical induction and also provided us with a way of updating the basic question usage report.

While the "Separated Concerns" approach appears promising, I opted against directly comparing its effectiveness to "non-Separated Concerns" in this study. Drawing upon existing literature, it was evident that students often struggle with the very concerns in Mathematical Induction. Given this prior knowledge, a direct comparison felt redundant, as research on mathematical induction already highlighted such student difficulties.

This is not a surprising result, but the study provides evidence that engaging with the weekly lecture quiz is correlated with successful in the quizzes and successful in the final exams. Furthermore, the results indicate that engaging with the practice materials that designed using Separated Concerns approach can reduce misconceptions that may arise in the quizzes later. In 2020/21, four concerns were identified in learning mathematical induction from students' responses in PPS course. It was found that students sometimes show a fragile understanding of what the statement $P(n+1)$ is. STACK questions were designed to test if students' exhibit this concern. The concern was addressed in a practice lecture quiz in PPS course which was used to prepare for learning mathematical induction in PPS course. STACK potential response trees were updated based on the identified concerns by adding more nodes. Perhaps more time on the original response tree could be spent in advance. Teachers need to consider not only how existing tools can be applied, but also whether they need to be modified (or even replaced) to make them suitable for use in educational settings (Hanna and Yan 2021). The use of computer-based proof technology may offer a new option for overcoming students' difficulty with proofs while at the same time creating an opportunity for mathematics education to reflect the practices of modern mathematics. The introduction of these new tools cannot be successful without considerable thought and research, including a thorough analysis of classroom instruction with and without these tools (Hanna and Yan 2021).

7.6.3 Conclusion

The study contributes to the field of proof instruction in two key ways. Firstly, it demonstrates the effectiveness of online learning materials in preparing students to learn induction proofs. This finding can inform curriculum design and potentially improve student success in proof-based courses. Secondly, the research explores how STACK's basic question usage report can be used to analyze student responses and identify common misconceptions. This analysis can then be used to refine STACK's potential response trees, leading to more targeted and effective feedback that directly addresses student needs. By analyzing student responses and tailoring feedback accordingly, educators gain valuable insights into individual learning needs, enabling them to personalize instruction and maximize student potential.

Chapter 8

Conclusion

The final chapter of this thesis provides some conclusions that one can draw from the research undertaken. First, the methodological contribution of the thesis and the three studies are summarised, detailing the main results of each. Next, some limitations of the three studies are reported.

This thesis introduces the unit coding scheme as a significant methodological contribution. The scheme offers a nuanced approach to assessing written work in online environments by evaluating both the quantity and quality of student explanations and proof construction. This innovative tool has the potential to transform online assessment practices by providing deeper insights into student understanding and fostering more effective learning experiences. The unit coding scheme was adapted from (Hodds, Alcock, and Inglis 2014) and (Toulmin 1958), combined with ideas of structured derivations from (R. J. Back 2010), and combined them to describe students' arguments (See Section 3.2). Toulmin's scheme of argumentation can provide an insight into the logical relationships within a proof and it has been used widely in the mathematics education literature. The difference between the unit coding scheme presented in this thesis and previous uses of the Toulmin's model is that the research reported in this thesis uses the unit coding scheme to break down the components of a mathematical arguments in order to discuss the effect of writing mathematical arguments in different formats using online assessment. Note that the unit coding scheme provides a qualitative method of gaining insight into how many units are typed and what kind of justification are written. The importance of the unit coding scheme is not only considering the number of units but also the quality and relevance of the statement within those units. In this thesis, the unit coding scheme is basically used as a tool for analysing a single argument in students' responses. Although students' arguments are varied and might be wrong. For example, some students might generate wrong warrants to connect their data to the conclusions, or they might give warrants that do not match the calculation steps. In general, the number of units considered a measure of the depth or quality of students' proof construction and that the other marking schemes used are used to balance limitations of this. The unit coding scheme can be applied into many subjects and fields when analysing students arguments particularly for the online assessment purposes. This thesis employs the unit coding scheme as a framework to analyze participants' responses in the two studies presented.

The first study took the suggestion from the literature that students are increasingly being expected to use online assessment systems as support for traditional courses (Sangwin 2013). Many of these online assessment or learning systems are implementing two-column input mechanisms as the examples in Figures 2.8 and 2.7 illustrated. The study investigated whether there is a format effect between traditional and two-column formats when students write mathematical arguments. I undertook an experiment to compare students' writing between unconstrained traditional arguments and arguments in a two-column format. Students from a first year calculus course were invited to take part in the study. Four experimental groups resulted which allowed a comparison to be made between the effects of formats and type of justifications when writing proofs. The analysis of the data obtained revealed there is a difference in students' marks, so that significant format effects can exist. Furthermore, using the unit coding scheme as a guide to analyse the participants' responses, it was found that students in the two-column group produced significantly more higher quality explanations than did the traditional group. It is specific types of explanations that are associated with subsequent marks which is in line with Ainsworth and Burcham (2007).

The results of the first study into the format effects when writing mathematics arguments are a promising indicator. Students are increasingly moving away from paper submission of assignments to working online, a trend accelerated in 2020/21 by the global pandemic. Online submission includes both automated online assessment and online submission of written work for human making. A natural question is therefore this: Is there a difference in performance and justifications between uploading handwritten and typing in writing mathematical responses? The second study investigated the difference and the format effects of typing vs handwriting mathematical responses to simple coursework tasks. Students from a first year of (ILA) course were invited to take place and responded to two relatively short problems, containing equivalent typing and uploading handwritten items. Students' reactions immediately after the tasks were obtained. These were then marked by two markers. Typing and uploading responses were distributed equally between the two markers so each marker dealt with both formats. Factors explored included overall score awarded, justifications provided, and number of steps or "units". The analysis of the data obtained from the study indicated that, overall scores were higher in the uploading handwritten group compared to the typing group. In addition, a higher level of marks is associated with a higher number of units. Students who wrote more tended to get slightly more marks. A secondary outcome of this study was the use and further development of the unit coding scheme.

The third study reported in this thesis illustrates two issues: 1) how engagement with the lecture quiz related to success on the weekly assessed quiz, and the course total, 2) how to explore the common mistakes made by students when using online materials (i.e., STACK questions) to prepare for mathematical induction? Secondary to the above research goals, the study illustrates how to use research to update STACK potential response trees in questions written to support learning mathematical induction based on Separated Concerns. The findings indicate that engaging with the practice STACK questions that were designed using the Separated Concerns format appears to reduce misconceptions that could arise in the assessed quizzes later. The analysis of this work has been carried out using data for the PPS course for two terms 2020/21 and 2021/22. In 2020/21, only 58% students were involved in participating in the lecture quiz, as participating in the lecture quiz was optional. It was found that students who engaged with the lecture quizzes performed better on the assessed quizzes. In 2021/22, the sittings of the lecture quiz were changed in order to encourage students to take the lecture quiz. Consequently, all students were required to complete the lecture quiz scoring at least 80% in order to access the assessed quiz. The results show that the mean score of the lecture quiz has increased in 2021/22 significantly compared to 2020/21. Surprisingly, the mean score of the assessed quiz increased by 10% in 2021/22 than the previous year 2020/21. The study provides evidence that engaging with the weekly lecture quiz is correlated with successful in the quizzes and successful in the final exams. Furthermore, the results indicate that engaging with the practice materials that were designed using Separated Concerns can reduce misconceptions that may arise in the quizzes later. In addition, the research has value for teachers preparing STACK questions to update PRT in STACK more effectively and efficiently. The research suggests that a two-year (or longer) development cycle for STACK questions is highly beneficial. First year cohort: get the question working, with the essential properties for correct/incorrect established reliably. Second year cohort: review students' data and update response trees accordingly, making sure the second (and subsequent) years benefit from better feedback (formative) or more nuanced partial credit (summative). In assessed quizzes, or online exams, marks are not available immediately. So we could improve the PRT before students marks are released to students. However, writing response trees is a significant amount of work. Experience suggests that a lot of time can be expended writing potential tests with feed-back which are never actually used by any significant proportion of the students. Time is then wasted trying to second-guess what students might do. Even then, a review of students' responses is necessary because students write answers we might not actually anticipate. Reviewing students' answers closes the learning cycle for teachers by allowing them to understand what students are doing. Creating online assessments is significant additional work, but once the questions have been created they require minimal work to maintain and can last for the lifetime of the course.

Based on the third study, this thesis contributes to the field of proof instruction in two key ways. Firstly, it demonstrates the effectiveness of online learning materials in preparing students to learn full induction proofs. This finding can inform curriculum design and potentially improve student success in proof-based courses. Secondly, the research explores how the STACK platform's basic question usage report can be used to analyze student responses and identify common misconceptions. This analysis can then be used to refine STACK's potential response trees, leading to more targeted and effective feedback that directly addresses student needs

Having presented the studies in this thesis, their limitations is discussed in the next section.

8.1 Limitations and future work

In this thesis, research was conducted in one university in the UK. As a result, students from different universities and countries will have different mathematical backgrounds. While I cannot claim that the sample is completely representative of the broader population of students undertaking mathematics modules at universities around the world, there is good reason to think that the results will be applicable in some other situations. A blend of pure and applied mathematics is taught in lectures and small group tutorials. Consequently, caution should be exercised about stating which students the effects observed in the studies apply to, but one should be reasonably confident that the format effect would influence the way in which mathematical arguments are written. The format might have an effect on students' marks, the written justification or the number of units. There is a lot of common ground amongst universities and so, despite this limitation, similar format effects are expected in other contexts.

Individual student circumstances may need to be considered in order to understand how the written format might affect students' written work. Students' mathematical backgrounds and prior teaching experiences might differently influence how the format effect exist on students' written arguments.

In study 2, students were assigned into two groups; typing and uploading handwritten responses. The results show that participants in the uploading group scored and provided units significantly higher than those in the typing group. A possible reason for this is that the constraining interface when typing using the editor might limit students' ability to express their reasoning freely using the mathematical symbols as well as some lack of familiarity in typing. Note that this study took place before the widespread change from COVID-19. According to Alarfaj, O'Hagan, and Sangwin (2022), electronic submission of hand-written work saw significant increases in use in 2020/21 compared with before the COVID-19 pandemic. Now students are much more familiar with photographing and uploading responses. Its also expected to conduct fully online examinations in future for many mathematical courses particularly for the method based courses in the earlier years. But the semi-automatic approach is a sensible compromise for better assessment coverage of intended learning outcomes. Therefore, using semi-automatic assessment approach as a means of assessing exams should be supported by practice, since students with no previous experience in an examination-like setting will prefer to stick with what they are familiar with (Mogey, Cowey, et al. 2012). The less experienced typist might find it useful to practice getting an idea of how much content they can generate in a fixed time. Possible avenues for future work involve the design of semi-automatic quizzes where students can submit both typed and handwritten responses. It would be also useful to encourage students to use LaTeX for the first assignment in a course and combine time spent becoming familiar with the technology with guidance or clarification, from tutors, about what constitutes a high-quality examination response. As this study was only from students prospective, It would also be helpful to ask teachers and course organisers about what advantages, difficulties and challenges of conducting semi-automatic tasks and quizzes.

The studies did not explicitly control for prior knowledge, which could have acted as a potential confounding factor and influenced the observed relationship between the format effect and students' performance. Students with varying prior knowledge could have experienced different levels of difficulty with the tasks, potentially inflating results in either format. While we acknowledge this limitation, recognizing the need for future research to explicitly control for prior knowledge using methods like pretests or matched groups.

It is important to acknowledge that the studies is also potentially subject to self-selection bias. Since not all students volunteered to participate, those who participated might have more confident in their mathematical abilities, potentially leading to better performance regardless of format. While acknowledging this potential bias, planning future studies with more representative samples can help reduce its impact on the validity of the findings. Students could be involve in different ways, maybe even choosing them randomly, so the findings reflect the whole class better.

In study 3, the findings apply exclusively and explicitly on applying Separated Concerns when designing STACK questions on learning mathematical induction. However, the Separated Concerns' approach could be applied on different mathematical topics. One limitation of this study is the difficulty of developing materials based on Separated Concerns. It was time consuming to identify common misconceptions in learning induction and to design STACK questions that address these concerns. Despite this limitation, creating online assessments is significant additional work, but once the questions have been created they require minimal work to maintain and can last for the lifetime of the course. It was also shown that the instructional settings of online materials should also be considered. In the first cohort of this study, the proposed questions were designed but the settings of the lecture quiz was not utilized effectively. Accordingly, it was found that 70% of students in the first cohort did not attempt to solve the lecture quiz, skipping some questions, or not answering the whole lecture quiz, resulted in unbalanced comparison among students who took the assessed quiz. Therefore, the settings of taking the lecture quiz for the next year has been changed from optional to required. Despite this limitation, conducting the first cohort of the study opens the door to consider the settings of the practice materials and how it affects the performance in the assessed quiz. The fist cohort helped to understand how to read and analyse raw data in STACK basic questions use report to identify common misconceptions to develop the potential response trees and the feedback scheme for the next year. Furthermore, the findings indicate that engaging with the practice materials can reduce misconceptions that may arise in the quizzes later. The introduction of new tools cannot be successful without considerable thought and research, including a thorough analysis of classroom instruction with and without these tools (Hanna and Yan 2021).

I have focused on my thesis on investigating format effect when students answer mathematical questions, and how to use and design questions for online assessment. From the findings of the research reported in this thesis it can be concluded that format effects exist, and the format does influence many students who participated in these studies when writing mathematical arguments using online assessment. Familiarity, some guidance or training on how to use these formats, might reduce these effects but nevertheless, we would expect some format effects to persist. It is important that teachers are aware of format effects and that material design takes account of effects, where they are understood.

Using online assessment as part of students' learning experience appears to have significantly improved students' performance in mathematics in the short-term, as well as offering potential longer-term benefits. More research will be needed to confirm these findings, given that the studies here involved participants from only one UK university on what would be considered as typical mathematics degree courses for the UK. However, these findings are promising and provide evidence of the effects of different formats on students' learning of mathematics using online assessment.

Other issues, such as societal acceptance of online examinations, institutional governance and conduct of fully online examinations, are very important, however my focus in the near future will be on students' behaviour and how do students interact with an automated feedback, especially investigating students' views about automated feedback. I am also interested to look into teachers' and course designers' attitudes toward designing automated feedback. What are the challenges and the opportunities of designing online tasks with automated feedback? The global pandemic has demonstrated the need for online high-quality assessment and interest in online assessment of mathematics is likely to significantly grow in the near future.

Glossary

Cognitive load theory is a psychological theory about learning built on the premise that since the brain can only do so many things at once, it should be intentional about what questions could be asked to do. The theory was developed in the 1980s by psychologist John Sweller to improve the teaching of mathematics and science. The assumptions of CLT are based on human cognitive architecture, which is characterised by a limited working memory and unlimited long term memory. 20, 21

constraint is a limit or restriction. 27

format is a method of organizing data. 15

knowledge component (KC) an acquired unit of cognitive structure or function that can be implied from performance on a set of related tasks. A knowledge component is used broadly to describe pieces of cognition or knowledge, such as schema or misconception as well as more common terms like concept, principle, fact or skill. 22

Moodle is a learning platform that provides educators, administrators, and learners with a robust, secure, and integrated system for creating individualized learning environments. 24

PPS *Proofs and Problem Solving* is a year 1, semester 2, module run at the University of Edinburgh. PPS is designed to introduce and develop fundamental skills needed for advanced study in Pure Mathematics. This includes precise language and the axiomatic method focusing on definition/theorem/proof. For more detail, see Appendix C. 92

SQA is the Scottish Qualifications and Authority, who are responsible for accrediting and awarding national qualifications in Scotland. See https://www.sqa.org.uk/. 82, 84

STACK a System for Teaching and Assessment using a Computer algebra Kernel, is an open source Computer Aided Assessment (CAA) system for mathematics, and other STEM subjects. The first version of STACK was developed in 2004 by Chris Sangwin in collaboration with Laura Naismith at the University of Birmingham. Since its first release, STACK has been continuously developed and is in widespread use particularly in higher education, notably by The University of Edinburgh, The Open University and Loughborough University. STACK focuses on accepting algebraic input from students. Stack is widely used in its community group. For more information, visit `https://stack-assessment.org/`. 31

two-column is a method of presenting a mathematical proof or argument by using a tabular layout with two-columns. 25

Appendix A

Ethics

For the data presented in this thesis, ethics approval was granted by the University of Edinburgh, School of Mathematics. Thus, the studies conducted were passed and approved by the University's ethics committee.

Appendix B

Materials

B.1 Items used in Study 1: Investigating a potential format effect with two-column proofs

This section contains materials and items for the study in Chapter 4.

B.1.1 Information to Participants

Dear students,

The University of Edinburgh is well-known for developing high-quality teaching, learning and assessment resources in mathematics. One way we achieve this is by undertaking educational research. If you agree, we would like to collect data from your workshop task this week as part of our educational research. The purpose of the research is to better understand students' reasoning and proof writing.

All data will be completely anonymised prior to analysis. Findings from the analysis might be combined with other research findings and published in a peer-reviewed journal or presented at a conference. However names, student numbers and any other personal identifiers will be removed from the data prior to analysis. Importantly, no one but you and your lecturer will know how you performed on the task.

We will make our findings publicly available in due course.

The workshop task is a compulsory part of your course. Whether your anonymised results are included in the data analysis is up to you. Your decision about this has no impact on your grade for this module, or what you are being asked to do.

If you have any questions regarding the study, you can ask the researcher now or later by contacting Maryam Alarfaj at `M.K.H.Alarfaj@sms.ed.ac.uk`

We do hope that you will consider agreeing to your anonymised data being used in the research project so that we can further improve the learning experience for students in the future.

Please choose to agree or disagree on each worksheet.

Thank you very much indeed,

Maryam Alarfaj

B.1.2 Instruction to Tutors

This week two of the workshop tasks will contribute to an educational research project.

The workshop task is a compulsory part of the course. We ask students for consent to include their work in the data analysis.

Please encourage the students to take the tasks seriously, and for their consent for their work to be included.

It is important the two tasks be completed in order as follows:

1. Give Task 1 to students first. Individual work, no textbooks.
2. Collect Task 1. Please make sure name/consent is completed.
3. Give Task 2 to students separately. Individual work, no textbooks.
4. Collect Task 2. Please make sure name/consent is completed.
5. Super tutors to collect folder and return to MA.
6. Continue with the tutorial problems...

We will return the work to you next week, together with worked solutions and some notes, including the rationale for the tasks.

Thank you for your help.

Maryam Alarfaj, Chris Sangwin

Please choose to agree or disagree:
I (agree / disagree) for my anonymised data to be used in the study.

Name: Student ID:

Signature:

Question 1. Explain why the following integral is improper and determine whether it converges or diverges.

$$\int_{1}^{2} \frac{x}{\sqrt{x-1}} \mathrm{d}x$$

Question 2. Explain why the following integral is improper and determine whether it converges or diverges.

$$\int_{-\infty}^{0} xe^{x} \mathrm{d}x$$

B.1.3 Presenting mathematical arguments

A mathematical argument is normally written in *correct* sentences and paragraphs.

Please read the following example carefully to help you understand how you can write a typical mathematical argument yourself.

Example

Evaluate the following using l'Hopital's Rule:

$$\lim_{x\to\infty} \frac{e^x}{x^2}$$

Solution

Consider

$$\lim_{x\to\infty} \frac{e^x}{x^2}.$$

We have $\lim\limits_{x\to\infty} e^x = \infty$ and $\lim\limits_{x\to\infty} x^2 = \infty$, so l'Hopital's Rule gives

$$\lim_{x\to\infty} \frac{e^x}{x^2} = \lim_{x\to\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}x}(e^x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x^2)} = \lim_{x\to\infty} \frac{e^x}{2x}.$$

Since $e^x \to \infty$ and $2x \to \infty$ as $x \to \infty$, the limit on the right side is also indeterminate, but a second application of l'Hopital's gives

$$\lim_{x\to\infty} \frac{e^x}{x^2} = \lim_{x\to\infty} \frac{e^x}{2x} = \lim_{x\to\infty} \frac{e^x}{2} = \infty.$$

Advice

Arguments should be self-contained and so should begin with the information that is provided. When writing an argument use correct and complete sentences to provide both a statement and any justification. It should be clear when you reach the final conclusion, showing that the statement has been proved. You can decide what level of detail is needed, and how to write steps in your argument.

B.1.4 Presenting arguments using a two-column format

A mathematical argument in *two-column format* is written in columns, separating out the statements in the argument from the justifications of those statements.

Please read the following example carefully to help you understand how to use this format yourself.

Example: Use l'Hospital's Rule to evaluate the given limit

$$\lim_{x\to 1} \frac{\ln x}{x-1}$$

Solution (using *two-column format*):

No.	Statement	Justification
1	Consider $\lim_{x\to 1} \frac{\ln x}{x-1}$	Problem statement
2	$\lim_{x\to 1} \ln x = 0$ and $\lim_{x\to 1} x - 1 = 0$	Direct evaluation gives $\frac{0}{0}$ which is an indeterminate form
3	Now consider $\lim_{x\to 1} \frac{\frac{\mathrm{d}}{\mathrm{d}x}\ln(x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x-1)}$	Attempt l'Hospital's Rule
4	$= \lim_{x\to 1} \frac{1/x}{1}$	Evaluate derivatives
5	$= 1$	Evaluate limit, which exists
6	So $\lim_{x\to 1} \frac{\ln x}{x-1} = \lim_{x\to 1} \frac{\frac{\mathrm{d}}{\mathrm{d}x}\ln(x)}{\frac{\mathrm{d}}{\mathrm{d}x}(x-1)}$	Apply l'Hospital's Rule
7	$\lim_{x\to 1} \frac{\ln x}{x-1} = 1$	5 & 6: Conclusion

Advice To use the two-column format, please apply the following steps:

1. Start with the given information
2. Number each step
3. Write justifications in the second column referring to step numbers as needed

You can decide what level of detail is needed, and when to combine small steps into a single step.

B.2 Items used in Study 2: Typing vs. Photograph

This section contains items and materials used in the study in Chapter 5.

Abbreviation	Definition
U_1	The first unit in the proof.
D_1	Data for U_1 in the proof.
C_1	The conclusion of D_1 in the proof.
W_{P_1}	A principle based warrant used to connect D_1 and C_1.
Q_1	Qualifier for C_1.
U_2	The second unit in the proof.
D_2	Data for U_2 in the proof.
U_3	The third unit in the proof.
C_2/D_3	Conclusion of U_2 is the data for U_3.
W_{P_2}	A principle based warrant used to connect D_2 to C_2.
U_4	The fourth unit in the proof.
C_3/D_4	Conclusion of U_3 is the data for U_4.
W_{R_3}	A paraphrasing warrant used to connect D_3 to C_3.
U_5	The fifth unit in the proof.
C_4/D_5	Conclusion of U_4 is the data for U_5.
C_5	Conclusion of D_5 in the proof.
U_6	The sixth unit in the proof.
D_6	Data for U_6 in the proof.
W_{N_6}	A noticing warrant used to connect D_6 to C_6.
U_7	The seventh unit in the proof.
C_6/D_7	Conclusion of U_6 is the data for U_7.
U_8	The eight unit in the proof.
C_7/D_8	Conclusion of U_7 is the data for U_8.
C_8	Conclusion of D_8 in the proof.

Table B.1: Key abbreviations for the model solution

B.2.1 Announcement to students

Dear students,

The University of Edinburgh is well-known for developing high-quality teaching, learning and assessment resources in mathematics. One way we achieve this is by undertaking educational research. We would like to ask you to volunteer in our study to better understand students' reasoning and proof writing. All data will be completely anonymised prior to analysis.

Findings from the analysis might be combined with other research findings and published in a peer-reviewed journal or presented at a conference. However names, student numbers and any other personal identifiers will be removed from the data prior to analysis. Importantly, no one but you and your lecturer will know how you performed on the task.

We will make our findings publicly available in due course.

Whether your anonymised results are included in the data analysis is up to you. Your decision about this has no impact on your grade for this module, or what you are being asked to do.

If you have any questions regarding the study, you can ask the researcher now or later by contacting Maryam Alarfaj at `M.K.H.Alarfaj@sms.ed.ac.uk`

We do hope that you will consider agreeing to your anonymised data being used in the research project so that we can further improve the learning experience for students in the future.

Please choose to agree or disagree on each worksheet.

Thank you very much indeed,

B.2.2 Instruction to Tutors

This quiz of this week will contribute to an educational research project.

The quiz a compulsory part of the course. We ask students for consent to include their work in the data analysis.

Please encourage the students to do the tasks, and for their consent for their work to be included.

Thank you for your help.

Maryam Alarfaj, Chris Sangwin

B.2.3 Instruction to markers

Dear markers,

We would like you to mark students' responses on a task on the topic "Subspaces and Spanning" from the course Introduction to Linear Algebra. The responses are either handwritten or typed.

Please mark the students' work according to the following instructions. In this research study it is important to have consistent marking according to this scheme. We are also interested in whether mark schemes like this (using ideas of data, warrant and conclusion) are reliable and helpful tools for teachers. If you have any questions about this scheme please ask.

1. First mark the answer out of 9 in traditional way, using the "traditional mark scheme" provided in Section B.2.4.

2. Second, work through again applying the "units coding scheme" provided in Section B.2.5.

3. Third, record the total marks and total number of units for each proof. The total number of units is equal to the number of conclusions in the units within the proof. The number of units will tell us how many steps students write to achieve their answer.

B.2.4 Traditional mark scheme

Please mark the answer out of 9 in traditional way, using this mark scheme.

The codes in this mark scheme are taken from a school exam board. E.g. notice we separate out issue like “method” from “accuracy” in this scheme.

1. M Marks awarded for attempting to use a correct Method.
2. (M) Marks awarded for Method; may be implied by correct subsequent working.
3. A Marks awarded for an Answer or for Accuracy: often dependent on preceding M marks.
4. (A) Marks awarded for an Answer or for Accuracy; may be implied by correct subsequent working.
5. R Marks awarded for clear Reasoning.
6. N Marks awarded for correct answers if no working shown.

Accuracy “A” is only awarded when there is evidence of a correct method. If there is no method, do not give “A” marks. If method is not needed, there will be “N” marks instead!

Below is a “model answer” written in a traditional way, and on the next page we have added in marks.

Prove $\text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\} = \text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$.

Response	Marks
Let $U := \text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$. Assume $\mathbf{v} \in W := \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$	M1: Attempt to show $A = B$ by showing $A \subseteq B \wedge B \subseteq A$.
then there exist $a, b, c \in \mathbb{R}$ such that $\mathbf{v} = a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x})$ $= (a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z}$	R1A1
so that if $\mathbf{v} \in W$ then $\mathbf{v} \in U$, i.e. $W \subseteq U$.	R1A1: Reasoning by taking an arbitrary member of W, and accuracy in using the definition of W)
Assume $\mathbf{v} \in U$ there exist $a, b, c \in \mathbb{R}$ such that	M1: Attempt to show $U \subseteq W$
$\mathbf{v} = a\mathbf{x} + b\mathbf{y} + c\mathbf{z}$ $= \frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W$	R1A1: Reasoning by taking an arbitrary member of U, and accuracy in using the definition of U
i.e. $U \subseteq W$.	R1
Total	[9 marks]

B.2.5 Units coding scheme

Please code the data using the following scheme. The scheme divides up a larger argument into smaller self-contained "units". Each unit has data (D), a conclusion (C) and a warrant (W) justifying why the conclusion follows from the data.

- Data (D): typically a statement is coded as data if it directly followed 'consider', 'if', 'let' or if it is mentioned as an obvious fact/hypothesis at the start of this unit without support.
- Conclusion (C): a statement is coded as a conclusion if it followed 'then' or 'therefore' or when it stated as a result of a calculation from previous data.
- Warrant (W): an explanation is coded as a warrant when it is used to connect data to conclusion in a way that explains how the data supported the conclusion.

  There are four categories for warrants

  1. Principle Based Explanation (W_P): when participants provide any explanation based on definitions, theorems, rules not explicitly mentioned in the proof. For example, when a student wrote "This is because by the definition of ..."
  2. Goal-Driven Explanation (W_G): when a participant gave an explanation that related to the structure of the proof (how it is used to reach the goal of the unit or wider proof). For example, student wrote "We use .... to evaluate ..."
  3. Noticing Explanation (W_N): when a participant gave explanation that linked to a previous idea used in the proof. For example, "... this is because in line 5 we used..."
  4. Paraphrasing (W_R): E.g. repeating or paraphrasing a calculation in words that has just been done in algebra. For example, when a student wrote "...separating, simplifying ..." .

To use this scheme identify the first data in the proof, and draw a circle around the data. The first data is usually information given in the question that students start with. Label it D1, and draw a circle around the first conclusion resulting from that data and call it C1 and so on.

The second conclusion is usually coded as the third data, (i.e C_2/D_3) as we did not reach the final answer yet. The final answer will be coded as C_n, where n is the number of the last conclusion in the proofs. Then, we can say that there are n Units in the proof.

Background: we are trying to use a scheme developed by Toulmin, and adapted here for practical marking. Toulmin's scheme identifies "data", "warrants" and "conclusions" in different parts of an argument. If you would like more information about Toulmin's work, we will provide it after you have completed this marking.

B.2.6 Model Answer

Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three linearly independent vectors. Explain, with justification, whether or not

$$\text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\} = \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}.$$

If $U \ = \text{span}\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ in $\mathbb{R}^n$ show that $U = \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$.
Typical response:

Assume $\mathbf{v} \in W := \text{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}$
then there exist $a, b, c \in \mathbb{R}$ such that

$$\begin{aligned}\mathbf{v} &= a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x}) \\ &= (a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z}\end{aligned}$$

so that if $\mathbf{v} \in W$ then $\mathbf{v} \in U$, i.e. $W \subseteq U$.
Assume $\mathbf{v} \in U$ then there exist $a, b, c \in \mathbb{R}$ such that

$$\begin{aligned}\mathbf{v} &= a\mathbf{x} + b\mathbf{y} + c\mathbf{z} \\ &= \frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W\end{aligned}$$

i.e. $U \subseteq W$.

B.2.7 The model answer after applying the Units coding scheme

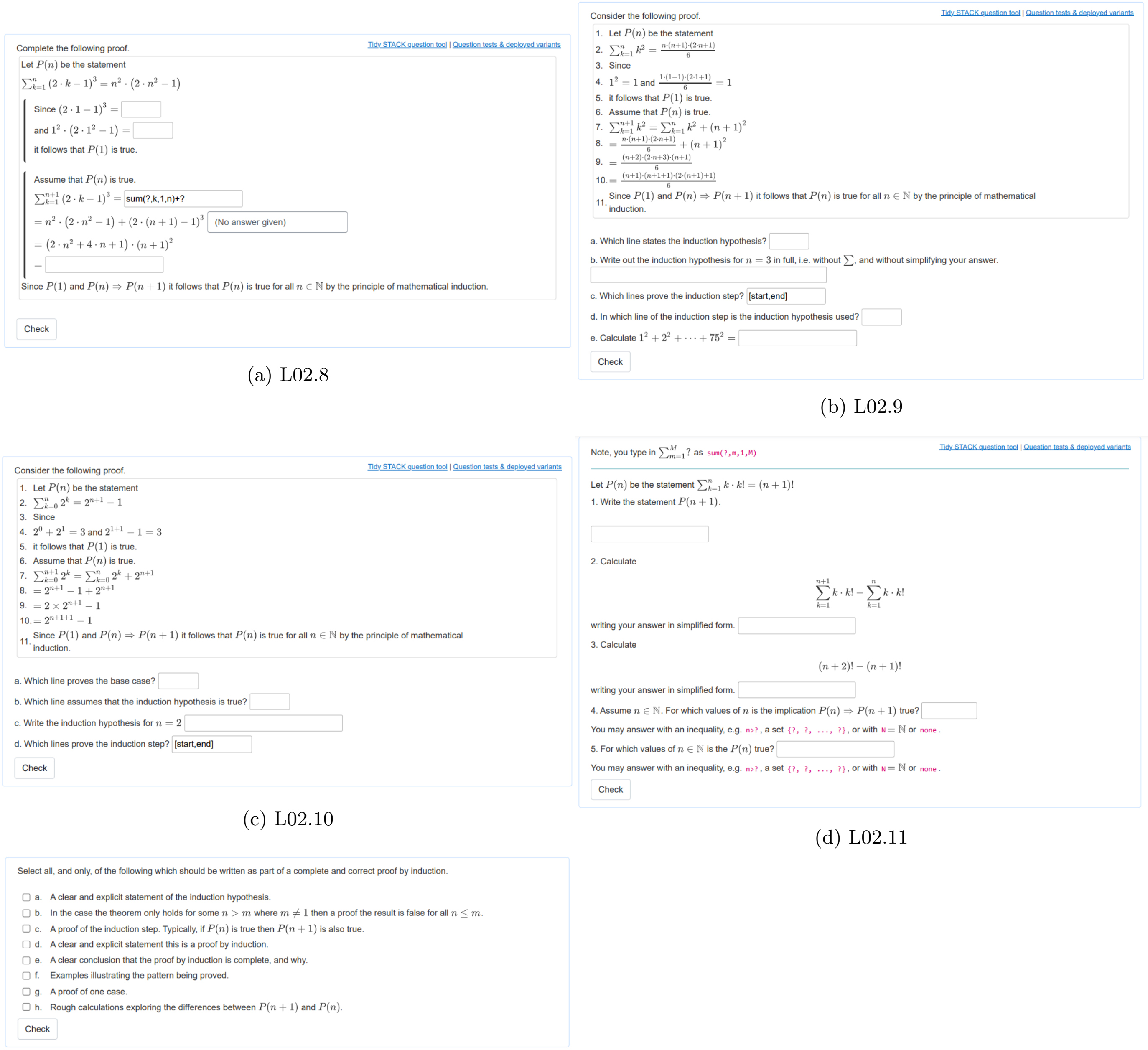
Response	Codes
Assume $\underset{D_1}{\boxed{W := \mathbf{v} \in \operatorname{span}\{\mathbf{x}+\mathbf{y}, \mathbf{y}+\mathbf{z}, \mathbf{z}+\mathbf{x}\}}}$	Data (D1): Attempt to show $A = B$ by showing $A \subseteq B \wedge B \subseteq A$.
then $\underset{C1/D2}{\boxed{\mathbf{v} = a(\mathbf{x}+\mathbf{y}) + b(\mathbf{y}+\mathbf{z}) + c(\mathbf{z}+\mathbf{x})}}$	$C1/D2$: The conclusion for the first unit is the data for the second unit.
$= \underset{C2/D3}{\boxed{(a+c)\mathbf{x} + (a+b)\mathbf{y} + (b+c)\mathbf{z}}}$	$C2/D3$:The conclusion for the second unit is the data for the third unit
$\boxed{\text{so that if } \mathbf{v} \in W \text{ then } \mathbf{v} \in U, \text{ i.e. } W \subseteq U}C3$	C3 : Reasoning by taking an arbitrary member of W, and accuracy in using the definition of W)
$\boxed{\text{Assume } \mathbf{v} \in U}D4$	D4: Attempt to show $U \subseteq W$.
then $\underset{C4/D5}{\boxed{\mathbf{v} = a\mathbf{x} + b\mathbf{y} + c\mathbf{z}}}$	$C4/D5$: The conclusion for the fourth unit is the data for the fifth unit
$= \underset{C5/D6}{\boxed{\frac{a+b-c}{2}(\mathbf{x}+\mathbf{y}) + \frac{b+c-a}{2}(\mathbf{y}+\mathbf{z}) + \frac{a-b+c}{2}(\mathbf{z}+\mathbf{x}) \in W}}$	$C5/D6$: The conclusion for the fifth unit is the data for the sixth unit
i.e. $\boxed{U \subseteq W}C6$	C6: the conclusion of the proof
Total: [6 Units]	The total number of units is 6 since there are 6 conclusions in the proof.

Table B.2: A model solution for Question 1

B.3 Items used in Study 3: Updating STACK PRT based on Separated Concerns

This section contains items and materials used in Study 3 in Chapter 7.

B.3.1 Lecture Quiz L02


Tidy STACK question tool | Question tests & deployed variants
Check

Complete the following proof.

Let $P(n)$ be the statement

$\sum_{k=1}^{n}(2\cdot k-1)^3=n^2\cdot(2\cdot n^2-1)$

Since $(2\cdot 1-1)^3=$ □

and $1^2\cdot(2\cdot 1^2-1)=$ □

it follows that $P(1)$ is true.

Assume that $P(n)$ is true.

$\sum_{k=1}^{n+1}(2\cdot k-1)^3=$ sum(?,k,1,n)+?

$=n^2\cdot(2\cdot n^2-1)+(2\cdot(n+1)-1)^3$ (No answer given)

$=(2\cdot n^2+4\cdot n+1)\cdot(n+1)^2$

$=$ □

Since $P(1)$ and $P(n)\Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n\in\mathbb{N}$ by the principle of mathematical induction.

(a) L02.8

Consider the following proof.

1. Let $P(n)$ be the statement
2. $\sum_{k=1}^{n}k^2=\frac{n\cdot(n+1)\cdot(2\cdot n+1)}{6}$
3. Since
4. $1^2=1$ and $\frac{1\cdot(1+1)\cdot(2\cdot 1+1)}{6}=1$
5. it follows that $P(1)$ is true.
6. Assume that $P(n)$ is true.
7. $\sum_{k=1}^{n+1}k^2=\sum_{k=1}^{n}k^2+(n+1)^2$
8. $=\frac{n\cdot(n+1)\cdot(2\cdot n+1)}{6}+(n+1)^2$
9. $=\frac{(n+2)\cdot(2\cdot n+3)\cdot(n+1)}{6}$
10. $=\frac{(n+1)\cdot(n+1+1)\cdot(2\cdot(n+1)+1)}{6}$
11. Since $P(1)$ and $P(n)\Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n\in\mathbb{N}$ by the principle of mathematical induction.

a. Which line states the induction hypothesis? □

b. Write out the induction hypothesis for $n=3$ in full, i.e. without $\sum$, and without simplifying your answer. □

c. Which lines prove the induction step? [start,end]

d. In which line of the induction step is the induction hypothesis used? □

e. Calculate $1^2+2^2+\cdots+75^2=$ □

(b) L02.9

Consider the following proof.

1. Let $P(n)$ be the statement
2. $\sum_{k=0}^{n}2^k=2^{n+1}-1$
3. Since
4. $2^0+2^1=3$ and $2^{1+1}-1=3$
5. it follows that $P(1)$ is true.
6. Assume that $P(n)$ is true.
7. $\sum_{k=0}^{n+1}2^k=\sum_{k=0}^{n}2^k+2^{n+1}$
8. $=2^{n+1}-1+2^{n+1}$
9. $=2\times 2^{n+1}-1$
10. $=2^{n+1+1}-1$
11. Since $P(1)$ and $P(n)\Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n\in\mathbb{N}$ by the principle of mathematical induction.

a. Which line proves the base case? □

b. Which line assumes that the induction hypothesis is true? □

c. Write the induction hypothesis for $n=2$ □

d. Which lines prove the induction step? [start,end]

(c) L02.10

Note, you type in $\sum_{m=1}^{M}?$ as `sum(?,m,1,M)`

Let $P(n)$ be the statement $\sum_{k=1}^{n}k\cdot k!=(n+1)!$

1. Write the statement $P(n+1)$.

□

2. Calculate

$$\sum_{k=1}^{n+1}k\cdot k!-\sum_{k=1}^{n}k\cdot k!$$

writing your answer in simplified form. □

3. Calculate

$$(n+2)!-(n+1)!$$

writing your answer in simplified form. □

4. Assume $n\in\mathbb{N}$. For which values of n is the implication $P(n)\Rightarrow P(n+1)$ true? □

You may answer with an inequality, e.g. `n>?`, a set `{?, ?, ..., ?}`, or with `N`$=\mathbb{N}$ or `none`.

5. For which values of $n\in\mathbb{N}$ is the $P(n)$ true? □

You may answer with an inequality, e.g. `n>?`, a set `{?, ?, ..., ?}`, or with `N`$=\mathbb{N}$ or `none`.

(d) L02.11

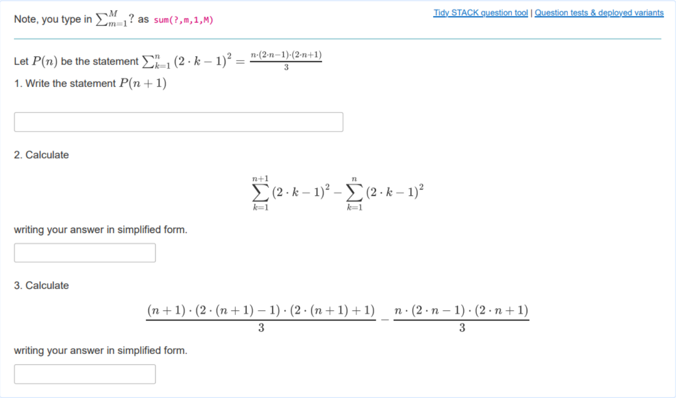
Select all, and only, of the following which should be written as part of a complete and correct proof by induction.

- □ a. A clear and explicit statement of the induction hypothesis.
- □ b. In the case the theorem only holds for some $n>m$ where $m\neq 1$ then a proof the result is false for all $n\leq m$.
- □ c. A proof of the induction step. Typically, if $P(n)$ is true then $P(n+1)$ is also true.
- □ d. A clear and explicit statement this is a proof by induction.
- □ e. A clear conclusion that the proof by induction is complete, and why.
- □ f. Examples illustrating the pattern being proved.
- □ g. A proof of one case.
- □ h. Rough calculations exploring the differences between $P(n+1)$ and $P(n)$.

(e) L02.12

Figure B.1: Lecture Quiz: L02

B.3.2 Assessed Quiz Q02


Tidy STACK question tool | Question tests & deployed variants

Note, you type in $\sum_{m=1}^{M} ?$ as `sum(?,m,1,M)`

Let $P(n)$ be the statement $\sum_{k=1}^{n}(2\cdot k-1)^2=\frac{n\cdot(2\cdot n-1)\cdot(2\cdot n+1)}{3}$

1. Write the statement $P(n+1)$

2. Calculate

$$\sum_{k=1}^{n+1}(2\cdot k-1)^2-\sum_{k=1}^{n}(2\cdot k-1)^2$$

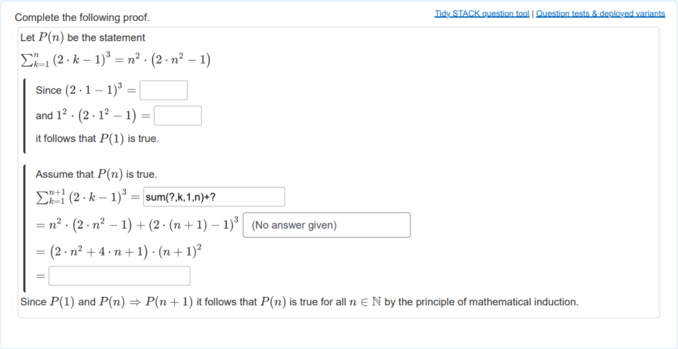
writing your answer in simplified form.

3. Calculate

$$\frac{(n+1)\cdot(2\cdot(n+1)-1)\cdot(2\cdot(n+1)+1)}{3}-\frac{n\cdot(2\cdot n-1)\cdot(2\cdot n+1)}{3}$$

writing your answer in simplified form.

(a) Q02.8


Tidy STACK question tool | Question tests & deployed variants

Complete the following proof.

Let $P(n)$ be the statement

$\sum_{k=1}^{n}(2\cdot k-1)^3=n^2\cdot(2\cdot n^2-1)$

Since $(2\cdot 1-1)^3=$

and $1^2\cdot(2\cdot 1^2-1)=$

it follows that $P(1)$ is true.

Assume that $P(n)$ is true.

$\sum_{k=1}^{n+1}(2\cdot k-1)^3=$ sum(?,k,1,n)+?

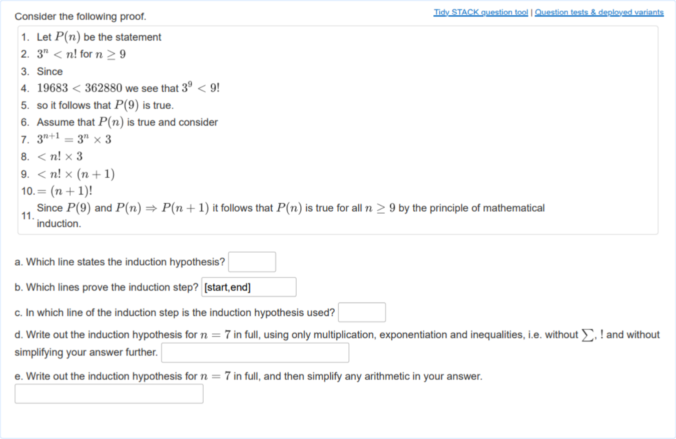
$=n^2\cdot(2\cdot n^2-1)+(2\cdot(n+1)-1)^3$ (No answer given)

$=(2\cdot n^2+4\cdot n+1)\cdot(n+1)^2$

$=$

Since $P(1)$ and $P(n)\Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n\in\mathbb{N}$ by the principle of mathematical induction.

(b) Q02.9


Tidy STACK question tool | Question tests & deployed variants

Consider the following proof.

1. Let $P(n)$ be the statement
2. $3^n<n!$ for $n\geq 9$
3. Since
4. $19683<362880$ we see that $3^9<9!$
5. so it follows that $P(9)$ is true.
6. Assume that $P(n)$ is true and consider
7. $3^{n+1}=3^n\times 3$
8. $<n!\times 3$
9. $<n!\times(n+1)$
10. $=(n+1)!$
11. Since $P(9)$ and $P(n)\Rightarrow P(n+1)$ it follows that $P(n)$ is true for all $n\geq 9$ by the principle of mathematical induction.

a. Which line states the induction hypothesis?

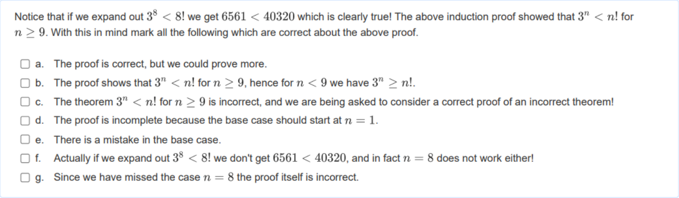
b. Which lines prove the induction step? [start,end]

c. In which line of the induction step is the induction hypothesis used?

d. Write out the induction hypothesis for $n=7$ in full, using only multiplication, exponentiation and inequalities, i.e. without $\sum$, ! and without simplifying your answer further.

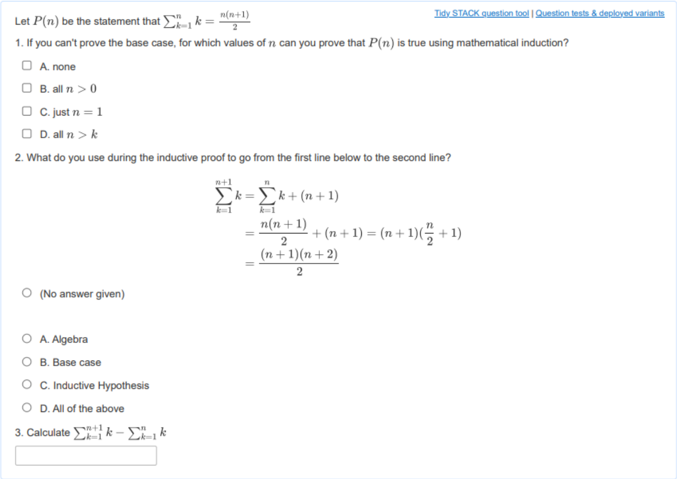
e. Write out the induction hypothesis for $n=7$ in full, and then simplify any arithmetic in your answer.

(c) Q02.10


Notice that if we expand out $3^8<8!$ we get $6561<40320$ which is clearly true! The above induction proof showed that $3^n<n!$ for $n\geq 9$. With this in mind mark all the following which are correct about the above proof.

- ☐ a. The proof is correct, but we could prove more.
- ☐ b. The proof shows that $3^n<n!$ for $n\geq 9$, hence for $n<9$ we have $3^n\geq n!$.
- ☐ c. The theorem $3^n<n!$ for $n\geq 9$ is incorrect, and we are being asked to consider a correct proof of an incorrect theorem!
- ☐ d. The proof is incomplete because the base case should start at $n=1$.
- ☐ e. There is a mistake in the base case.
- ☐ f. Actually if we expand out $3^8<8!$ we don't get $6561<40320$, and in fact $n=8$ does not work either!
- ☐ g. Since we have missed the case $n=8$ the proof itself is incorrect.

(d) Q02.11


Tidy STACK question tool | Question tests & deployed variants

Let $P(n)$ be the statement that $\sum_{k=1}^{n}k=\frac{n(n+1)}{2}$

1. If you can't prove the base case, for which values of n can you prove that $P(n)$ is true using mathematical induction?

- ☐ A. none
- ☐ B. all $n>0$
- ☐ C. just $n=1$
- ☐ D. all $n>k$

2. What do you use during the inductive proof to go from the first line below to the second line?

$$\begin{aligned}\sum_{k=1}^{n+1}k&=\sum_{k=1}^{n}k+(n+1)\\&=\frac{n(n+1)}{2}+(n+1)=(n+1)(\frac{n}{2}+1)\\&=\frac{(n+1)(n+2)}{2}\end{aligned}$$

- ○ (No answer given)

- ○ A. Algebra
- ○ B. Base case
- ○ C. Inductive Hypothesis
- ○ D. All of the above

3. Calculate $\sum_{k=1}^{n+1}k-\sum_{k=1}^{n}k$

(e) Q02.12

Figure B.2: Assessed quiz: Q02

B.3.3 Consent form

We are seeking your consent to use anonymised responses in the online quizzes in PPS this semester as part of a research study into the effectiveness of online assessment.

Participation in this study is voluntary and you have the right to withdraw from it at any point. Participation doesn't affect your grade and participation in this study does not affect what activities you will be asked to do. There are no known risks to participation in this survey.

Of course, we normally do a careful analysis of students' answers with a view to improving teaching. We need to seek your consent to potentially publish an anonymous analysis to help improve mathematics education and help colleagues elsewhere. The data gathered from this survey will be anonymised before the analysis.

The anonymised data will be made publicly available for use in other research.

This project has been approved by the School of Mathematics Ethics procedure. If you have any concerns regarding your own rights as a participant, you can contact the project supervisor, Chris Sangwin (C.J.Sangwin@ed.ac.uk.), or the Deputy Director of Research, Arend Bayer (arend.bayer@ed.ac.uk).

I have read the above and agree to allow my data to be used in this research.

Select one:

True

False

Appendix C

Courses Details

C.1 Calculus and its Applications (CAP)

Calculus is one of the most fundamental tools in mathematics and its applications. This course presents an introduction to the two main branches of calculus: differential calculus and integral calculus. At the heart of both lies the notion of the limit of a function, sequence, or series. In addition to promoting a conceptual appreciation of these foundations of calculus, the course will develop calculational facility, both of which are essential for further mathematical study.

A suggested syllabus for the course is as follows. Functions. Limits and continuity. Differentiation: techniques and applications. Inverse functions. Integration: techniques and applications. Fundamental theorem of calculus. Sequences and series. Taylor and Maclaurin series. Differential equations, moments, and exponential growth.

The course will be assessed 60% on a final examination and 40% on coursework; the coursework component will consist of biweekly written homework (20%) and weekly online quizzes (20%).

C.2 Introduction to Linear Algebra (ILA)

This course picks up on ideas some of which are likely to be familiar from previous study: vectors, matrices and simultaneous equations, but carries them much further. Here are some questions that you will find answers to during the course.

- How do we work with vectors in 4,5 or more dimensions, and why might we want to?
- Are there things like “straight lines” and “planes” in higher dimensions?
- How can we solve 1000 linear equations in 1000 unknowns? And would anybody ever need to?
- Googling linearity gives about the same number of hits as googling ”calculus” (around 10 million). What is linearity and why is it important?
- How could we describe a rotation of 4-dimensional space?
- If A is a $2x2$ matrix, how can we find vectors v with the property that Av is in the same direction as v? (And why is this important?)

This course also starts your transition from school to university mathematics. We will use ”definitions” which are simply precise statements establishing exactly what we mean by a term. A ”theorem” is a statement we believe because we have seen and understood a ”proof”, which is a checkable record of some reasoning that establishes its truth, based often on definitions and previous theorems.

As well as studying known theorems and their proofs we learn how to prove things for ourselves and do a lot of other problem solving.

The fundamental ideas of linear algebra that we study permeate all areas of mathematics, both pure and applied. The concepts and techniques in the course are essential for many advanced mathematics courses and also necessary for a lot of courses in other disciplines that use mathematical techniques. Coursework 100% Examination 0%. The assessment for this course will involve regular coursework throughout the assessment (probably weekly) with a combination of online assessments, written hand-in assessments, and synoptic coursework to be completed at the end of the semester.

C.3 Proofs and Problem Solving (PPS)

Proofs and Problem Solving (PPS) is a year 1, semester 2, module run at the University of Edinburgh. PPS is designed to introduce and develop fundamental skills needed for advanced study in Pure Mathematics. This includes precise language and the axiomatic method focusing on definition/theorem/proof. PPS follows on from a semester 1 course *Introduction to Linear Algebra* (ILA). PPS is a compulsory course for all degree programs in the School of Mathematics, including Mathematics (BSc Hons). The principal areas of study which are both essential foundations to Mathematics and which serve to develop the skills mentioned above are sets and functions, and number systems and their fundamental properties.

In 2020-21 PPS had the following weekly activities.

1. *Weekly course notes* were written by the course team, and a PDF file of approximately 15-20 pages released online on the Friday the week before.
2. *Online live lecture (1).* Monday 13:10–14:00.
3. *Lecture quiz* (STACK). Students were required to complete the quiz between Monday 14:00 and Wednesday 10:00 (between the two live lectures).
4. *Online live lecture (2).* Wednesday 10:00–10:50.
5. *Workshops/tutorials:* Thursday and Fridays. This was an opportunity to meet a tutor in small groups (fewer that 12 students per tutor) for 90 minutes. Workshops have specific unseen tasks related to each week. Tutors also discuss past hand-in work, current material, and answer students' questions.
6. *Assessed quiz* (STACK), was released on Friday at 18:00 after the workshops. The deadline was 10:00 on Wednesday of week + 1.
7. Weekly hand-in work. Students upload their written solutions to problems to Gradescope. Problems were released on Wednesday at 12:00 (after lectures, but before workshops). Work was handed-in Wed 10:00 one week later, and tutors were asked to mark and provide specific feedback one week later.

PPS was assessed using coursework (50% of the final grade, taken from the hand-in and assessed quizzes), and an exam (50% of the final grade). The examination is "force fail", which means students had to pass the exam with a grade of 40% and needed an overall course average of 40% to pass the course. An "alternative assessment" (resit examination) is available in August for students who did not pass the course at a first attempt.

Bibliography

Aberdein, A. (2005). “The Uses of Argument in Mathematics”. In: *Argumentation* 19.3, pp. 287–301. DOI: 10.1007/s10503-005-4417-8.

Aberdein, A. (Mar. 2009). “Mathematics and Argumentation”. In: *Foundations of Science* 14, pp. 1–8. DOI: 10.1007/s10699-008-9158-3.

Ainsworth, S. and S. Burcham (2007). “The impact of text coherence of learning by self-explanation”. In: *Learning and Instruction* 17 (3), pp. 286–303. DOI: 10.1016/j.learninstruc.2007.02.004.

Alarfaj, M. K., S. O’Hagan, and C. J. Sangwin (2022). “Changes made to the teaching of linear algebra and calculus courses in the UK in response to the COVID-19 pandemic”. In: *MSOR Connections*.

Alarfaj, M. K. and C. J. Sangwin (Dec. 2020). “Mathematical Induction in Advanced Higher Mathematics”. In: *Scottish Mathematics Council Journal* 50, pp. 79–86.

Alarfaj, M. K. and C. J. Sangwin (2021). “Investigating a Potential Format Effect with Two-Column Proofs”. In: *Teaching Mathematics and its Applications*. DOI: https://doi.org/10.1093/teamat/hrab028.

Alarfaj, M. K. and C. J. Sangwin (Dec. 2022). “Updating STACK Potential Response Trees Based on Separated Concerns”. In: *International Journal of Emerging Technologies in Learning (iJET)* 17.

Alcock, L. and M. Inglis (2008). “Doctoral students’ use of examples in evaluating and proving conjectures.” In: *Educational Studies in Mathematics*, pp. 111–129.

Alcock, L. and A. Simpson (2009). *Ideas from mathematics education: An introduction for mathematicians*. MSOR Network.

Alcock, L. and N. Wilkinson (2011). “e-Proofs: Design of a Resource to Support Proof Comprehension in Mathematics”. In: *Educational Designer* 1.4, pp. 1–19.

Aleven, V. and K. R. Koedinger (2002). “An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor”. In: *Cognitive Science* 26.2, pp. 147–179. DOI: 10.1016/S0364-0213(02)00061-7.

Arzarello, F. and C. Sabena (2011). “Semiotic and theoretic control in argumentation and proof activities”. In: *Educational Studies in Mathematics* 77.2/3, pp. 189–206. ISSN: 00131954, 15730816. URL: http://www.jstor.org/stable/41485925 (visited on 12/12/2022).

Aviatl, S. and S. Libeskind (1978). “Mathematical induction in the classroom: Didactical and mathematical issues”. In: *Educational Studies in Mathematics* 9, pp. 429–438. URL: https://doi.org/10.1007/BF00410588.

Azzouni, J. (2004). “The Derivation-Indicator View of Mathematical Practice†”. In: *Philosophia Mathematica* 12.2, pp. 81–106. ISSN: 0031-8019. DOI: 10.1093/philmat/12.2.81.

Back, R., J. Grundy, and J. Von Wright (1997). “Structured calculational proof”. In: *Formal Aspects of Computing* 9, pp. 469–483. DOI: https://doi.org/10.1007/BF01211456.

Back, R. J. (2010). “Structured derivations: a unified proof style for teaching mathematics”. In: *Formal Aspects of Computing* 22.5, pp. 629–661. DOI: 10.1007/s00165-009-0136-5.

Back, R. J. (2016). *Teaching Mathematics in the Digital Age with Structured Derivations*. Turku, Finland: Four Ferries Publishing.

Back, R. J., L. Mannila, and S. Wallin (2010). “‘It takes me longer, but I understand better’ – student feedback on structured derivations”. In: *International Journal of Mathematical Education in Science and Technology* 41.5, pp. 575–593. DOI: 10.1080/00207391003605221.

Baker, J. D. (1996). “Students’ Difficulties with Proof by Mathematical Induction”. In: *Proceedings of the Annual Meeting of the American Educational Reseach Association*.

Beeson, M. (1998). “Design Principles of Mathpert: Software to support education in algebra and calculus”. In: *Computer-Human Interaction in Symbolic Computation*. Ed. by N. Kajler. Texts & Monographs in Symbolic Computation. Vienna, Austria: Springer-Verlag, pp. 89–115. DOI: 10.1007/978-3-7091-6461-7.

Brancker, T., J. Pell, and J. H. Rahn (1668). *An Introduction to Algebra*. London, UK: Printed by W.G. for Moses Pitt.

Brousseau, G. (1997). *Theory of Didactical Situations in Mathematics: didactiques des mathématiques, 1970–1990*. N. Balacheff, M. Cooper, R. Sutherland, and V. Warfield (Trans.) Kluwer.

Brown, J. R. (2008). *Philosophy of Mathematics: A Contemporary Introduction to the World of Proofs and Pictures*. New York: Routledge. DOI: 10.4324/9780203932964.

Chase, C. I. (1968). “The impact of some obvious variables on essay test scores”. In: *Journal of Educational Measurement* 5, pp. 315–318.

Chase, C. I. (1986). “Essay test scoring: Interaction of relevant variables”. In: *Journal of Educational Measurement* 23, pp. 33–41.

Chi, M. et al. (1989). “Self-explanation: how students study and use examples in learning to solve problems”. In: *Cognitive Science* 13.2, pp. 145–183. DOI: 10.1207/s15516709cog1302_1.

Cooper, B. and M. Dunne (2000). *Assessing Children’s Mathematical Knowledge: Social Class, Sex and Problem-Solving*. Open University Press.

Crisp, G. (2007). *The e-Assessment Handbook*. London, Continuum.

Delius, G. (2004). “Conservative approach to computerised marking of mathematics assignments”. In: *MSOR Connections*, pp. 42–47.

Engelbrecht, J. and A. Harding (2005). “Teaching undergraduate mathematics on the Internet. Part 1: Technologies and taxonomy”. In: *Educational Studies in Mathematics*, pp. 235–252.

Ericsson, K. A. et al., eds. (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, England ; New York, N.Y.: Cambridge University Press. ISBN: 052184097X.

Ernest, P. (May 1984). “Mathematical induction: A pedagogical discussion”. In: *Educational Studies in Mathematics* 15.2, pp. 173–189. DOI: 10.1007/BF00305895.

Evens, H. and J. Houssart (2004). “Categorizing pupils’ written answers to a mathematics test question: ’I know but I can’t explain’”. In: *Educational Research* 46.3, pp. 269–282. DOI: 10.1080/0013188042000277331. eprint: https://doi.org/10.1080/0013188042000277331. URL: https://doi.org/10.1080/0013188042000277331.

Haese, M. et al. (2019). *Mathematics: Core Topics HL*. Adelaide, Australia: Haese Mathematics. ISBN: 978-1-925489-58-3.

Hanna, G. (1991). “Mathmatical proof”. In: *Advanced mathematical thinking*. Ed. by D. Tall. Kluwer Academic Publishers. Chap. 4, pp. 54–61.

Hanna, G. and X. Yan (Nov. 2021). “Opening a Discussion on Teaching Proof with Automated Theorem Provers”. In: *For the Learning of Mathematics* 41.3, pp. 42–46.

Hassmén, P. and D. P. Hunt (1994). “Human self-assessment in multiple choice”. In: *Journal of Educational Measurement* 31.2, pp. 149–160.

Herbst, P. G. (2002). “Establishing a custom of proving in American school geometry: evolution of the two-column proof in the early twentieth century”. In: *Educational Studies in Mathematics* 49.3, pp. 283–312. DOI: 10.1023/A:102026490.

Hersh, R. (Dec. 1993). “Proving is convincing and explaining”. In: *Educational Studies in Mathematics* 24, pp. 389–399. DOI: 10.1007/BF01273372.

Hodds, M. (June 2014). “Improving Proof Comprehension in Undergraduate Mathematics”. PhD thesis.

Hodds, M., L. Alcock, and M. Inglis (2014). “Self-Explanation Training Improves Proof Comprehension”. In: *Journal for Research in Mathematics Education* 45.1, pp. 62–101. DOI: 10.5951/jresematheduc.45.1.0062.

Howard, J. and J. Beyers, eds. (2020). *Teaching and Learning Mathematics Online*. New York: Chapman and Hall/CRC. DOI: 10.1201/9781351245586.

Hunt, T. J. (2012). “Computer-Marked Assessment in Moodle: Past, Present and Future”. In: *In Proceedings of CAA 2012 International Conference, Southampton*.

Iannone, P. et al. (2011). “Does generating examples aid proof production?” In: *Educational Studies in Mathematics* 77, pp. 1–14.

Ingleby, E. (July 2012). “Research methods in education, Cohen, L. L. Manion, and K. Morrison. Professional Development in Education 2012, 38, 3, 507-509”. In: *Professional Development in Education* 38. DOI: 10.1080/19415257.2011.643130.

Inglis, M., J. P. Mejia-Ramos, and A. Simpson (2007). “Modelling mathematical argumentation: the importance of qualification”. In: *Educational Studies in Mathematics* 66, pp. 3–21. DOI: 10.1007/s10649-006-9059-8.

Inglis, M. and J. P. Mejía-Ramos (Nov. 2009). “On the persuasiveness of visual arguments in mathematics”. In: *Foundations of Science* 14.1–2, pp. 97–110. DOI: 10.1007/s10699-008-9149-4. URL: https://link.springer.com/article/10.1007/s10699-008-9149-4.

James, A. W. (1927). “The effect of handwriting on grading”. In: *English Journal* 16, pp. 180–205.

Jordan, S. (2013). “E-assessment: Past, present and future”. In: *New Directions in the Teaching of Physical Sciences* 9.1, pp. 87–106. DOI: 10.11120/ndir.2013.00009.

Kalyuga, S., P. Ayres, et al. (2003). “The expertise reversal effect”. In: *Educational Psychologist* 38.1, pp. 23–31. DOI: 10.1207/S15326985EP3801_4.

Kalyuga, S., R. Rikers, and F. Paas (2012). “Educational Implications of Expertise Reversal Effects in Learning and Performance of Complex Cognitive and Sensorimotor Skills”. In: *Educational Psychology Review* 24.2, pp. 313–337. DOI: 10.1007/s10648-012-9195-x.

Kinnear, G. (2018). “Improving an online diagnostic test via item analysis”. In: *Proceedings of the Fifth ERME Topic Conference on Mathematics Education in the Digital Age*. University of Copenhagen, pp. 315–316.

Kinnear, G. (2019). “Delivering an online course using STACK”. In: *Contributions to the 1st International STACK conference 2018 in Fürth, Germany*. Zenodo. DOI: 10.5281/zenodo.2565969. URL: https://doi.org/10.5281/zenodo.2565969.

Kinnear, G., A. K. Wood, and R. Gratwick (2021). “Designing and evaluating an online course to support transition to university mathematics”. In: *International Journal of Mathematical Education in Science and Technology* 0.0, pp. 1–24. DOI: 10.1080/0020739X.2021.1962554.

Kirschner, P.A., J. Sweller, and R.E. Clark (2006). “Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching”. In: *Educational psychologist* 41.2, pp. 75–86.

Koedinger, K. R., A. T. Corbett, and C. Perfetti (2012). “The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning”. In: *Cognitive Science* 36, pp. 757–798. DOI: 10.1111/j.1551-6709.2012.01245.x.

Krause, U., R. Stark, and H. Mandl (2009). “The effects of cooperative learning and feedback on e-learning in statistics”. In: *Learning and Instruction* 19.2, pp. 158–170. ISSN: 0959-4752. DOI: https://doi.org/10.1016/j.learninstruc.2008.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0959475208000376.

Krummheuer, G. (1995). “The ethnography of argumentation. In P. Cobb and H. Bauersfeld (Eds.), The emergence of mathematical meaning, interactions in classroom cultures”. In: pp. 229–269.

Kuechler, W. and M. Simkin (2003). “How well do multiple choice tests evaluate student understanding in computer programming classes”. In: *Journal of Information Systems Education*, pp. 389–399.

Laming, D. (2004). “Marking University Examinations: Some Lessons from Psychophysics”. In: *Psychology Learning & Teaching* 3.2, pp. 89–96. DOI: 10.2304/plat.2003.3.2.89.

Leron, U. (1983). “Structuring Mathematical Proofs”. In: *The American Mathematical Monthly* 90.3, pp. 174–185. DOI: 10.2307/2975544.

Livingston, S. A. and S. L. Rupp (Nov. 2004). *Performance of Men and Women on Multiple-Choice and Constructed-Response Tests for Beginning Teachers*. Research Report 04-48. Educational Testing Services.

Markham, L. R. (1976). “Influences of handwriting quality on teacher evaluation of written work”. In: *American Educational Research Journal* 13, pp. 277–283.

Mason, J. and S. Johnston-Wilder (2004). *Fundamental constructs in mathematics education*. Routledge Falmer.

Mazzeo, J., A. P. Schmitt, and C. A. Bleistein (1993). *Sex-related performance differences on constructed-response and multiple-choice selections of Advanced Placement Examinations*. College Board Report 92-7. College Entrance Examination Board, New York.

Mejia-Ramos, J. P., E. Fuller, et al. (2012). “An assessment model for proof comprehension in undergraduate mathematics”. In: *Educational Studies in Mathematics* 79.1, pp. 3–18. DOI: doi.org/10.1007/s10649-011-9349-7.

Mejia-Ramos, J. P., K. Lew, et al. (2017). “Developing and validating proof comprehension tests in undergraduate mathematics”. In: *Research in Mathematics Education* 19.2, pp. 130–146. DOI: 10.1080/14794802.2017.1325776.

Michaelson, M. (2008). In: *Australian Senior Mathematics Journal* 22.2, pp. 57–62. URL: https://search.informit.org/doi/10.3316/informit.410329014071111.

Michalewicz, Z. and M. Michalewicz (2008). *Puzzle-based Learning: Introduction to critical thinking, mathematics, and problem solving*. Hybrid Publishers.

Mogey, N., J. Cowey, et al. (2012). “Students’ choices between typing and handwriting in examinations”. In: *Active Learning in Higher Education* 13.2, pp. 117–128. DOI: 10.1177/1469787412441297.

Mogey, N., J. Paterson, et al. (2010). “Typing compared with handwriting for essay examinations at university: letting the students choose”. In: *ALT-J, Research in Learning Technology* 18.1, pp. 29–47. DOI: 10.1080/09687761003657580.

Moons, F., E. Vandervieren, and J. Colpaert (2022). “Atomic, reusable feedback: a semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers.” In: *Computers and Education Open* 3, p. 100086. ISSN: 2666-5573. DOI: https://doi.org/10.1016/j.caeo.2022.100086. URL: https://www.sciencedirect.com/science/article/pii/S2666557322000143.

Newman, M. H. A. and et.al. (1957). *The teaching of algebra in sixth forms: a report prepared for the Mathematical Association*. London, UK: G. Bell and sons, Ltd.

Nihalani, P., M. Mayrath, and D. Robinson (Nov. 2011). “When Feedback Harms and Collaboration Helps in Computer Simulation Environments: An Expertise Reversal Effect”. In: *Journal of Educational Psychology - J EDUC PSYCHOL* 103. DOI: 10.1037/a0025276.

Panza, M. (2003). “Mathematical Proofs”. In: *Synthese* 134.1/2, pp. 119–158. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20117328 (visited on 11/20/2022).

Pelc, A. (Oct. 2008). “Why Do We Believe Theorems?†”. In: *Philosophia Mathematica* 17.1, pp. 84–94. ISSN: 0031-8019. DOI: 10.1093/philmat/nkn030. eprint: https://academic.oup.com/philmat/article-pdf/17/1/84/4238129/nkn030.pdf. URL: https://doi.org/10.1093/philmat/nkn030.

Pitcher, N., J. Goldfinch, and C. Beevers (2002). “Aspects of computer-based assessment in mathematics”. In: *Active Learning in Higher Education*, pp. 159–178.

Poole, D. (2011). *Linear Algebra: a modern approach*. Third. Brooks/Cole, Cengage learning.

Rasmussen, C. and M. Stephen (2007). “Modeling mathematical argumentation: the importance of qualification”. In: *Educational Studies in Mathathematics* 66.1, pp. 3–21. DOI: 10.1007/s10649-006-9059-8.

Rav, Y. (2007). “A critique of a formalist-mechanist version of the justification of arguments in mathematicians proof practices. ” In: *Philosophia Mathematica* 15.3, pp. 291–320. URL: https://doi.org/10.1093/philmat/nkm023.

Renkl, A. (1997). “Learning from worked-out examples: a study on individual differences”. In: *Cognitive Science* 21.1, pp. 1–29. DOI: 10.1016/S0364-0213(99)80017-2.

Renkl, A. (2002). “Worked-out examples: instructional explanations support learning by self-explanations”. In: *Learning and Instruction* 12.5, pp. 529–556. DOI: 10.1016/S0959-4752(01)00030-5.

Renkl, A., R. K. Atkinson, and C. S. Gross (2004). “How Fading Worked Solution Steps Works – A Cognitive Load Perspective”. In: *Instructional Science* 32, pp. 59–82. DOI: 10.1023/B:TRUC.0000021815.74806.f6.

Richey, J. Elizabeth and Timothy J. Nokes-Malach (2013). “How much is too much? Learning and motivation effects of adding instructional explanations to worked examples”. In: *Learning and Instruction* 25, pp. 104–124. ISSN: 0959-4752. DOI: https://doi.org/10.1016/j.learninstruc.2012.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0959475212001016.

Rønning, Frode (Feb. 2017). “Influence of computer-aided assessment on ways of working with mathematics”. In: *Teaching Mathematics and its Applications: An International Journal of the IMA* 36.2, pp. 94–107. ISSN: 0268-3679. DOI: 10.1093/teamat/hrx001. eprint: https://academic.oup.com/teamat/article-pdf/36/2/94/17724097/hrx001.pdf. URL: https://doi.org/10.1093/teamat/hrx001.

Rowland, T. (2002). “Generic proofs in number theory”. In: *Learning and Teaching Number Theory: Research in Cognition and Instruction*. Ed. by S. Campbell and R. Zazkis. Westport, CT: Ablex Publishing, pp. 157–184.

Roy, S., L. Alcock, and M. Inglis (2010). "Undergraduates Proof Comprehension: A Comparative Study of Three Forms of Proof Presentation". In: *Proceedings of the 13th Conference on Research in Undergraduate Mathematics Education.*

Russell, M. and W. Tao (2004). "Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum and Ramsey". In: *Practical Assessment, Research and Evaluation* 1.

Sangwin, C. J. (2004). "Assessing mathematics automatically using computer algebra and the internet". In: *Teaching Mathematics and its Applications* 23.1, pp. 1–14. DOI: 10.1093/teamat/23.1.1.

Sangwin, C. J. (2013). *Computer Aided Assessment of Mathematics.* Oxford, UK: Oxford University Press. ISBN: 978-0-19-966035-3.

Sangwin, C. J. and I. Jones (2017). "Asymmetry in student achievement on multiple choice and constructed response items in reversible mathematics processes". In: *Educational Studies in Mathematics* 94, pp. 205–222. DOI: 10.1007/s10649-016-9725-4.

Sangwin, C. J. and G. Kinnear (2022). "Coherently Organized Digital Exercises and Expositions". In: *PRIMUS* 32.8, pp. 927–938. DOI: 10.1080/10511970.2021.1999352. eprint: https://doi.org/10.1080/10511970.2021.1999352. URL: https://doi.org/10.1080/10511970.2021.1999352.

Sangwin, C. J. and N. Köcher (2016). "Automation of mathematics examinations". In: *Computers and Education* 94, pp. 215–227. DOI: 10.1016/j.compedu.2015.11.014.

Sangwin, C. J. and P. Ramsden (2007). "Linear syntax for communicating elementary mathematics". In: *Journal of Symbolic Computation* 42.9, pp. 902–934. DOI: 10.1016/j.jsc.2007.07.002.

Schulze, A. and F. Sevenoak (1913). *Plane Geometry.* Revised. MacMillan.

Selden, A. and J Selden (2003). "Validations of Proofs Considered as Texts: Can Undergraduates Tell Whether an Argument Proves a Theorem?" In: *Journal for Research in Mathematics Education* 34.1, pp. 4–36. (Visited on 11/20/2022).

Simpson, A. (2015). "The anatomy of a mathematical proof: Implications for analyses with Toulmin's scheme". In: *Educational Studies in Mathematics* 90, pp. 1–17. DOI: 10.1007/s10649-015-9616-0.

Singh, A. et al. (2017). "Gradescope: A Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work". In: *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale.* L@S '17. Cambridge, Massachusetts, USA: Association for Computing Machinery, pp. 81–88. ISBN: 9781450344500. DOI: 10.1145/3051457.3051466. URL: https://doi.org/10.1145/3051457.3051466.

Soloff, S. (1973). "Effect of non-content factors on the grading of essays". In: *Graduate Research in Education and Related Disciplines* 6, pp. 44–54.

Stedall, J. A. (2002). *A discourse concerning algebra: English algebra to 1685.* Oxford, UK: Oxford University Press.

Stewart, J. (2007). *Essential Calculus: Early Trancendentals.* International Student Edition. Thomson.

Stylianides, G. J. (2007). "Proof and Proving in School Mathematics". In: *Journal for Research in Mathematics Education* 38.3, pp. 289–321.

Stylianides, G. J., J. Sandefur, and A. Watson (2016). "Conditions for proving by mathematical induction to be explanatory". In: *The Journal of Mathematical Behavior* 43, pp. 20–34. DOI: 10.1016/j.jmathb.2016.04.002.

Sweller, J. (Apr. 2006). "The worked example effect and human cognition". In: *Learning and Instruction - LEARN INSTR* 16, pp. 165–169. DOI: 10.1016/j.learninstruc.2006.02.005.

Sweller, J., P. Ayres, and S. Kalyuga (2011). *Cognitive Load Theory.* Springer. ISBN: 10.4018/978-1-60566-014-1.ch084.

Sweller, J. and G. Cooper (1985). "The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra". In: *Cognition and Instruction* 2.1, pp. 59–89. DOI: 10.1207/s1532690xci0201_3. URL: https://doi.org/10.1207/s1532690xci0201_3.

Sweller, J., J. van Merriënboer, and F. Paas (2019). "Cognitive Architecture and Instructional Design: 20 Years Later." In: *Educational Psychology Review* 31.2. DOI: 10.1007/s10648-019-09465-5.

Thompson, D. (1996). "Learning and Teaching Indirect Proof". In: *The Mathematics Teacher* 89.6, pp. 474–482. URL: http://www.jstor.org/stable/41485925.

Toulmin, S. E. (1958). *The Uses of Argument.* Cambridge, United Kingdom: Cambridge University Press.

Tuovinen, J. and J. Sweller (1999). "A comparison of cognitive load associated with discovery learning and worked examples". In: *Journal of Educational Psychology* 91, pp. 334–341.

Van Gerven, P.W.M et al. (2002). “Cognitive load theory and aging: effects of worked examples on training efficiency”. In: *Learning and Instruction* 12.1, pp. 87–105.

Weber, K. (2015). “Effective Proof Reading Strategies for Comprehending Mathematical Proofs”. In: *International Journal of Research in Undergraduate Mathematics Education* 1.3, pp. 289–314. DOI: 10.1007/s40753-015-0011-0.

Weiss, M., P. Herbst, and C. Chen (2009). “Teachers’ perspectives on “authentic mathematics” and the two-column proof form”. In: *Educational Studies in Mathematics* 70.3, pp. 275–293. DOI: 10.1007/s10649-008-9144-2.

Wittwer, J. and A. Renkl (2010). “How Effective are Instructional Explanations in Example-Based Learning? A Meta-Analytic Review.” In: *Educ Psychol Rev* 22, pp. 393–409. DOI: https://doi.org/10.1007/s10648-010-9136-5.

Wolfe, M. B. W. and S. R. Goldman (2005). “Relations between adolescents’ text processing and reasoning”. In: *Cognition and Instruction* 23.4, pp. 467–502. DOI: 10.1207/s1532690xci2304_2.

Yang, Kai-Lin and Fou-Lai Lin (2008). “A Model of Reading Comprehension of Geometry Proof”. In: *Educational Studies in Mathematics* 67.1, pp. 59–76. ISSN: 00131954, 15730816. URL: http://www.jstor.org/stable/40284640 (visited on 01/30/2023).

Young, J. W. A. (1908). “On Mathematical Induction”. In: *The American Mathematical Monthly* 15.8, pp. 145–153. DOI: 10.2307/2969864.