# Comparative Judgment and Proof

by

**Benjamin Davies**

A Doctoral Thesis

Submitted in partial fulfillment of the requirements
for the award of
Doctor of Philosophy of Loughborough University

November 12, 2019

# Abstract

Proof is a central concept in mathematics, pivotal both to the practice of mathematicians and to students' education in the discipline. The research community, however, has failed to reach a consensus on how proof should be conceptualised. Moreover, we know little of what mathematicians and students think about proof, and are limited in the tools we use to assess students' understanding.

This thesis introduces comparative judgment to the proof literature via two tasks evaluated by judges performing a series of pairwise comparisons. The Conceptions Task asks for a written explanation of what mathematicians mean by proof. The Summary Task asks for a summary of a given proof, available to respondents as they complete the task.

Having established robust evidence supporting the reliability and validity of both tasks, I then use these tasks to develop an understanding of the conceptions of proof held by mathematicians and students. I also generate insights for assessment, leading to an argument for the unidimensionality of proof comprehension in early undergraduate mathematics.

In conducting this research I adopt a mixed methods approach based on the philosophy of pragmatism. By using a range of methodological approaches, from statistical modelling to thematic analysis of interviews with judges, I develop a multi-faceted understanding of both the validity of the tasks, and the behaviours and priorities of the participants involved.

The Conceptions Task outcomes establish that mathematicians primarily think of proof in terms of argumentation, while students emphasise the arguably more philosophically naive notion of certainty.

The Summary Task outcomes establish that references to the method of proof and key mathematical objects are most valued by mathematician judges. Further, from correlational analyses of various quantitative measures, I learn that the Summary Task scores are meaningfully reflective of local proof comprehension but are not related to more general measures of mathematical performance.

Several open questions are identified. In particular, there is still much to learn about judges' decision-making processes in comparative judgment settings, the dimensionality of proof comprehension, and the range of proofs for which the Summary Task is applicable. Future work on these questions is outlined in the final chapter, alongside the practical applications and theoretical implications of this work.

# Acknowledgements

First, I would like to thank my supervisors, Ian Jones and Lara Alcock. You have both been extremely influential in my academic and personal growth over the last three years and I will be forever grateful for your guidance and support in producing this thesis. I would also like to thank the entire Mathematics Education Centre. It has been a privilege to be surrounded by such a welcoming, passionate and diverse group of researchers.

To Pablo Mejia-Ramos and colleagues at Rutgers University, thank you for the enthusiastic welcome during my brief stay. I thoroughly enjoyed my time and took away many exciting new ideas. I look forward to crossing paths in the future.

Next, thank you to my parents, Emma and John, and my sister, Jess, for your love and support over the last 3 years, as well as the preceding 22. While none of you share (or perhaps understand) the joy I get from mathematics, you have been unwavering in your enthusiasm about my enthusiasm and have always supported me to pursue the ideas that excite me most.

I also want to thank my Bubba, who helped teach me to read and who has always encouraged my love of mathematics. I will always remember being picked up from a Fibonacci workshop in Takapuna, eager to explain the beauty of constructing a spiral. I will always be grateful for the way you helped foster that excitement and, like my parents, encouraged me to pursue the ideas in which I find the most joy. It is fitting that one of three proofs integral to my thesis invokes the same sequence that we discussed 15 years ago.

Finally, to my partner and best friend, Adam Clearwater, thank you for everything. Thank you for spending weekends helping me think through various analyses, finding typos or correcting my use-of hyphens. Thank you for your patience during long evenings and for your unwavering faith in my ability to finish this project. I am extremely grateful for what we have and I look forward to navigating the next chapter of our lives together.

# Declaration

I, the author, declare that the work presented in this thesis is my own and has not been submitted for a degree at any other institution. None of the work has previously been published in this form.

# Contents

**Part One**

**Part Two**

**Part Three**

Part Four

Appendices

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Proof is central to mathematics and has attracted substantial attention from the education community. Recent research on proof comprehension has identified many aspects of proof that students find difficult, and many plausible accounts for those observed difficulties. Yet, despite its importance, we have very few accounts of what students and mathematicians explicitly say about proof (Stylianou et al., 2015; Weber and Czocher, 2019) and the tools we use to capture students' understanding of proof are understood by mathematicians and educators to be inadequate (Mejia-Ramos et al., 2017).

Convergence on a singular definition of proof has eluded mathematicians and philosophers for as long as deductive reasoning has existed in mathematics. Various definitions from the strictly formalistic (Hilbert, 1931) to the more socially oriented, context-dependent (Davis and Hersh, 1981), have been proposed. However, for every reasonable definition, it seems there is an equally reasonable counter-example or counter-argument (Weber et al., 2014a; Czocher and Weber, in press). This apparent epistemic diversity can be viewed as problematic for mathematics education, as it limits the ability of researchers to build on one another's work due to the unexplored variety of implicitly adopted conceptions of proof (Balacheff, 2008). Others have questioned the extent and importance of the diversity. For example, Weber and Czocher (2019) reported that mathematicians agreed on a proof verification task for 'typical cases' (p. 12), but disagreed on fringe cases like computer-based and visual proofs.

Perhaps as a result of the difficulty of defining proof, researchers have also struggled to generate reliable and valid measures of proof comprehension. Advanced mathematics is often taught following a 'definition-theorem-proof' structure (Weber and Alcock, 2004; Moore, 1994; Davis and Hersh, 1981) and a large proportion of traditional assessment focuses on written constructions of proofs, either identical or similar to those presented in class (Weber, 2012; Rowland, 2001). Such assessments have been criticised for their over-reliance on recall and near-transfer tasks, limiting their validity as meaningful measures of holistic proof comprehension (Weber, 2012; Mejia-Ramos et al., 2017; Cowen, 1991; Conradie and Frith, 2000). Yet, despite widespread agreement on the problem, robust solutions are few and far between.

In my research, I use *comparative judgment* to address these gaps in the literature. Comparative judgment relies on Thurstone's (1927) observation that humans are better at side-by-side comparisons than they are at evaluating an object against criteria. For a classical example, consider the estimation of weight. If given two objects of similar weight, most humans are more reliable when asked to identify the heavier object than they are when asked to identify the weight of a given object (Jones et al., 2019). In education, these objects are most often replaced by student-produced texts and the notion of weight is replaced by experts' perceptions of merit. Regarding proof, in particular, the application of comparative judgment intends to harness (and locate) any collective understanding of mathematicians about what proof is and how proof-related tasks should be performed. I propose this approach in contrast to the traditional measures where criteria for evaluation must be precisely defined and agreed upon, difficult in topics related to proof.

The research presented in this thesis focuses on two tasks, both assessed using comparative judgment. The *Conceptions Task* asks for an explanation of what mathematicians mean by proof in 40 words or fewer. The *Summary Task* asks for a summary of an elementary proof in either number theory or real analysis, also in 40 words or fewer. Responses to both tasks are evaluated by a cohort of judges, each of whom makes a series of pairwise comparisons used to generate a unique score for each script. The technical details of this modelling process are set out in Chapter 3.

## 1.1 Research aims

This research has two primary aims:

1) To evaluate the value of two comparative judgment-based tasks as measures of proof-related understanding. This is accomplished by exploring the reliability and validity of the scores produced by the comparative judgment of each task.

2) To use these tasks to learn about the conceptions and behaviours of mathematicians and students in proof-related contexts. This is accomplished via systematic analyses of responses from mathematicians and students, seeking patterns in the features of responses most valued by the various judging cohorts.

## 1.2 Outline of thesis

This thesis is presented in four parts. Part One, consisting of Chapters 2 and 3, establishes the theoretical and methodological foundations of my research. Part Two, Chapters 4 and 5, focuses on the Conceptions Task. Part Three, Chapters 6 through 9, focuses on the Summary Task. Each of the empirical chapters in Parts Two and Three presents data oriented toward different types of validity. Finally, Part Four provides summative remarks on each of the research questions set out in Chapter 2, and explores the implications of this research.

In Chapter 2, I review the literature on proof, proof comprehension assessment, and comparative judgment. I provide a brief history of mathematical proof, before extending the above discussion on the various conceptions of proof in the philosophical and educational literature. I then discuss traditional and recent approaches to proof comprehension assessment, before exploring the comparative judgment literature and its established applications, including differing content domains and research purposes.

Chapter 3 explores the methodological issues associated with my research. This begins with a discussion of mixed methods and an associated pragmatic philosophical approach. I then explore topics in assessment and research validity, and their implications for the data collection and analyses presented in the chapters that follow.

Chapter 4 begins Part Two, presenting responses to the Conceptions Task from undergraduate students and research-active mathematicians. This study is presented in four phases, with judges of different mathematical expertise at each phase. The reliability of the proof Conceptions Task is established via the standard quantitative measures for comparative judgment-based tasks. Validity is examined through a series of quantitative and qualitative analyses including comparisons with established measures, content analysis of the responses themselves, a regression analysis using content-based codes as predictors of task score, and by varying the mathematical expertise of the judging cohort.

Chapter 5 presents a longitudinal study in which students completed the

Conceptions Task at the beginning and end of an introduction to Proof module for undergraduates. Based on the assumption that students' conceptions should improve through their participation in the module, I infer validity evidence based on the capacity of the Conceptions Task to reflect this improvement. I, again, present a content analysis of the responses, facilitating comparison across contexts in conjunction with the previous chapter. This acts as a first step toward understanding the generalisability of the findings across the two chapters. This chapter also considers the implications of comparative judgment on the measurement of beliefs and conceptions and the generalisability of this work to other content domains within and beyond mathematics.

Chapter 6 begins Part Three and presents the first of four empirical studies associated with the Summary Task. This chapter is based on the *uncountability proof*, demonstrating the uncountability of the open unit interval using Cantor's diagonalisation argument. Here, I explore the reliability and validity of the comparative judgment-based assessment of students' summaries of this proof. Similar to Chapter 4, this chapter features a quantitative comparison between Summary Task scores and established proof comprehension measures, a content analysis of students' summaries, and a regression analysis using content-based codes as predictors of summary scores.

Chapter 7 presents a study focused on the *primes proof*, demonstrating the infinitude of the prime integers via contradiction. Analyses are similar to the previous chapter. However, the measures used for quantitative comparison are more numerous and diverse, providing a more detailed view of the criterion validity of the Summary Task scores. As before, this chapter also features content analyses for summaries of each proof and associated regression analysis using content-based codes as predictors.

Chapter 8 presents two studies focused on the *Fibonacci proof*, demonstrating the evenness of every third Fibonacci number via the principle of mathematical induction. The first is a small study, with data collected alongside those presented in Chapter 6. The second is a larger study, using data collected alongside the data presented in Chapter 7. As well as criterion and content validity analyses, this chapter also features a summative element, considering the implications of the work presented across all three proofs.

Chapter 9 presents interviews focused on judges' decision-making in evaluating students' summaries of the uncountability proof. Despite revisiting data from Chapter 6, this study is presented last to highlight the departure of this chapter from the format of those before it. In this chapter, I present a thematic analysis of interview transcripts, identifying themes influencing judges in choosing one proof summary over another.

Chapter 10 concludes this thesis by summarising the findings and considering the relationships between the empirical studies. This chapter also explores directions for future research and the implications of this research, both in practical classroom-based settings, and for the wider literature on proof and assessment.

# Chapter 2

# Literature review

In this chapter, I review two bodies of literature on *proof* and *comparative judgment*. I survey the literature on proof, beginning with a brief history of proof and its various conceptions in the philosophy, mathematics and education literatures. I then survey the work on proof comprehension and its assessment. Having established the need for alternative approaches to proof comprehension assessment, I turn attention to comparative judgment and its possible role in a solution. I explore the origins of comparative judgment, first as a psychometric tool, and more recently as an assessment tool in education. I discuss its various applications across education, before discussing the variety of approaches to evaluating reliability and validity seen in the literature.

## 2.1 A brief history of mathematical proof

Proof is commonly understood to be central to mathematics. Yet, despite its importance, there is little consensus on what proof is, or how it functions within the system of mathematical knowledge. A brief history illustrates the variety of proof conceptions in the literature and suggests that the range of conceptions within mathematics education is not only unsurprising but may be a necessary consequence of the historical and epistemological evolution of mathematics itself.

### The rise and fall of uncertainty

For the majority of the period from ancient Greece to approximately 1850, mathematics was built on Euclid's 'Elements' and was viewed as establishing self-evident truths associated with the laws of nature (Stedall, 2012). This belief in *a priori* mathematical truths lent itself to the belief that mathematics was the pinnacle of certainty – above empirical sciences – based on its unique reliance solely on logic and argumentation (Geist et al., 2010).

This Platonic idea of an *a priori* mathematics was first challenged in the mainstream consciousness by the discovery of non-Euclidean geometries early in the 19th century (Kline, 1980). For the first time, the mathematical world had multiple sets of axioms, leading to acceptable yet contradictory inferences in the sensible world. The discovery of non-Euclidean geometries led to a gradual loss of certainty in the objectivity of a singular *a priori* mathematics, prompting many mathematicians to re-examine their Platonic positions. As a result, it became unpopular to speak of mathematics as a descriptor of a natural truth (Marcus and McEvoy, 2016). This loss of certainty led to a closer examination of the foundations of mathematics that had gone largely unquestioned for centuries, resulting in the discovery that mathematics 'contained [many] false proofs, slips in reasoning, and inadvertent mistakes' (Kline, 1980, p. 5).

The latter half of the 19$^{\text{th}}$ century saw a rigorisation of mathematics, in which many mathematical flaws were discovered and fixed, and multiple new branches of mathematics began to flourish (Shapiro, 2000). It was not long, however, before several further contradictions emerged, demonstrating further epistemic issues in the new mathematics. Attempts to address these new problems fragmented the community. According to Shapiro (1997), at least three new branches of mathematical philosophy gained significant followings during the early 20$^{\text{th}}$ century: *formalism, intuitionism* and *logicism.* Each had its own epistemological assumptions about mathematics and the world around us and proponents of each school of thought attempted to prove the logical consistency of their version of mathematics.

### The influence of Gödel's incompleteness theorems

All these projects were interrupted when Gödel's incompleteness theorems, coupled with the paradoxes discovered in the preceding decades, 'showed both that the mathematical realm was stranger... than had been imagined, and raised questions regarding the relation between mathematical truth and mathematical proof' (Marcus and McEvoy, 2016, p. 245). Since then, the philosophy of mathematics has diversified further with the development of a host of new theories (see Marcus and McEvoy, *ibid.*, for a non-exhaustive list including fictionalism, modalism, naturalism, and experimentalism). Despite numerous attempts, none has addressed the relationship between truth and proof in a manner sufficient to dominate the philosophical landscape (Shapiro, 2000).

**The Jaffe-Quinn debate**

Following the controversial paper of Jaffe and Quinn (1993) on the epistemic role of informal proof in mathematics, there has been active recent discussion about mathematics' return to a pre-rigorous epistemology on proof. Jaffe and Quinn drew on the relationship between theoretical physics, in which informal speculation is common and fruitful, and mathematics, highlighting modern mathematicians' contrasting aversion to informal work. To this end, the authors proposed a framework for mathematical practice that they argue, with careful implementation, 'should give a positive context for speculation in mathematics' (*ibid*, p. 13).

This influential paper led to a body of recent work on the philosophy of mathematical practice, attempting to understand and develop a theory of mathematical epistemology within which one can accept informal proofs as epistemic contributors. Further discussion of the Jaffe-Quinn debate can be found in Atiyah (1994), and overviews of the wider philosophy of mathematical practice literature can be found in Mancosu (2008) and Larvor (2012). For this thesis, it suffices to conclude that there is active and recent debate on the epistemology of mathematics and the role of proof therein.

## 2.2 Conceptions of proof

Having established a historical context for discussion of proof and mathematics, I now turn to the specific conceptions of proof present in the literatures on philosophy and education. Borrowing language from Weber and Czocher (2019), I frame this discussion through two perspectives, casting various works as advancing either a *pluralistic* or *consensus* view of proof. I explore theoretical and empirical arguments for both epistemic positions, before discussing recent work attempting to bridge the gap between the two. Finally, this section concludes with a discussion of students' proof conceptions and attempts to evaluate the merits of such conceptions in education research.

### 2.2.1 The pluralistic view

> 'From antiquity onwards, mathematicians disagreed on how best to do mathematics, what methods to use for attacking and establishing results... There were in general no established criteria of what constitutes an acceptable proof, and perceptions of the role of proof varied considerably' (Kleiner and Movshovitz-Hadar, 1997, p. 16).

Consistent with its turbulent history, definitions of proof have been contested for as long as mathematicians have valued proof. Rav (2007) argued this is an inevitable product of the 'historical and methodological wealth of proof practices' (p. 299), and that attempts to condense mathematical proof to a 'unique and uniform' (p. 299) perspective were and remain fundamentally misguided. This pluralistic view of proof was also explicitly advanced by Czocher and Weber (in press), who noted the 'severe challenges' (p. 4) associated with the oft-seen approach of defining proof through a set of necessary and sufficient conditions. At least according to Weber (2014), for every plausible set of conditions there exists a counter-example that satisfies these conditions but 'is not a proof or is a proof but fails to satisfy these [conditions]' (p. 1). Pluralists often point to new or fringe cases in mathematics, such as computer-assisted or diagrammatic proofs, to support their arguments (Aberdein, 2009; Czocher and Weber, in press; Bundy et al., 2005). Aberdein, for example, introduced the term *proof** to deal with these fringe cases that he refers to as 'alleged proofs' (p. 1). These grey areas divide the mathematical community and are hence used to advance the idea that mathematicians hold differing conceptions of proof.

Proponents of this pluralistic viewpoint also point to empirical cases where mathematicians disagree on the validity of purported proofs. In an online study with 109 mathematicians, Inglis et al. (2013) found that mathematicians often disagreed on the validity of a purported proof demonstrating that $\int x^{-1}dx = \ln(x) + c$. The purported proof had several potential shortcomings or ambiguities, yet was deemed valid by 27% of the participants who either did not notice these potential issues or deemed them not substantial enough to invalidate the proof. This was an extension of a smaller eye-movement study (Inglis and Alcock, 2012) in which 12 mathematicians were shown six brief purported proofs of similarly elementary mathematical statements. In three cases, a non-trivial level of disagreement was found in mathematicians' appraisals.

In a selective review of the education literature on proof, Balacheff (2008) identified no fewer than five distinct conceptions of proof prominent in the literature, each with their own emphases and implications for research and practice, see Table 2.1. Balacheff's discussion was not intended as an exhaustive review, but rather to highlight the diversity in perspectives prominent in the community. Czocher and Weber (in press) provided a similar list of diverse definitions, each built on different epistemic assumptions.

*Table 2.1*

*Balacheff's (2008) summary of influential texts on proof.*

| Author | Conception of proof |
|---|---|
| Fawcett (1938) | 'The concept of proof is one for which the pupil should have a growing and increasing understanding. It is a concept which not only pervades his work in mathematics but is also involved in all situations where conclusions are to be reached and decisions to be made. Mathematics has a unique contribution to make in the development of this concept' (p. 120). |
| Harel and Sowder (1998) | 'A person's proof scheme consists of what constitutes ascertaining and persuading for that person ... As defined, ascertaining and persuading are entirely subjective and can vary from person to person, civilisation to civilisation, and generation to generation within the same civilisation' (p. 242). And finally: 'one's proof scheme is idiosyncratic and may vary from field to field, and even within mathematics itself' (p. 275). |
| Healy and Hoyles (1998) | 'Proof is the heart of mathematical thinking, and deductive reasoning, which underpins the process of proving, exemplifies the distinction between mathematics and the empirical sciences' (p. 1). |
| Hanna and Janke (1996) | 'The most significant potential contribution of proof to mathematics education is the communication of mathematical understanding. ... A mathematics curriculum which aims to reflect the real role of rigorous proof in mathematics must present it as an indispensable tool of mathematics rather than at the very core of that science' (p. 877-879). |
| Mariotti (1997) | 'A geometrical fact, a theorem ... is acceptable only because it is systematised within a theory, with complete autonomy from any verification or argumentation at an empirical level' (p. 22). |

At the level of the individual researcher, the pluralistic understanding promoted by Rav (2007) or Kleiner and Movshovitz-Hadar (1997) is neither remarkable nor clearly problematic. At the community level, the multiplicity of conceptions of proof is a source of concern, as it limits the ability of researchers to build on each other's work and the ability of the discipline to make meaningful advances in understanding (Balacheff, 2008). Like philosophers and mathematicians, education researchers notoriously disagree on the nature of mathematical proof. There is substantive overlap in the language used (e.g. 'proof, argumentation, justification, validation', Balacheff, 2008, p. 10). Yet, 'for each of them, we have in mind slightly different meanings when taking mathematics as a reference' (*ibid*).

Both Balacheff (2008) and Weber (2014) called for more thoughtful consideration of these discrepancies and the various definitions implicitly adopted by researchers. One approach to understanding the differences between various conceptions is to understand their scope, as well as the nature of the overlap between them. The next section discusses evidence for a view of proof based on consensus among experts. By examining the locus of the consensus view of proof, the pluralistic view can be better understood by elimination.

### 2.2.2 The consensus view of proof

In contrast to the literature discussed above, other researchers are more cautious about the extent of the disagreement. In this section, I discuss the claim that many education researchers implicitly adopt a consensus view without examination (Weber and Czocher, 2019), and the empirical evidence supporting a consensus-oriented position.

According to Weber and Czocher (2019), the consensus view of proof is implicit in substantive bodies of mathematics education research. As Inglis et al. (2013) did with expert mathematicians, it is common to evaluate proof comprehension by asking students to verify purported proofs (Alcock and Weber, 2005; Healy and Hoyles, 2000; Ko and Knuth, 2013, 2009; Weber, 2010), then scoring responses as correct or incorrect based on comparison with the accepted evaluation of the given proof. In the absence of a consensus amongst experts, these tasks arguably become nonsensical. One must assume that the designers of such studies believe that such a consensus exists, at least for the specific set of arguments they used. Similarly reliant on a consensus view of proof is the theoretical work of Dawkins and Weber (2017), who discussed the learning of mathematical proof as an enculturation process for students to come to understand and appreciate the values and norms of the mathematical community. Again, for this work to make sense, one must assume that at least on some level, there is a shared understanding of mathematics into which students can be enculturated.

Taking an empirical approach to the topic, Lai et al. (2012) and Miller et al. (2018) both found substantial agreement amongst mathematicians, at least in pedagogical contexts. Lai et al. examined mathematicians' perceptions of good pedagogical proofs, and found agreement amongst eight mathematicians on features of good pedagogical proofs for presentation to second- and third-year undergraduate students. These features included clear introductory and concluding sentences, explicit statements of the main idea and the avoidance of extraneous or redundant information. Although focused on the appraisal of students' proofs rather than those produced for students, Miller et al. (2018) also found consistency amongst 10 mathematicians in evaluating the correctness of students' proofs.

Weber and Czocher (2019) reported an online study with a focus similar to Lai et al. (2012) and Miller et al. (2018). Weber and Czocher asked 109 mathematicians to evaluate five proofs with varied characteristics. Of the five, two were described as prototypical textbook proofs, one was exclusively visual, one was computer-assisted and one was based exclusively on empirical evidence. Consistent with previous literature, Weber and Czocher (2019) found that mathematicians were divided on the visual and computer-assisted proofs, deemed valid by 39% and 62% of participants, respectively. For the other three proofs, however, Weber and Czocher found near-total agreement ($> 98\%$). The prototypical proofs were accepted and the empirical argument unanimously rejected. These findings straddle the line between pluralistic and consensus views of proof. To the pluralists, Weber and Czocher added evidence of disagreement amongst mathematicians, and empirically confirmed Aberdein's (2009) assertions regarding the divisiveness of computer-assisted and visual proofs. To the consensus view, Weber and Czocher showed that mathematicians may largely agree on the proofs that they 'typically encounter' (p. 12). When divisions do arise, they do so only in contexts the mathematicians found to be atypical, and when asked, mathematicians demonstrated awareness that these purported proofs were controversial.

This literature paints a picture of a centralised consensus with ambiguities at its extremities. From this perspective, the debate between pluralistic and consensus views can be framed as one of relative size of the domains generating (dis-)agreement and the epistemic importance one places on the disagreements when they arise. Empirical work on experts' conceptions of proof is still relatively sparse and further research is needed to clarify when and where mathematicians agree, as well as the specific nature of the disagreements where they exist. Chapters 4 and 5 present comparative judgment-based work with implications for the scope of the consensus view in undergraduate mathematics.

### 2.2.3 Proof as a cluster category

This section presents the definition of proof promoted by Czocher and Weber (in press)[1]. This definition responds to the concerns about theoretical variety (Balacheff, 2008; Weber et al., 2014b), and is used to theorise about the empirical work presented later.

Based on a theoretical discussion of the likely pluralistic nature of proof, the authors argued that a good definition of proof should accommodate the following empirical observations:

> - 'Mathematicians generally agree on what constitutes a proof, but there are particular kinds of justifications, such as computer-assisted proofs*, that lead to intense disagreement.
> - There are properties shared by many proofs. For instance, most proofs remove all doubts about the veracity of a theorem, employ *a priori* reasoning, and are sanctioned by one's community. Yet none of the properties is shared by every proof.
> - There appear to be some members of the proof category that are more prototypical than other members. For instance, verbal-symbolic proofs containing algebraic manipulation are regarded as more prototypical than a computer-assisted proof* or a visual proof*'.
>
> (Czocher and Weber, in press, p. 15)

These authors further argued that a classical account, defined by Lakoff (1987) as a collection of properties shared by members of the group and no other, cannot result in an adequate definition of proof. In lieu of a classical account, Czocher and Weber (in press) turn to Lakoff's cluster category, born out of Wittgenstein's (1953) notion of family resemblance. Wittgenstein's family resemblance rests on the premise that an object with more properties consistent with membership is more likely to belong than one with fewer. Importantly, this is a probabilistic (as opposed to deterministic) structure that denies the binary notion of belonging inherent in classical accounts. Lakoff's notion of a *cluster category* builds on Wittgenstein's idea, describing a cluster category by the collection of properties that an object can satisfy, 'counting toward membership of the given category' (Czocher and Weber, in press, p. 17). To this end, Czocher

[1] This idea was initially published in Weber (2014) using the language of cluster concepts. This work has evolved to now be phrased in terms of categories, as is presented in Czocher and Weber (in press).

and Weber (in press) nominated the following five properties contributing to what they labelled a *proof cluster definition*:

> - 'A proof is a convincing justification that will remove all doubt that a theorem is true for a knowledgeable mathematician.
> - A proof is a perspicuous justification that is comprehensible by a knowledgeable mathematician and provides the reader with an understanding of why a theorem is true.
> - A proof is an *a priori* justification that shows that a theorem is a logically necessary consequence (i.e., a deductive consequence) of axioms, assumptions, and/or previously established claims.
> - A proof is a transparent justification where any sufficiently knowledgeable mathematician can fill in every gap (or believes in principle that he or she can do so given sufficient time, motivation, and content knowledge), perhaps to the level of being a formal derivation.
> - A proof is a justification that has been sanctioned by the mathematical community'.
>
> (Czocher and Weber, in press, p. 20)

The authors noted that this list of five properties is highly speculative and likely to spark disagreement. While seemingly reasonable, the specifics of this account are less important for this thesis than its structure and consequences. This account adequately reflects the coexistence of both pluralistic and consensus conceptions of proof. Fundamentally, a cluster account is a pluralistic view, acknowledging and building on the multitude of context-dependent classical accounts of proof. Czocher and Weber's cluster account is also consistent with the empirical work supporting a consensus viewpoint and the notion that in typical/elementary cases, most mathematicians appear to agree most of the time. Moreover, unlike their problematic status for a classical account, the divisive, atypical notions on the periphery of the cluster category become the expected consequence of the probabilistic structure.

Adopting this definition has several consequences for the mathematics education literature. I discuss three of them here[2]. First, describing classroom

[2] Czocher and Weber (in press) also note implications for task design, experimental design and pedagogical practices.

practices through a variety of lenses becomes easier by allowing multiple positions to coexist. This can be achieved by explicating which properties of mathematical proof one is adopting/referring to in a given moment.

Second, this cluster account also responds to concerns raised by Balacheff about researchers' ability to build upon one another's work. In particular, Czocher and Weber's cluster account allows for greater communication between researchers who may view one account of a student's activity as incommensurate with their own. By acknowledging that perhaps the two researchers are operating in different corners of the same room, bridging the gap between them likely becomes easier.

Finally, I note a further implication pertinent to the comparative judgment aspects of this thesis. As is discussed later in this chapter (see Section 2.4), comparative judgment is particularly valuable in contexts for which clear criteria are hard or impossible to produce. In such cases, comparative judgment-based approaches assume a shared understanding of the phenomenon of interest amongst the judges to produce reliable estimates of the relative quality of the responses being evaluated. No agreed-upon criteria exist in the context of proof, making it a prime candidate for comparative judgment. In the absence of the cluster account of proof, it appears that any comparative judgment-based approach would have to assume a consensus view of proof. Otherwise, agreement between judges would likely be capturing only non-mathematical properties such as handwriting and linguistic form. Through a cluster account of proof, it becomes reasonable to rely on the collective expertise of the judges, even if variation in proof conceptions exists.

### 2.2.4 Students' conceptions of proof

Substantive omissions until now have been the study of the conceptions of proof held by students, and the various attempts researchers have made to capture and evaluate these conceptions. In this section, I first discuss the research documenting students' generally 'non-availing'[3] (Muis, 2004, p. 64) conceptions of mathematical proof, before exploring an argument tempering such conclusions.

#### Students' non-availing conceptions of proof

In an online study of 220 undergraduate students, Mejia-Ramos and Inglis (2011) found that students held two conflicting conceptions of proof. The first

[3] While not explicitly focused on proof, Muis argued for the use of this intentionally non-value laden language to characterise students' potentially unproductive beliefs. By explicitly labelling beliefs as good and bad, Muis claimed that researchers limit their scope to see merit where they did not expect it.

related to conviction and was generally prompted by tasks containing the verb form 'to prove'. The second related to validity, activated by the noun form 'proof'. That these competing conceptions of proof can be activated by subtle differences in the semantic phrasing of a task suggests that these students did not hold particularly robust conceptions of proof and that these competing conceptions are likely a substantive barrier to progress for these individuals.

In a slightly different vein, Weber et al. (2014b) focused on mathematics students' beliefs about proof reading. The authors found that students expected to be able to understand a good proof in less than 15 minutes, that they would not be expected to independently produce justifications or diagrams and that understanding a proof was equivalent to understanding each isolated step.

It is also common to explore students' conceptions of proof via the proof schemes framework of Harel and Sowder (1998). This framework has been used numerous times (e.g. Kanellos et al., 2018; Segal, 1999; Recio and Godino, 2001) to describe and categorise the types of arguments students make and the types of arguments they find convincing. In such research students often demonstrated an over-reliance on empirical evidence and appeals to authority. Researchers have consistently reported a discrepancy between the proof schemes employed by research-active mathematicians and those employed by students. For example, in a large-scale study of undergraduate students, Recio and Godino found that more than 40% of 400 first-year university students resorted to producing empirical arguments when asked to prove elementary statements from algebra and geometry. The authors inferred that students likely hold empirical proof schemes and find empirical arguments convincing.

### A word of caution on inference from written responses

Weber (2010) was critical of the conclusions above, noting that students are influenced by a variety of non-epistemological factors. Weber observed that Recio and Godino's written artefacts are products of, amongst other things, a social context in which students likely experience pressure to produce an answer, either to receive partial credit or to please the researcher. In such cases, the implicit assumption that students are convinced by their own work, or that they find empirical arguments convincing is misguided. Weber presented empirical evidence supporting this view. In a smaller study with 28 undergraduate mathematicians, students were asked to evaluate the conviction they gained in the truth of 10 statements based on certain pre-selected arguments, Weber concluded that his students were not convinced by empirical arguments. He also found, however, that students were convinced by diagrammatic arguments, and that they often accepted invalid arguments without recognising their logical flaws.

### Evaluating students' proof conceptions

Of those interested explicitly in students' conceptions of proof, only Stylianou et al. (2015) and Healy and Hoyles (2000) quantitatively examined the relative merit of students' responses. Stylianou et al. used a multiple-choice test of proof conceptions comprising five pre-judged items. To understand the role of proof conceptions in learning, the authors compared their test of proof conceptions to self-efficacy beliefs and attainment on more standard proof comprehension tasks. They found that the quality of students' proof conceptions, as captured by their test, were strong predictors of mathematics attainment but were unrelated to their self-efficacy beliefs or self-reported experience of themselves as learners of proof. These findings were consistent with Healy and Hoyles, who implemented a suite of proof-related tasks with middle school students, attempting to understand the conceptions of proof they held. Healy and Hoyles implemented two quantitative surveys, designed to elicit students' conceptions of proof by asking them to evaluate given arguments, to select arguments most similar to those they would expect to produce, and to provide a written description of proof and its purpose. The authors categorised students' descriptions according to the purposes of proof set out by de Villiers (1990) and used these data as predictors of performance on other proof-related tasks. Healy and Hoyles imposed no direct value-judgment on the relative quality of each category and found no relationship between their coding and other measures of mathematics attainment or beliefs.

The approaches of both Stylianou et al. (2015) and Healy and Hoyles (2000) are useful for understanding how students view the world of mathematics through pre-determined criteria. Both are limited, however, by the granularity of their data and their reliance on pre-set categories for student responses. Stylianou et al. (2015) used a closed-instrument that lends itself to immediate systematic analysis, but does not capture the valuable diversity achieved by other methods. On the other hand, Healy and Hoyles (2000) allowed for a wide diversity of responses by setting at least one open-ended task, but lacked a fined-grained tool to systematically analyse this aspect of their data. As is discussed later in this chapter, comparative judgment achieves the best of both worlds, facilitating open-ended tasks with a quantitative measure of response quality. The reliability and validity of such a measure are open questions that are considered empirically and theoretically in Chapters 4 and 5.

## 2.3 Proof comprehension

Having discussed the proof conceptions of philosophers, mathematicians and students, I now turn to the education-focused literature on students' comprehension of particular proofs. I first discuss *traditional mathematics assessment*, as labelled by Weber (2015) and Iannone and Simpson (2011). I then discuss the metrics of proof comprehension apparent in the recent education literature before turning attention to a line of research explicitly addressing the problem of systematically generating reliable and valid measures of proof comprehension. This line of research is given particularly careful attention for its pioneering status in the field and its importance to the empirical work in later chapters.

In taking an assessment-oriented view of the literature, I exclude several tangentially related bodies of work on proof within mathematics education. In particular, I discuss little of the recent work developing the teaching and learning of mathematical proof, attending only to particular works with interesting contributions to evaluating progress and attainment in proof comprehension. For wider reviews of the field, see Reid and Knipping (2010) and Stylianides et al. (2017).

### 2.3.1 Traditional assessment

I infer a definition of *traditional assessment*, from the literature on *traditional instruction*. In a study of mathematicians' pedagogical practices, Weber (2004) referred to traditional instruction as following a 'definition-theorem-proof' (language from Davis and Hersh, 1981) mode of teaching, with the main goal of having students become 'capable of producing rigorous proofs about the covered mathematical concepts' (p. 116). From this, I infer that *traditional assessment* is that which measures students' ability to produce rigorous proofs. This is consistent with Iannone and Simpson (2011) who characterised closed-book examinations as traditional assessments of mathematics[4].

This traditional view, while consistent with the wider assessment literature, may turn out to be an unfair characterisation of the care and attention mathematicians' give to their pedagogical practice. Seeking a more nuanced view of mathematicians' behaviour, Weber (2012) interviewed nine mathematicians,

[4] According to Iannone and Simpson (2011), until as recently as the 18th century, mathematics had a long-standing tradition of oral assessment. It was only with the increased focused on individual cognition, coupled with alleged corruption in the oral assessment system and the rise of Newtonian mechanics, that written assessments took over. As recently as 2011, this status quo remained largely undisturbed, with all 11 UK universities from a representative survey using closed-book examinations as their primary mode of assessment. Within institutions, percentages ranged from 43-92% with a mean of 72% (ibid).

asking how they assess their students, and in particular, their understanding of proof. Five mathematicians said they asked near-transfer tasks in which students were required to prove novel statements similar to those presented in class (e.g. asking students to prove $\sqrt{3} \in \mathbb{R} \setminus \mathbb{Q}$, having shown the same for $\sqrt{2}$ during the course), two said they asked students to reproduce proofs from the course and two said that they did not assess proof at all. Importantly, those who indicated that they assess proof noted the inadequacy, or shallow nature, of their assessments. Yet, consistent with the claims of Iannone and Simpson (2011), this mode of assessment continues to dominate current practice, leaving a status quo wherein mathematicians knowingly administer inadequate assessments.

In recent decades, the mathematics education community has attempted to address this problem in two ways. One adopts a more rigorous consideration of reliability and validity. The other calls for increased diversity in 'the assessment diet' (Iannone and Simpson, 2011) of mathematics students, implicitly accepting the imperfections in the various approaches and arguing that by diversifying the assessment structure, we naturally attain a more holistic view of students' understanding. The majority of the literature in this latter direction addresses general mathematics assessment, rather than proof comprehension assessment itself, and is omitted here. Detailed reviews of this literature can be found in Iannone and Simpson (2015) and Iannone and Simpson (2011). Here, I focus on research explicitly addressing proof comprehension.

### 2.3.2 Recent approaches to proof comprehension

In the previous section, I established that mathematicians and mathematics educators are dissatisfied with the traditional approach to mathematics assessment and agree that, at best, these measures provide only a superficial understanding of students' understanding (Mejia-Ramos et al., 2017). In attempting to generate a deeper understanding of students' proof comprehension, I identify three strands of literature. The first is based on proof construction tasks. The second, and more recent, is on proof reading. The third strand of literature focuses on generating models of students' proof comprehension and the use of these models in generating reliable assessment. The first two strands are explored in this section, the third is reserved for the more in-depth discussion that follows.

#### Proof construction

Early education research on proof comprehension focused on the use of qualitative methods to evaluate students' proof constructions. For the most part, research on construction tasks embraced the status quo of traditional assessment, using largely qualitative methods to gain deeper insights into students' understanding than is offered by traditional evaluations. I summarise the findings from this body of work in terms of three barriers to success: 1) lack of content knowledge (Moore, 1994; Ko and Knuth, 2009), 2) lack of strategic knowledge (Weber, 2001; Hoyles and Healy, 2007), and 3) over-reliance on inappropriate argument forms (Harel and Sowder, 1998, 2007; Küchemann and Hoyles, 2006).

Moore (1994) identified insufficient content knowledge as a significant barrier to students' proof constructions in an upper-undergraduate course on logic and set theory. In a study of 16 students, Moore found seven sources of difficulty in constructing proofs, four of which I interpret as demonstrations of insufficient content knowledge. Moore observed that students did not know the definitions, had little intuitive understanding of associated concepts, were unable to generate examples, and had insufficient concept images (in the sense of Tall and Vinner, 1981) of related concepts. Similarly, in their interview-based study of undergraduate students, Ko and Knuth (2009) reported that students lacked content knowledge. In particular, their inability to recall key definitions limited their capacity to produce proofs and counter-examples to given theorems.

Beyond content knowledge, Weber (2001) found that students lacked strategic knowledge. From a study of four undergraduates and four doctoral students, Weber found that undergraduates lacked sufficient strategic knowledge to know where their entirely adequate content knowledge should be applied. Hoyles and Healy (2007) reported a similar finding with secondary school students who struggled to identify an appropriate starting point when constructing a proof, limiting their ability to implement their relevant content knowledge.

Finally, the third strand of research identified over-reliance on inappropriate argument forms as a barrier to students' proof constructions. Harel and Sowder (1998, 2007) and Küchemann and Hoyles (2006) reported an over-reliance on empirical data and concrete examples, limiting students' ability to construct arguments recognised by the mathematical community as proofs. Using qualitative methods on a limited number of specific tasks at the lower tertiary and upper secondary school levels, all three papers found that students did not see the limitations of empirical approaches, and accordingly failed to see the benefits of potentially more sophisticated approaches.

While shedding new light on students' understanding of proof, and greater specificity about students' difficulties, this line of research takes us no closer to more effective modes of assessment suitable for practitioners. In search of further insights and classroom-appropriate assessments, researchers have begun turning to reading tasks for solutions.

### Reading purported proofs

More recently, research has incorporated more work on students' reading of given proofs and, in particular, their ability to appropriately evaluate the validity of purported proofs. There is a strong consensus that many students do not read proofs effectively and do not appropriately evaluate purported proofs. The research in this area often involves multi-methods designs in which small groups of students are given validation tasks generating both qualitative and quantitative data. As before, the findings from this literature are discussed in three classes: 1) inappropriate focus on surface features (Selden and Selden, 2003; Ko and Knuth, 2009, 2013; Weber, 2012), 2) insufficient attention to localised detail (Weber, 2010; Ko and Knuth, 2013), and 3) failure to apply new information to closely related problems (Shepherd et al., 2012).

Both Selden and Selden (2003) and Alcock and Weber (2005) observed that some students did not pay attention to a purported proof's global purpose or structure, focusing instead on verifying calculations or specific implications. When reading proofs, many students focused too much on surface-level features, failing to pay sufficient attention to links between statements and global argument structure. Investigating students' ability to evaluate the validity of given proofs, Selden and Selden presented eight undergraduate mathematics students with arguments for four theorems generated by their peers. The students had limited success in evaluating the validity of these arguments, reportedly as a result of an inappropriate focus on surface features of the arguments. Students' surface-level focus was also reported by Ko and Knuth (2009) and Ko and Knuth (2013), who studied 16 mathematics undergraduates and identified an inability to recognise 'global-structure' as a significant barrier to successful evaluation of given arguments. These largely qualitative findings were corroborated with the use of eye-tracking technology by Inglis and Alcock (2012). By comparing the eye-movements of 12 mathematicians and 18 first-year undergraduates, Inglis and Alcock found that the undergraduates were less inclined to switch their attention back and forth between lines of an argument, suggesting they spend less time identifying implicit links between statements and/or attending to global structure.

Weber (2010) and Ko and Knuth (2013) both found that mathematics majors did not pay sufficient attention to detail and did not adequately identify localised logical flaws. Weber, in a study with 28 mathematics majors in a ‘transition-to-proof’ course, found students did not identify logical gaps in deductive arguments. Similarly, the mixed methods study by Ko and Knuth involving 16 undergraduate mathematics students supported the claim that students struggle to identify what the researchers termed localised logical gaps in deductive arguments. These largely interview-based findings were corroborated using eye-movement analyses by Inglis and Alcock (2012). Inglis and Alcock found that undergraduates were less inclined than mathematicians to switch their attention back and forth between lines in a purported proof, suggesting that students were less attentive to the global structure of proofs or the implicit links between logical statements. All these findings suggest that students might not fully comprehend proofs they are expected to read. In a similar vein, Hodds et al. (2014) reported that self-explanation training, in which participants were encouraged to identify and elaborate on the main ideas of the proof before developing their own explanations, significantly increased students’ success rate in appropriately evaluating purported proofs. This is consistent with the claim that students do not evaluate proofs appropriately because of their insufficient attention to detail.

Finally, Shepherd et al. (2012) focused on high-achieving mathematics students and their ability to transfer new knowledge to an immediately related context. Even high achieving students had significant difficulties with such tasks, despite their good general reading ability (as measured by the Constructively Responsive Reading metric). Supporting several findings discussed earlier, Shepherd et al. identified insensitivity to localised errors, insufficient content knowledge, and insufficient attention to detail as barriers to participants’ success.

The literature presented in this section highlights barriers to students’ success in coming to understand proof and the various methods researchers have used to evaluate students’ understanding. This work represents meaningful progress in proof comprehension assessment. In particular, unlike what is learnt from traditional assessment of proof, this body of work has advanced understanding of students’ difficulties with proof and their specific shortcomings in both constructing and reading proofs.

Thus far in the discussion, there has been little consideration of the reliability and validity of the conclusions. The majority of the research discussed is based on studies with fewer than 30 participants and investigations of reliability and validity were based primarily on qualitative methods. Further, while this work has advanced understanding of students’ difficulties, it has done little to advance the design of future assessments for use in the classroom. The strand of literature discussed in the following section addresses this.

### 2.3.3 Generating a model for proof comprehension assessment

In this section, I discuss a line of research that has resulted in three reportedly reliable and valid multiple-choice tests, rigorously designed to assess students' reading comprehension of three proofs in number theory and real analysis. These tests result from an extensive mixed methods design and are based on the seven-aspect model of proof comprehension assessment generated in the process (Mejia-Ramos et al., 2012, 2017). I first discuss the origins and specifics of this model, before discussing the multiple-choice tests and their utility for future research, including my own. I identify the line of research initiated by Conradie and Frith (2000) and later extended by Mejia-Ramos et al. (2012) as the only comprehensive effort to understand assessment of proof comprehension. Mejia-Ramos et al. drew on both Conradie and Frith, and Yang and Lin (2008) to provide an assessment model for proof comprehension that is an important source for much of the research presented later.

#### A seven-aspect model for proof comprehension

The model described below was the product of an extensive two-part literature review and a series of systematic interviews with mathematicians. Mejia-Ramos et al. (2012) first considered the functions and purposes of proof, then reviewed the recommended alternative methods of presenting proof.

The resulting seven-part model comprised two parts, the *local* aspects and the *holistic* aspects, see Table 2.2. Local aspects of proof comprehension are those for which the reader is required to consider particular mathematical statements either in isolation, or in relation to a small number of other statements situated nearby. Holistic aspects are those which can only be understood by considering the proof globally, and 'cannot be gleaned by examining a small number of statements' (p. 6).

All three local aspects are adaptations from the framework in Yang and Lin (2008), developed in the context of high school geometry. Yang and Lin included one final aspect referred to as encapsulation, described as the stage at which students understand the generality and applicability of the proof at hand. Mejia-Ramos et al. (2012) deemed this insufficient for the more general setting of undergraduate mathematics, differentiating between four holistic aspects of proof comprehension at this level.

*Table 2.2*

*Mejia-Ramos et al.'s (2012) proof comprehension assessment model (p. 15).*

| | Aspect | Assessment evidence |
|---|---|---|
| | *Local* | |
| 1. | Meaning of terms and statements | Understanding of key terms and statements in the proof. |
| 2. | Logical status of statements and proof framework | Knowledge of the logical status of statements in the proof and the logical relationship between these statements and the statement being proven. |
| 3. | Justification of claims | Comprehension of how each assertion in the proof follows from previous statements in the proof and other proven or assumed statements. |
| | *Holistic* | |
| 4. | Summarising via high-level ideas | Grasp of the main idea of the proof and its overarching approach. |
| 5. | Identifying the modular structure | Comprehension of the proof in terms of its main components/modules and the logical relationship between them. |
| 6. | Transferring the general ideas or methods to another context | Ability to adapt the ideas and procedures of the proof to solve other proving tasks. |
| 7. | Illustrating with examples | Understanding of the proof in terms of its relationship to specific examples. |

### Three multiple-choice comprehension tests

One product of this model was a series of three 12-question multiple-choice tests targeting comprehension of specific proofs. The design process was outlined in Mejia-Ramos et al. (2017). Similar to the assessment model itself these tests were designed via a resource-intensive mixed methods approach, comprising a series of validation phases including interviews with students and mathematicians and large-scale quantitative implementations (Mejia-Ramos et al., 2017). The final 12-question tests were cut down from original sets of 20 questions, some of which were found to be redundant. Each test attained a Cronbach's $\alpha > .7$ in the final large-scale trial, indicating acceptable internal consistency.

To my knowledge, these tests are the first rigorously developed measures of proof comprehension. Despite the numerous advantages of this approach, the resources required in developing tests for the variety of contexts required by practitioners limit its potential scope. As such, my research uses these tests as benchmarks against which to evaluate the merits of a new, perhaps more generalisable, approach to proof comprehension assessment. Before moving on to the comparative judgment literature, I discuss the notion of dimensionality in proof comprehension.

## A note on dimensionality

Thus far, I have used the phrase 'proof comprehension' as an implicitly unidimensional entity. Consistent with much of the literature on proof, I have permitted language regarding 'students' understanding of proof', implicitly accepting that such a notion can at least theoretically be considered a singular construct. This assumption of unidimensionality is not trivial and is worthy of further examination.

One alternative to this potentially problematic assumption is to distinguish between types of assessment, or between particular tasks. For example, it seems theoretically possible to argue that there is nothing in the literature fundamentally binding students' ability to construct proofs to their ability to read and/or evaluate them. There are two problems with this line of reasoning. First, even if one accepts such a distinction, the dimensionality of these new, smaller entities (proof construction and proof reading) remains unclear. Several possible factors influence an individual's ability to produce acceptable proofs and there is no evidence to suggest proficiency with one necessitates proficiency with another. The second problem comes from the language common in the literature. Regardless of the specific task- or content-specific context, researchers continue to use the same overlapping language of proof comprehension, indicating that at least implicitly, educationalists believe they are investigating a unidimensional entity.

The three multiple-choice tests of Mejia-Ramos et al. (2017), developed from the seven-aspect model of comprehension, provide an interesting testing ground for such thinking. Mejia-Ramos and Weber (2016) presented preliminary findings in a study with fewer than 150 students taking all three tests, reporting strong correlations between any two of the tests. The authors suggest that their results support a unidimensional view of proof comprehension, but are careful to note that further work is needed before making definitive claims to this end.

Although preliminary, these empirical results present an interesting theoretical conundrum in light of the seven-aspect model presented by Mejia-Ramos et al. (2012). On the one hand, these findings support the until now implicitly assumed unidimensionality of proof comprehension and serve to allay concerns about possible oversights in the literature. On the other hand, this quantitative indication of unidimensionality suggests that each aspect of the assessment model evaluates the same thing. This leads to the seemingly illogical conclusion that a task asking students to, for example, recall a definition from the proof (aspect 1 from the Mejia-Ramos et al.'s model) is somehow equivalent to one asking students to summarise the high-level ideas (aspect 4). This challenges seemingly reasonable intuitions about the nature of proof comprehension tasks and the necessity for a certain level of complexity to meaningfully evaluate students' comprehension. Interpreting simple quantitative evidence in education must be done with caution, but at the very least these findings provide motivation for future research.

In this thesis, I introduce a comparative judgment-based approach to proof comprehension, generating a new tool for assessing students' comprehension of a possibly unlimited number of proofs. Where each comprehension test of Mejia-Ramos et al. (2017) requires a rigorous design stage, my comparative judgment-based approach could, in principle, be applied to any proof. In the presence of sufficient evidence regarding the reliability and validity of the resulting scores, this has substantive potential to contribute to research on the dimensionality of proof comprehension by allowing researchers to consider and compare a wide variety of mathematical proofs. I return to the potential of comparative judgment in the realm of proof and proof comprehension at the end of the following section on the applications and purposes of comparative judgment in educational settings.

## 2.4 Comparative judgment

Recall that comparative judgment is a method for quantifying subjective psychological experiences. After collecting responses to what is often a short, open-ended task, judges are recruited to perform a series of pairwise comparisons, selecting the 'better' response from each pair. A statistical model is then used to generate a score for each response, understood to be an estimate of the quality of each response.

First introduced by Thurstone (1927), comparative judgment methods are based on the observation that humans are better at pairwise comparison tasks than they are at evaluation tasks based on explicit criteria. In the realm of education, Thurstone proposed applications to qualities like handwriting quality and children's drawings, but never published empirical work on these topics. Before Thurstone's work, many psychologists had steered away from the measurement of such subjective phenomena (Bramley, 2007), lacking the mathematical tools to model these high-variance phenomena that cannot be directly observed.

The technical details of comparative judgment are explained in Section 3.5. In this section, I discuss the education-focused literature on comparative judgment, beginning with a survey of the content-driven applications of comparative judgment in educational settings. While many studies focus exclusively on validating a particular comparative judgment-based assessment, there is substantive diversity in the purposes of the published research. These purposes are discussed next, before consideration of the various aspects of validity examined in the literature. This method-focused presentation of the literature reflects the aims of this thesis and lays the foundations for understanding the empirical work presented in later chapters. This section concludes by returning to the literature on proof and cluster concepts, justifying the use of comparative judgment in this context.

### 2.4.1 Applications of comparative judgment

The first application of comparative judgment in education is widely understood to be Pollitt and Murray (1993), who investigated spoken language proficiency using video-recorded excerpts judged by linguistics experts. By requiring judges to justify each decision aloud, the authors inferred an understanding of the validity by identifying themes in the judges' motivations.

In the intervening years, comparative judgment has been applied to a wide variety of content domains in languages (Heldsinger and Humphry, 2010, 2013), design and technology (Seery et al., 2012; Bartholomew et al., 2019; Kimbell, 2012), chemistry (McMahon and Jones, 2015), visual art portfolios (Newhouse, 2014), history (Holmes et al., 2018), engineering (Williams, 2012) and mathematics. Within mathematics, researchers have focused on problem solving (Jones et al., 2015; Jones and Inglis, 2015), conceptual understanding of algebra (Jones et al., 2019; Bisson et al., 2016), $p$-values (Bisson et al., 2016), calculus (Bisson et al., 2016; Jones and Alcock, 2014) and general GCSE mathematics[5] (Jones and Inglis, 2015). Others focused more generally on mathematical thinking (Hunter and Jones, 2018) or conceptual explanations (Jones and Karadeniz, 2016).

In all cases, researchers deemed their comparative judgment-based scores to be reliable and valid. Given the scarcity of universally successful quantitative measures within education, the ubiquity of the successful findings across the

[5] General Certificate of Secondary Education (GCSE) qualifications are subject-specific examinations taken by UK-based secondary students, typically aged 16.

literature raises a question of a 'file-draw problem' within this relatively young literature. This is not explored here, but the absence of 'boundary cases' is worth considering when reading the wider comparative judgment literature. I return to this topic in the final chapter.

### 2.4.2 Purposes of comparative judgment

There are numerous motivations for comparative judgment-based research, each aiming to capitalise on differing strengths of comparative judgment. I identify four primary focuses of the published work on comparative judgment: 1) validating new assessments, 2) evaluating and improving existing assessments, 3) developing non-comparative judgment-related assessments, and 4) understanding the behaviour of participants.

The first focus, validation of new assessments, has attracted the most research and is perhaps the most obvious application. Research of this ilk often argues that traditional assessments are inadequate, either for their lack of validity as authentic assessments of the target domain (Bisson et al., 2016), or for the constraints on task designers of traditional criteria-based assessments (Jones and Inglis, 2015). Comparative judgment research in education has typically focused on areas in which traditional, criteria-based assessment can be said to have fallen short, attempting to validate a comparative judgment-based assessment to fill the gap. Jones et al. (2015) and Jones and Inglis (2015) focused on validating a comparative judgment-based assessment of secondary school students' mathematical problem-solving. Bisson et al. (2016), Hunter and Jones (2018), Jones and Karadeniz (2016) and Jones et al. (2019) did similarly for tasks of the form 'Explain concept X' (Jones et al., 2019, p. 672).

Others have focused on evaluating and improving existing assessments in areas with existing tools (Jones et al., 2015; McMahon and Jones, 2015; Benton et al., 2018; Jones et al., 2016). Alongside their focus on problem-solving, Jones et al. used comparative judgment to evaluate traditional standard assessments of secondary school mathematics, finding their comparative judgment-based evaluation of responses was more time-efficient than traditional marking procedures, without sacrificing reliability or validity. Similarly, McMahon and Jones reported on one teacher's journey in implementing comparative judgment across an array of internal assessments of secondary school chemistry. They reported that comparative judgment was as reliable, more efficient and arguably generated better (more valid) assessment outcomes than the traditional assessments. Jones et al. also used comparative judgment to evaluate existing assessments, considering changing standards in UK-wide secondary school mathematics examinations.

The third focus of the comparative judgment-based literature has been the development of non-comparative judgment-based assessments. For example, Heldsinger and Humphry (2013) and Heldsinger and Humphry (2010) used comparative judgment to calibrate writing samples to be used as benchmarks against which to evaluate other samples in the absence of a comparative judgment-based process. Both Bramley (2007) and Thurstone (1927) also noted the potential of this application although its use has not been widespread.

The final focus of the comparative judgment literature has been on understanding the behaviour of participants, both via their responses to comparative judgment-based tasks and the decisions of judges. Hunter and Jones (2018), for example, used a comparative judgment-based approach to examine primary students' mathematical thinking. While the task design and methodology of their study were akin to those interested in evaluating validity, Hunter and Jones assumed a valid output, and used the resulting analysis to develop insights into the behaviour of both students and judges. Bartholomew et al. (2019) adopted a similar approach in developing an understanding of design values across cultures. In both studies, by considering the contents of written responses, researchers were able to identify the priorities or values of the judging cohort by identifying characteristics that corresponded with high comparative judgment-based scores.

In this thesis, I am motivated by the first and last elements of this list. That is, the two tasks featured in my empirical work generate comparative judgment-based evaluations for topics where I have argued that traditional assessment has failed. I consider the reliability and validity of these tasks. In each case, I then invert the attention to the participants, generating understanding of the students and mathematicians involved through the comparative judgment data.

### 2.4.3 Evaluating validity

Given its pivotal place in this thesis, I now focus on the literature about the validity of comparative judgment. In the sections that follow, I outline the several approaches to validity adopted by different researchers, before outlining the theoretical potential of comparative judgment in the realm of proof and proof comprehension. I discuss three common approaches to generating validity evidence in the literature on comparative judgment: *expert testimony, content analysis*, and *comparative analysis.* A detailed discussion of validity theory and its evolution in education research is presented in the following chapter on methodologies.

### Expert testimony

It is common to investigate the validity of comparative judgment scores using expert testimony (Jones et al., 2015; Jones and Inglis, 2015; Hunter and Jones, 2018; Davies et al., 2012; Pollitt and Murray, 1993). This testimony usually serves dual purposes within the research design. Most immediately, researchers compare expert testimony to theoretically expected ideas. For example, in their investigation of primary students' portfolio-based tasks in science and technology, Davies et al. drew positive conclusions regarding the validity of their task based on the theoretically appropriate priorities expressed by the judges. Jones and Inglis also relied primarily on expert testimony collected via two closed-answer questionnaires in evaluating a comparative judgment-based secondary school assessment designed as an alternative to a standard British GCSE assessment. The first questionnaire asked teachers how well their alternative assessment addressed primary curriculum features. The second asked teacher judges which elements of students' responses most influenced their decision-making. In a less structured manner, Jones et al. (2019) considered validity through discussions with the 'project advisory panel' (p. 674) of the task design and responses from a pilot study.

### Content analysis

It is also common to pair expert testimony with qualitative analyses of the responses being judged (Jones et al., 2015; Jones and Inglis, 2015; Hunter and Jones, 2018). In these cases, researchers have evaluated validity based on comparisons between judges' testimony and a coded content analysis of the task responses.

In investigating primary students' free-response explanations of mathematical concepts, Hunter and Jones (2018) conducted a content analysis for a sample of six students' responses. They reported symmetry between the comparative judgment-based scores, expert testimony from interviews with teachers, and the qualitative features of the sampled responses. On this basis, they concluded that their comparative judgment-based assessment had demonstrated validity in this case.

Content analyses have also been used in the absence of expert testimony, instead directly comparing content analysis with comparative judgment-based scores using statistical modelling. For example, Jones and Karadeniz (2016) investigated secondary students' conceptual understanding with a series of open-ended questions evaluated using comparative judgment. In evaluating the validity of their measure, they conducted a qualitative analysis of students' responses, coding them for five important traits predetermined from the literature. After conducting a multiple linear regression predicting comparative judgment scores using these five codes (as well as file size and performance on a standard test on fractions), Jones and Karadeniz concluded that their comparative judgment-based evaluation had demonstrated acceptable validity.

### Comparative analysis

The third approach to validity comes from quantitative comparison with theoretically similar measures. Bisson et al. (2016) considered the validity of their comparative judgment-based assessments of students' conceptual understanding by comparing their new measure with outputs from existing validated measures of theoretically similar entities. For their investigation of students' understanding of $p$-values, they benchmarked comparative judgment scores against performance on the RPASS-7 test (Lane-Getaz, 2013). When investigating students' comprehension of the derivative, they used the Calculus Concept Inventory (Epstein, 2013) as the benchmark. It is also common to benchmark comparative judgment-based assessments against standard measures of attainment for the population from which participants are recruited. For example, Jones and Alcock (2014) evaluated their assessment of conceptual understanding in introductory real analysis using the summative assessment scores attained at the end of the module on which students were enrolled. Jones et al. (2015) did similarly with scores in evaluating the validity of their secondary school assessment of mathematical problem-solving.

In a similar vein, Jones et al. (2019) performed a randomised control trial using a comparative judgment measure of students' algebra performance alongside a suite of standardised measures of procedural understanding, conceptual understanding and general achievement, as well as writing skills and mathematics anxiety. Similar to Bisson et al., they considered the correlation between comparative judgment-based scores and their standardised measure of algebra performance. Jones et al. also considered the capacity of their comparative judgment-based scores to detect the effect of their RCT intervention (known to exist through their algebra measure), expecting to find divergence between their control and intervention groups.

Jones et al. (2019) further pursued this notion of divergence by comparing comparative judgment-based scores of algebra performance with writing skills, finding a moderate correlation. This was explained as a function of the primary school setting in which students' ability to produce coherent sentences was likely related to their ability to respond to the comparative judgment prompt. Nevertheless, the authors drew no explicit conclusions about validity from this evidence. With a similar method, Jones and Karadeniz (2016) found that comparative judgment scores correlated more strongly with general mathematics achievement than with reading achievement, and hence that they had found further evidence for the validity of their measure of conceptual understanding.

Jones and Alcock (2014) also considered divergence as a measure of validity, albeit via a different mechanism. By judging their introductory analysis assessment with expert, external non-expert and peer judges, they evaluated the extent to which the scores produced were based on inherently mathematical features. Upon finding a significant difference between models produced by the judging cohorts, these authors concluded their data demonstrated validity as an assessment of mathematics, rather than non-mathematical features upon which the non-experts judges were assumed to have focused.

Finally, I consider the predictive capacity of a measure as an indicator of validity. Benton et al. (2018) evaluated students' writing samples using adaptive comparative judgment via comparison with later scores of language proficiency. While not common in the comparative judgment literature, I return to this notion of predictive validity in the empirical component of this thesis (Chapters 5 and 8).

### 2.4.4 Comparative judgment and proof comprehension

As discussed earlier, comparative judgment is most suited to areas of mathematics in which criteria are difficult to generate. In these cases, comparative judgment offers a flexible alternative for quantifying understanding in the absence of better options.

The mathematics education research community has developed a short list of robust, closed-form measures for conceptual understanding of particular mathematical concepts. These include the RPASS-7 test of introductory statistics (Lane-Getaz, 2013), the Calculus Concept Inventory (Epstein, 2013), and the multiple-choice tests of proof comprehension developed by Mejia-Ramos et al. (2017). By comparing comparative judgment-based outputs with such established measures, we gain insight into the scope of comparative judgment's domain-specific applications, potentially circumventing the need to replicate the resource- and time-intensive processes required to rigorously generate and validate such conceptual measures.

Given the diversity of proof conceptions, and the absence of an agreed-upon definition, proof comprehension is a prime candidate for such an approach. As discussed in Section 2.2.3, this approach to assessment is consistent with the notion of proof as a cluster category, acknowledging and benefiting from the coexistence of pluralistic and consensus conceptions of proof. From the pluralistic perspective, an approach relying on the collective expertise of the judges is necessary to avoid the unreliability likely to result from any criteria-based approach. From the consensus perspective, I gain the theoretical comfort that grounding validity in the collective expertise of judges is a reasonable approach, and moreover, is likely to generate reliable scores.

One aim of this thesis is to understand the validity of two comparative judgment-based assessments. I report evidence from expert testimony and content analyses of responses, as well as comparing scores with students' performance on the multiple-choice Proof Comprehension Tests of Mejia-Ramos et al. (2017). The availability of these three tests provides a productive starting point for understanding the scope of comparative judgment in proof comprehension assessment in general. If it is the case that all three assessments are adequately related to comparative judgment-based assessments, this will be interpreted as evidence that this approach to assessment may generalise to other proofs for which no such psychometrically validated tests yet exist. The extent to which such findings can and should be generalised is an important consideration and is discussed at several points later in this thesis. For now, it suffices to say that this approach has the potential to generate whole classes of assessments with drastically lower resource requirements than the design approach proposed by Mejia-Ramos et al. (2017).

## 2.5 Research questions

The following research questions address the various gaps in the literature discussed above.

***Research question 1***:
*What do students and mathematicians write when explicitly asked about their conceptions of proof?*

***Research question 2a***:
*What do mathematicians most value when evaluating the written proof conceptions of others?*
***Research question 2b***:
*What do mathematicians most value when evaluating students' proof summaries?*

***Research question 3a***:
*Do written proof conceptions, scored using comparative judgment, generate a reliable and valid output?*
***Research question 3b***:
*Do proof summaries, scored using comparative judgment, generate a reliable and valid output?*

Questions 1, 2a and 2b serve two purposes. First, they represent interim focal points, of interest in isolation. Answers to these questions, however, also contribute to the more general questions of reliability and validity addressed by questions 3a and 3b. Table 2.3 summarises the empirical work addressing each question.

This thesis offers parallel contributions to the literatures on comparative judgment and proof. To the comparative judgment literature, it offers a series of methodological considerations regarding the applicability of this assessment approach in as-yet-unexplored content domains. By evaluating the reliability and validity of the scores produced by the two tasks, it generates an understanding of the strengths and weaknesses of comparative judgment-based approaches and draws general methodological conclusions about comparative judgment itself. To the literature on proof, it introduces and evaluates a new measurement approach for proof conceptions and proof comprehension. Regarding conceptions and beliefs, this approach has the potential to quantitatively evaluate responses in domains where previous research has been limited. Regarding proof comprehension, comparative judgment has the potential to offer an efficient assessment approach that captures students' understanding of a wide range of proofs. The line of research presented in this thesis also provides insight into the nature of proof itself.

*Table 2.3*

*Thesis outline.*

| Chapter | Data | Analytical focus | Research questions |
|---|---|---|---|
| **Part One** | | | |
| 2: Literature review | | | |
| 3: Methodology | | | |
| **Part Two** | | | |
| 4: Conceptions I: Preliminary investigation | Written conceptions, module scores, expert judgments, non-expert judgments | Reliability, discriminant and content validity | 1, 2a, 3a |
| 5: Conceptions II: A longitudinal study | Written conceptions (beginning/end of module), expert judgments | Predictive and content validity | 1, 2a, 3a |
| **Part Three** | | | |
| 6: Summaries I: The uncountability proof | Proof summaries, expert judgments, module scores | Reliability, convergent and content validity | 2b, 3b |
| 7: Summaries II: The primes proof | Proof summaries, expert judgments, module and SAT scores | Reliability, convergent and content validity | 2b, 3b |
| 8: Summaries III: The Fibonacci proof | Proof summaries, expert judgments, module and SAT scores | Convergent and content validity | 2b, 3b |
| 9: Summaries IV | Judge interviews | Judges' decision-making, content validity | 2b |
| **Part Four** | | | |
| 10: Final discussion | | | |

# Chapter 3

# Methodology

> 'Reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected'.
>
> (Peirce, 1878, as cited in Menand, 1997, p. 5)

In this chapter, I address the various methodological decisions embedded in this thesis. I position the two empirical strands (on the Conceptions and Summary Tasks) as mixed methods investigations, based on the ideas of pragmatism.

I first discuss mixed methods research, the necessity for such a paradigm in my research, and its philosophical underpinnings. This is followed by a section on validity, including glossaries of the relevant terms in assessment and research validity. I then use this language to discuss the various data and analyses featured in the following chapters. This chapter ends with a brief note on sources of ethical approval for this research. The specific methods of each study can be found in Chapters 4 to 9.

## 3.1 Mixed methods research

In recent decades, mixed methods research has established itself as a third major research paradigm alongside purely qualitative and quantitative approaches. In this section, I first identify a definition from Johnson and Onwuegbuzie (2004) before exploring the diversity of the mixed methods paradigm. I then position my work within the field using the eight-part taxology of Leech and Onwuegbuzie (2009). Finally, I introduce pragmatism as a philosophical underpinning for this work and discuss its implications for mixed methods research.

### 3.1.1 Defining mixed methods research

Johnson and Onwuegbuzie (2004) defined mixed methods research as 'the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study' (p. 17).

As a relatively young paradigm, there is substantive diversity in the literature about what constitutes mixed methods (or simply 'mixed') research and how it should be conducted. In reviewing the definitions provided by 24 leading researchers in the field, Johnson and Onwuegbuzie (2007) identified five themes: what is mixed, when/where is it mixed, why is it mixed, the breadth of applications and the orientations of those doing the mixing. Regarding what is being mixed, the authors found near-consensus that mixed research involved the combining of qualitative and quantitative approaches/tools. Beyond this most basic consideration, experts definitions varied on the other four themes. Some argued for mixed research involving mixing in the design stage, others prioritised the data collection or analysis stages (when/where). Some said that mixed research necessarily involves mixing at all stages (breadth). Some said the purpose of mixed research is corroboration, while others emphasised depth and richness of understanding (whys). Finally, some definitions oriented their approach as bottom-up, in the sense that mixed research is necessarily driven by research questions, while others espoused a top-down model where the researchers' goals and orientations are seen as the epistemic driving force.

According to Johnson and Onwuegbuzie (2007), 'definitions can and will usually change over time as the [field] continues to grow' (p. 112). In the meantime, researchers should attempt to be specific about the operational and epistemic assumptions upon which they base their work (*ibid*).

#### Three dimensions for identifying mixed methods research

Building on the work of Johnson and Onwuegbuzie (2007), Leech and Onwuegbuzie (2009) identified three dimensions along which mixed research can operate: a) breadth of mixing, b) time-orientation of mixing and c) the relative emphasis between approaches.

By breadth of mixing, the authors refer to the location of the mixing, either in the design, collection or analysis phases of a study. A fully mixed study involves mixing at all three phases, while a partially mixed study does so at only one or two of the three. Time-orientation refers to the timing of the data collection as either concurrent or sequential. A concurrent study involves qualitative and quantitative data collected simultaneously, while a sequential study involves data types collected one after the other. Finally, the relative emphasis between approaches refers to the importance of the quantitative/qualitative approaches embedded in the research. A quantitative-dominant study views the quantitative data as most important, and uses qualitative data to inform understanding of the primarily quantitative focus. Insofar as this dimension can be viewed as dichotomous, a qualitative-dominant study does the opposite.

**My research**

In this thesis, Chapters 4 through 8 present a series of partially mixed, sequential, quantitative-dominant studies, each oriented toward either research question 3a or 3b (see page 34). Research questions 1a, 1b, and 2 are also addressed at various points in each study but are not the primary thrust of this research. All studies are partially mixed in the sense that the mixing takes place only at the analysis stage in most studies. They are sequential in the sense that each data collection focused exclusively on either primarily qualitative or quantitative evidence. Finally, they are quantitative-dominant in the sense that all but one study begins with quantitative analyses, implementing more qualitative tools only after initial conclusions have been established. The exception is the interview-based study presented in Chapter 9. This, the final empirical chapter, is presented as a qualitative-dominant study, with mixed analysis coming only in the final stages via a multiple regression on the codes generated from a thematic analysis of interview transcripts.

### 3.1.2 Justifications for mixed methods research

Justifications for mixed methods research have been outlined in full several times (Doyle et al., 2016; Leech and Onwuegbuzie, 2009; Onwuegbuzie and Johnson, 2006). This section highlights four pertinent justifications from Doyle et al.'s list.

*Design flexibility.* Mixing methods allows researchers to ask and answer different questions that cannot be adequately addressed by quantitative or qualitative methods in isolation (Creswell and Plano Clark, 2011; Doyle et al., 2016). Some questions require a variety of approaches to generate robust evidence. Questions of assessment validity necessitate this design flexibility and should be approached from multiple angles.

*Triangulation.* Triangulation facilitates increased validity through corroboration of findings from different philosophical positions (Pinto, 2010). Corroborating conclusions from varied backgrounds increases confidence in the findings.

*Offsetting weaknesses.* Every source of evidence is built on (often unknowable) combinations of bias and contextual features, each with its own set of strengths and weaknesses. By gathering data from multiple epistemic sources, the probability that conclusions are the product of a specific combination of unknown influences is reduced. To this end, mixed methods researchers explicitly design studies with offsetting blind-spots with the intention of covering more ground.

*Completeness.* Similar to the flexibility highlighted earlier, Doyle et al. also emphasised the strengths of mixed methods designs in generating a 'complete and comprehensive picture of the study phenomenon' (p. 178).

The investigation of judges' decision-making processes illustrates the value of this mixed methods approach. Chapters 6, 7, and 8 generate conjectures about judges' priorities in evaluating proof summaries through a content analysis of summaries (and their features) most likely to be rewarded by judges. However, this primarily quantitative investigation ignores much of the contextual information surrounding judges' decision-making processes and relies exclusively on statistical inference to understand the complexity of human decision-making. In contrast, Chapter 9 presents an interview-based study in which transcripts are explored using thematic analysis to generate a richer, more authentic account of the judges' processes. By triangulating between the findings from the two approaches, it is possible to generate a more complete understanding of the situation by using methods with offsetting weaknesses to investigate the same phenomenon.

### 3.1.3 The philosophy of mixed methods research

#### Paradigm wars

To understand the foundations of mixed methods research, it is necessary to understand its origins and its birth as a 'third-way' alternative to strictly quantitative and qualitative paradigms. The debate of the 1980s and 90s, often referred to as the paradigm wars (Gage, 1989), led researchers from both traditions to adopt purist and, at times, dogmatic philosophies dismissive of the merits of the other (Johnson and Onwuegbuzie, 2004). On one side, quantitative purists espoused a philosophical position often associated with logical positivism. They argued that social science should be treated just as physical science, in that the observer can meaningfully be separated from the observed and that research can and should be conducted from an objective vantage-point (Howe, 1988). Researchers from this methodological tradition rely heavily on statistical evidence to test clearly defined hypotheses.

At the other end of the spectrum, qualitative purists rejected these positivist notions and argued that social science cannot be productively approached as such. Rather, purist qualitative researchers (often labelled constructivist or interpretivist, Johnson and Onwuegbuzie, 2004) claimed that research and researcher are inextricably linked and that researchers should be explicit about their underlying assumptions, biases and the full context of the research setting. To this end, qualitative paradigms permit very different forms of evidence, based on rich descriptions of naturalistic observations.

For each paradigm, the research conclusions produced by the other are either incomprehensible or invalid (Gage, 1989). To the positivist, the interpretivist's evidence lacks rigour, specificity and generalisability. To the interpretivist, the positivist lacks the capacity to paint sufficiently nuanced pictures to have relevance to the real world (Doyle et al., 2016; Gage, 1989).

### The incompatibility thesis

The incompatibility thesis, built on ideas from the paradigm wars, asserts that qualitative and quantitative paradigms cannot be mixed. They are built on different philosophical foundations, and blending these approaches leads to ill-defined territory in need of either new foundations, or a detailed account of the marriage between two seemingly opposed positions. The phrase 'incompatibility thesis' was coined by Howe (1988) as a description of the criticisms levied at those promoting the parallel use of qualitative and quantitative methods. Debate on the merits of this objection is on-going (Doyle et al., 2016; Hathcoat and Meixner, 2017).

### Pragmatism

Most mixed methods research is based on *pragmatic* philosophical positioning (Johnson and Onwuegbuzie, 2004). Pragmatism attempts to find a happy medium between warring paradigms, benefitting from a pluralism of evidential forms and rejecting the incompatibility thesis on the basis that the merits of inquiry can only be assessed in relation to the questions they attempt to answer.

Originating with Peirce's (1878) pragmatic maxim, pragmatists assert that one should 'consider the practical effects of the objects of [their] conception'. Extending Peirce's original work, James (1907) wrote that 'the pragmatic method is primarily a method of settling metaphysical disputes that otherwise might be interminable. The pragmatic method in such cases is to try to interpret each notion by tracing its respective practical consequences' (p. 18).

In the social sciences, Dewey (1948) interpreted these ideas as suggesting that researchers should choose the combination of methods and analyses best suited to answering the questions at hand. The pragmatist then views their day-to-day findings as 'provisional truths' and permits variety in evidential forms: qualitative, quantitative or a mix. These provisional truths are therefore likely based in differing paradigms and hence have differing philosophical justifications. In combining these provisional truths into more robust 'absolute *Truths*' (Johnson and Onwuegbuzie, 2004, p. 18), the pragmatist claims that an argument should not be viewed as a chain whose strength is determined by its weakest link, but as 'a cable whose fibres may be ever so slender, provided they are sufficiently numerous and intimately connected' (Peirce, 1878, as cited in Menand, 1997, p. 5).

There are many modern versions of pragmatism. Here, I follow the philosophy of Johnson and Onwuegbuzie (2004), referred to in their later work as a 'pragmatism of the middle' (Johnson and Onwuegbuzie, 2007, p. 125), as an attempt to split the difference between realist and pluralist pragmatisms (*ibid*). In their 2004 paper, the authors offer an extensive list of the general characteristics of their middle-of-the-road pragmatism (p. 18). In short, my interpretation of Johnson and Onwuegbuzie's pragmatism views knowledge as being 'both [socially] constructed *and* based on the reality of the world we experience (original emphasis)', while truth and meaning are regarding as dynamic entities, changing with time as new knowledge emerges from research.

### Criticisms of pragmatism and mixed method research

Pragmatism as a philosophical position is not without its shortcomings and does not claim to solve the many existing debates between more established paradigms. However, as has been argued by many scholars, pragmatism does offer a productive paradigm for integrating mixed methods research and its practical benefits outweigh these drawbacks in many research settings (Johnson and Onwuegbuzie, 2004; Doyle et al., 2016). Much like Czocher and Weber's (2019) cluster account rescues proof from a potentially intractable diversity of proof conceptions (see Section 2.2.3), pragmatism offers a productive perspective from which certain research questions can profitably be approached. Here, I explore two classical critiques of pragmatism, one based on the incompatibility thesis, the other focused on the absence of specificity in integrating and evaluating evidence.

First, many critics of pragmatism point at the incompatibility thesis (Hathcoat and Meixner, 2017; Howe, 1988; Johnson and Onwuegbuzie, 2004) as presenting a problem not yet accounted for by the literature. Other established paradigms have well-articulated philosophical accounts for truth. Critics claim that pragmatists firmly assert that it is permissible to accept multiple sources of evidence in a single investigation. However, they are often not forthcoming with a philosophical justification for truth and knowledge in their new paradigm. I identify the pragmatic literature as offering two distinct responses. The first is to dismiss the importance of the incompatibility thesis. Some argue that the approach leads to productive research on many topics that other paradigms have failed to penetrate and that alone makes it merit-worthy (Hathcoat and Meixner, 2017; Johnson and Onwuegbuzie, 2004). Johnson and Onwuegbuzie explicitly note that a more robust philosophical justification is desirable and that researchers should continue work in this direction, but the absence of traditionally robust philosophy should not deter researchers from mixed methods approaches.

Others are more hostile to the incompatibility thesis, not just dismissing its importance, but rather claiming that it is built on a faulty premise. In particular, Howe (1988) asserted that the incompatibility thesis is based on the untenable claim that 'abstract paradigms should determine research methods in a one-way fashion' (p. 10). Howe advanced the view that paradigms must demonstrate their 'worth in terms of how they inform, and are informed by, the research methods' with which they associate (*ibid*). Thus, Howe promoted a 'what works' (p. 14) approach where each isolated line of inquiry is based on isolated considerations specific to the needs of the research question at hand. Howe's defence can be considered in terms of Peirce's metaphor for arguments as cables, not chains. In Peirce's metaphor, I claim that the weakest links are the philosophical relationships between forms of evidence, and the incompatibility thesis targets exactly this, the weakest links. However, Peirce asserts that one should not think in terms of weakest links, but in terms of the numerous and intimately connected strands of an argument.

In responding to the incompatibility thesis, pragmatists claim that it is acceptable to mix methods because it is productive. However, these justifications are light on specificity and critics have questioned the ability of pragmatism to inform how, when and where research should integrate its evidential sources. For example, when the canonical pragmatist asserts that researchers should use methods most suited to the question at hand, Mertens (2003) demanded more specificity on for whom the methods are suitable. In response, Johnson and Onwuegbuzie (2004) explicitly note the value-laden, researcher-centric nature of many mixed methods investigations and claim that, consistent with other naturalistic paradigms, this is not problematic as long as researchers are explicit about their assumptions and biases where relevant. In a similar light, pragmatism provides no direct guidance on the evaluation of evidence sources, leaving it to the researcher to determine what constitutes convincing evidence and how differing (both corroborating and contradictory) sources of evidence should be integrated (Johnson and Onwuegbuzie, 2004). In the absence of a single methodological approach to research, the tasks of evaluating and integrating evidence are left without substantive guidance.

To compensate for this gap in the philosophical literature, I turn to the literature on (assessment and research) validity for guidance on how to evaluate and integrate evidence in my research in particular.

## 3.2 Validity

In this section, I first discuss topics in assessment validity, adopting the notion of construct validity as further justification for the mixed methods approach used in this thesis. I then define six forms of assessment validity that form the basis of later discussions, both in contrasting assessment and research validity, and in understanding the empirical work in the following chapters. I define two broad types of research validity in terms of the preceding assessment language. This language is then used in the following section outlining the forms of data collection and analysis that feature in this work.

### 3.2.1 Assessment validity

#### Early positivist conceptions of assessment validity

Early conceptions of validity emerged at the turn of the 20th century and were grounded in a positivist tradition of education research (e.g. Spearman, 1904). According to Shaw and Crisp (2011), validity in this era was viewed largely as a statistical entity capturing a test's capacity to produce scores that correlate with established measures with a theoretically similar premise. The quantities measured were assumed to be static, objective entities with a definite value for each individual and validity was viewed as the accuracy of a measure's estimates of those values. In contrast to more modern conceptions of validity, these earlier works viewed validation as a question that could be definitively answered, and that with sufficient evidence, it was possible to unconditionally declare a measure to be valid (Kane, 2001).

### A modern conception of validity

According to Shaw and Crisp (2011), modern assessment validity is based on a model of *construct validity*, introduced by Messick (1989). Messick's construct validity was built on three claims that distinguish his work from that which came before it. First, Messick positioned the process of validation as the construction of an argument, rather than the deciding of a property to be assigned to a particular test. Second, in describing validation as an argument, Messick permitted and encouraged a full gambit of analytical approaches, asserting that 'construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores' (*ibid*, p. 23). This excerpt also alludes to a third epistemic feature of Messick's conception, explicitly shifting the focus from validating tests to validating scores. Further, Messick noted that test scores are contextual and their interpretation is only meaningful insofar as they measure something. Hence, it is meaningless to talk of a test, or set of scores, as valid or invalid. Rather, we should talk of test scores as (in-)valid measures of something and that something must be viewed in its social and political context (Onwuegbuzie and Johnson, 2006).

Consistent with the pragmatism of mixed methods research discussed earlier, Messick's construct validity is often referred to as 'post-positivist' (Shaw and Crisp, 2011; Kane, 2001), without further specification about its philosophical underpinnings. As Messick wrote, the role of test validation is 'to marshall evidence and arguments in support of, or counter to, proposed interpretations' (p. 43) and can legitimately call on any mode of inquiry. 'After all, within loose limits of scientific respectability, the issue is not the source of the evidence and arguments but, rather, their nature and quality.' (*ibid*). In this light, I aim to understand the validity of the scores generated by comparative judgment-based assessments via a series of mixed methods studies.

To facilitate later discussion of topics in assessment validity, I provide a brief glossary of terms subsumed under construct validity, and their relation to my review of the literature presented earlier.

#### Construct validity

Construct validity is an abstract overarching entity, incorporating all of the operational forms of validity described below. According to Messick (1989), construct validity is a framework for generating conclusions 'based on an integration of any evidence that bears on the interpretation or meaning' of the measure (p. 23).

#### Content validity

Content validity is the degree to which a measure represents the domain about which the conclusions are derived (Messick, 1989). This is most frequently evaluated using expert testimony and a range of primarily qualitative methods (*ibid*). See Section 2.4.3 for the corresponding comparative judgment-focused literature labelled 'expert testimony' and 'content analysis'.

#### Concurrent validity

Sometimes called convergent validity (Johnson and Onwuegbuzie, 2007), concurrent validity concerns comparisons between two or more measures collected simultaneously (Messick, 1989). This is a fundamental quantitative tool in establishing preliminary validity of new assessment tools and is considered in several places through the empirical chapters

#### Predictive validity

Predictive validity is the extent to which a measure predicts future performance. This can be examined in a longitudinal study of change using the same measure (as in Chapter 4), or as a prediction of external measures collected at a later date (as in Chapter 9).

#### Discriminant validity

Sometimes called divergent validity (Johnson and Onwuegbuzie, 2007), discriminant validity is the extent to which a measure discriminates between theoretically distinct entities. I investigate discriminant validity using comparisons between expert and non-expert judgments, and between mathematicians' and students' performance (Chapter 4).

#### Criterion validity

Criterion validity is the extent to which a measure replicates theoretically associated external measures (or criteria). Concurrent, predictive and discriminant validity are all types of criterion validity. The comparative judgment literature on criterion validity is summarised in Section 2.4.3, under the label 'Comparative analysis', and is usually based on quantitative analyses.

### 3.2.2 Research validity

Consistent with Messick's (1989) construct validity, my empirical work builds arguments for using both quantitative and qualitative methods. As discussed earlier, the pragmatist's positioning permits a multiplicity of evidential sources, but it does not permit the researcher to omit considerations of robustness and quality for each of the respective sources of evidence.

In this section, I address these concerns through the language of research validity. Having established a glossary of relevant forms of assessment validity, I now address topics in research validity pertinent to the data collection and analysis to come. For the purposes of this thesis, I refer simply to internal and external validity, and explore their relationship with the literature on assessment and comparative judgment. For a full exploration of validity in mixed methods research, see Onwuegbuzie and Johnson (2006), in which the authors present more than 50 interrelated concerns and provide several models for understanding their relationships.

#### Internal validity

Internal validity refers to the relationship between the data-driven conclusions and the phenomenon of interest. In purely quantitative terms, this is often phrased in terms of accuracy and within-assessment validity. In assessment, this is most commonly considered in terms of criterion, and more specifically convergent, validity. Comparative judgment research often examines convergent validity by comparing comparative judgment-based scores with established measures of theoretically similar constructs. Reliability measures, such as Scale Separation and inter-rater reliability (described later), can also be viewed as measures of internal validity because they are measures of variation between judges and hence can be seen as measures of accuracy.

In more qualitative terms, internal validity is more complex, and is the realm of credibility, confidence and plausibility (Cohen et al., 2000). In this thesis, this is particularly pertinent to the collection and analysis of interview data and is discussed in detail in sections on data collection and analysis.

**External validity**

External validity is often phrased as the business of generalisation (Cohen et al., 2000; Pinto, 2010). In quantitative research, this is a question of isolating and controlling phenomena observable in carefully selected samples of a population. This is closely connected to discriminant validity in terms of assessment. By comparing outputs to theoretically distinct measures, we can investigate and begin to eliminate the possibility of theoretically plausible confounds.

In qualitative settings, external validity is a more subtle process wherein many interpretivists in the paradigm wars rejected its relevance (Johnson and Onwuegbuzie, 2004; Howe, 1988). However, according to Cohen et al. (2000), external validity is about 'comparability and translatability' (p. 109). Here, generalisability can be profitably viewed as the capacity of a given evidential source to be interpreted and integrated with the findings of others to produce theoretical insights for the field. I return to this notion of integration in discussing the mixing of analytical tools in the following section.

## 3.3 Data collection

In this section, I consider the types of data collected in this thesis, potential threats to the quality of these data and the steps taken to evaluate and/or improve their quality.

### 3.3.1 Written responses

All studies conducted in this thesis are based on written responses to either the Summary or Conceptions Tasks. These responses were collected in classroom settings and for practical and ethical reasons, no explicit extrinsic motivation was provided in compelling students to complete the tasks.

I consider two internal validity concerns here. The first is the extent to which participants' responses to these tasks are reflective of their behaviour in other assessment settings. In the absence of extrinsic motivations, particularly module credits, the validity of these tasks as measures of performance may be questioned. This likely increases the noise in the data as some participants inevitably take voluntary tasks more seriously than others. This has consequences for examining the relationship between comparative judgment-based tasks and other measures, leading to possible underestimates of their true relationships.

By interpreting the data in light of this threat to validity, the integrity of the conclusions are preserved.

My second internal validity concern comes from the notion of written responses in general. Written assessment is not the only, nor necessarily best, form of assessment. The extent to which written tasks capture a holistic view of understanding has been questioned many times (Iannone and Simpson, 2011, 2013). This is particularly pertinent in the capturing of proof conceptions using a word-limited written task. To this concern, I concede that an ideal investigation would indeed comprise a wider variety of assessment modes for triangulation including oral and project-based assessments. This was a necessary trade-off in generating sufficiently large data sets suitable for the variety of analysis presented. This concern is mitigated by the variety of concurrent and divergent validity explorations presented.

From an external validity standpoint, I also consider questions of sampling from both student and mathematics communities. In recruiting students for this work, I took advantage of convenient module cohorts accessible at relevant institutions. The undergraduate participants were all first- and second-year students of mathematics from two academic institutions and are not necessarily representative of the entire student population at these institutions or beyond. However, the main focus of my research is on the evaluation of assessment approaches and developing understanding of expert communities, so the precise sampling of the student population is not of paramount importance. It is necessary that these participants are approximately representative of undergraduate mathematicians. However, given that the primary thrust of generalisation is not to the wider undergraduate community, but rather to other content domains, this research can tolerate the imprecision inherent in this work.

### 3.3.2 Comparative judgment

The written responses discussed above are evaluated using comparative judgment multiple times throughout this thesis. Given its centrality to this thesis, validity of comparative judgment is considered separately at the end of this chapter, presented alongside the technical details of the computations.

### 3.3.3 Multiple-choice Proof Comprehension Tests

The multiple-choice Proof Comprehension Tests used in this thesis are products of extensive research into the validity of their outputs (Mejia-Ramos et al., 2012, 2017). These tests feature in Chapters 6 - 9 as external benchmarks against which comparative judgment-based scores can be evaluated to provide insight into the concurrent validity of the measures in question.

I provide a brief theoretical account of the issues here, starting with three threats to the internal validity of these tests as they appear in this thesis: extrinsic motivation, the ceiling effect and random guessing. Extrinsic motivation was discussed in section 3.3.1. I address the other two here. Each test only features 12 items and suffers from a ceiling effect whereby the test cannot differentiate between the highest performing participants (Resch and Isenberg, 2018). This introduces an asymmetry between the tests and the comparative judgment-based scores with which they are compared, again leading to a possible underestimate of the true relationship.

At the other end of the spectrum, students may guess at random and will be correct approximately one-quarter of the time[1]. These correct guesses are indistinguishable from answers based on robust (or partial) knowledge, introducing noise inherent to all multiple-choice tests (Resch and Isenberg, 2018). This is a well-known limitation of multiple-choice assessment, but is particularly problematic when performance is low, as was the case for at least one dataset presented here.

In considering the external validity, the participant cohort must be carefully considered. In the relevant aspects of my research, I recruited participants in their first and second years of study in undergraduate mathematics degrees. This was done, in part, to mimic the academic background of the US-based participants from the original work on these tests (Mejia-Ramos et al., 2012). In this way, the tests should have been at an appropriate level to yield meaningful comparisons with the original work, and to have statistical properties making them appropriate for comparison with other measures. However, this was not always empirically the case. The internal consistency (captured by Cronbach's alpha) and mean scores differ substantively in some cases from that presented in the original publication (Mejia-Ramos et al., 2017), leading to a difficult problem for the elements of my research designed to rest upon these reportedly robust tests.

[1] All items have four options, while a small subset require students to 'select all that apply' dropping the expected score slightly below 25% on each test.

### 3.3.4 Module data

As a measure of performance, module scores[2] suffer from an inter-rater reliability problem and from a content validity problem as outlined by Mejia-Ramos et al. (2017) and Weber (2012) who argued that traditional mathematics assessment is often too narrow or shallow to fully capture proof comprehension. These data are used to evaluate concurrent validity as a measure secondary to the more robust Proof Comprehension Tests. Its inclusion is a product of practical considerations and availability. I concede that the internal validity is not perfect in this case and consider the strength of the conclusions from the resulting analyses in this light.

### 3.3.5 Interview data

In investigating judges' priorities when judging students' proof summaries, I conducted a series of semi-structured interviews with mathematician judges. These are presented in Chapter 9 with the intention of triangulating these interview data with the quantitatively oriented results in the preceding three chapters. My role as researcher and interviewer influences these data in two ways. First, as a researcher, I bring with me a series of assumptions and preconceptions to the interactions based on previous experiences (Persaud, 2010b). In particular, in light of the triangulation purpose of these interviews, I necessarily take in a set of expectations about the nature and type of answers I am likely to receive and the various approaches judges will take. To mitigate this influence, I followed the advice of Persaud (2010a) by predetermining an interview schedule and a series of intentionally neutral responses to expected answers where possible. However, it is impossible to be entirely objective and it is important to understand the inherently personal nature of the interactions that take place (*ibid*).

There are further internal validity concerns from the perspective of the participant judges. Judges were first asked to make a series of 20 decisions, before being asked about the features influencing their decision in abstract then concrete terms. In asking for judges to justify particular decisions I, as the interviewer, unavoidably created a social dynamic wherein judges may experience pressure to produce a cogent answer to the question (Hevey, 2010). In cases where there was little to choose between two summaries, this presents a prob-

[2] I use the term module to refer to a period of (undergraduate) study focused on a particular sub-discipline, usually lasting between 10 and 15 weeks. These are often referred to as courses in other parts of the world. In the UK, a standard undergraduate degree (or course) requires students to complete approximately 36 modules over 3 years.

lematic situation whereby judges are likely to produce post hoc justifications that may not have been present in the moment and may not be meaningfully reflective of that judges' decision-making. Further discussion of this potentially problematic feature of the data is presented in the discussion section of Chapter 9.

There are also external validity topics to discuss based on the laboratory setting in which these interviews were conducted. For practical reasons, the structure of the exercise was explained to each participant at the beginning of the session, before their initial judgments. This was deemed necessary to promote active engagement in the semi-structured interview to come (Hevey, 2010). However, this may have detracted from the external validity of the responses. Unlike comparative judgment data collected remotely, judges were warned that they would be asked to justify their decisions.

Further attempts to improve the validity of the interview data are explored in discussions of transcription and thematic analysis discussed in the next section.

## 3.4 Data analysis

This section concerns the approaches to data analysis featured in the chapters that follow. Again, I consider potential threats to validity and the steps taken to minimise their impact where possible. As in the previous section, topics related to comparative judgment are omitted here, to be included in their own section alongside other comparative judgment-related discussions.

### 3.4.1 On the mixing of methods

On several occasions in this thesis, I mix several data sources into a single analysis. In its simplest form, this takes the form of regression modelling, predicting comparative judgment-based scores using the coded data resulting from content/thematic analyses of responses. Here, I return to the taxology of Johnson and Onwuegbuzie (2007) and the framing of this work as 'partially mixed, sequential, quantitative dominant' research. While fundamentally quantitative, these analyses are embedded in the pragmatic mixed methods paradigm and questions of validity can be thought of as an amalgamation of imperfect constituent parts (Johnson and Onwuegbuzie, 2007).

### 3.4.2 Transcription

I consider the transcription process here as part of the analytical (rather than the data collection) process because of the interpretative acts embedded in deciding which aspects of the interview to include in a written transcription (Braun and Clarke, 2006). In transcribing interview data, it is important to note the inability of the transcript to fully capture the social dynamics inherent in the conversational nature of a two-person interaction. I explore the specific methods of interview transcription and analysis further in Chapter 9, but note here the care required to generate faithful (internally valid) transcripts that capture the interview process. While much of the contextual evidence is lost in the transcription process, by attending to the precise phrasing of the participants' contributions, it is possible to generate data productive for understanding judges' explicitly stated motivations in choosing one response over another. There are subtextual motivations that remain invisible to the researcher, and the researchers' influence on the interview itself must be considered throughout the analysis. Again, the pragmatist's mixed methods paradigm tolerates these imperfections through the triangulation of findings with other analyses, offsetting strengths and weaknesses.

### 3.4.3 Thematic analysis

In Chapter 9, I present a thematic analysis (Braun and Clarke, 2006) of the interview transcripts discussed above. In doing so, I aim to generate an understanding of judges' decision-making in evaluating proof summaries. In contrast to the quantitative analyses, this work is necessarily embedded in more constructivist traditions in which knowledge and meaning are constructed through shared interactions between individuals (Willig, 2013). Particularly pertinent to this form of investigation is the understanding that the researcher necessarily brings to any research an inescapable set of biases and predispositions that inform their work. In particular, the analysis presented in Chapter 9 is informed by a series of pre-determined research questions (set out in Chapter 9) and is motivated by the wider investigations of validity. In this light, I neither claim nor intend to present a holistic analysis of all the interviews' nuance and detail, but rather address specific notions of judges' motivation in a particular context. Insofar as this study attempts to understand the motivation of expert mathematicians, reported analyses are my interpretations of the actions and utterances of others, who have their own sets of biases and predispositions.

Details of the methods used in conducting this thematic analysis are presented in Chapter 9.

### 3.4.4 Content analysis

In attempting to understand the written responses discussed in 3.3.1, I present content analyses in several chapters. These analyses are based on the principles of thematic analysis discussed by Braun and Clarke (2006) with two important distinctions regarding the generation of coding schemes. Unlike in more comprehensive thematic analyses where a familiarity with the entire dataset is necessary to generate initial codes (Braun and Clarke, 2006), initial content analysis code schemes were generated using only a subset of the responses. This is appropriate in the majority of studies presented in this thesis, given the length of the responses in question and the concrete nature of the relevant content. In content analyses of proof summaries, initial coding schemes were based on subdivisions of the original proof, and were then amended as needed throughout the analysis process. Regarding proof conceptions, initial coding schemes came from examining a subset of the responses received in light of the established literature on the topic. While this abbreviated analytical process risks internal validity as a faithful representation of the entire dataset, this approach was deemed fit for purpose for its balance between efficiency and rigour.

The second distinction from Braun and Clarke's thematic analysis was in the number of researchers involved. Where thematic analysis emphasises rigorous personal reflection, the content analyses presented here were conducted by at least two researchers on every occasion. What was lost in internal validity through isolated reflection is replaced by checks of inter-coder reliability. Details of the specific methods used in each case are discussed in Chapters 4 through 8.

### 3.4.5 Corroboration via presentation

A final aspect of my data analysis process was the presentation of preliminary findings at various conferences, department-wide seminars and research group meetings. On each occasion, the feedback received was instrumental in shaping the analysis itself, and in understanding the most effective modes of communicating the results I perceived to be most important. This process was particularly instrumental in analysing and learning to communicate the interview data presented in Chapter 9.

## 3.5 Comparative judgment

Finally in this chapter, I return to comparative judgment, the methodological tool upon which the majority of my research is based. I present the technical details of comparative judgment, providing commentary on the theoretical and practical implications where necessary.

Comparative judgment is a tool for estimating subjective human perceptions by quantitative values indicative of those perceptions. Recall from the end of Chapter 2 that research questions 1a and 1b invoke comparative judgment as a research tool for gathering insight into the behaviour of mathematicians, while questions 3a and 3b focus on comparative judgment as a tool for assessment, addressing the reliability and validity of the resulting scores.

I first discuss the theoretical model of human perception underpinning the translation of subjective perceptions to pairwise comparisons. I then explain the computations necessary in transforming the pairwise comparisons (now referred to as judgments) to numerical estimates of the perceived quality of each response. I then turn to theoretical and computational notions of the reliability of the scores produced.

### 3.5.1 Using mathematics to model human perceptions

In this thesis, I am interested in estimating the merit of mathematical texts using experts' pairwise comparisons. This requires a mathematical model of the way experts perceive these texts in isolation. This model of perception, known as the law of comparative judgment (Thurstone, 1927), can then be used in conjunction with empirical judgment data to generate an estimate of the likelihood of a judge choosing one text over another and, eventually, a numerical estimate of the quality of each text.

Each time a judge encounters a text, $A$, it is perceived as lying somewhere on a continuum of merit. I say $A_i$ is the merit assigned to text $A$ in encounter $i$. Thurstone (1927) called the process of assigning that merit the *discriminal process*. This may be unstable in time as a judge may perceive the merit of a given text as different in different encounters (perhaps influenced by mood, time of day, or the other texts that judge has recently encountered).

The $A_i$ are assumed to be normally distributed, with standard deviation $\sigma_A$, about the mean, $v_A$. I use $v_A$ here to connote the collective 'value' assigned to the text $A$, via the various encounters with $A$. We call these normal distributions discriminal dispersions (Bramley, 2007). See Figure 3.1, showing the discriminal dispersions for texts $A$ and $B$ with distributions $\mathcal{N}(\sigma_A, v_A)$ and $\mathcal{N}(\sigma_B, v_B)$.

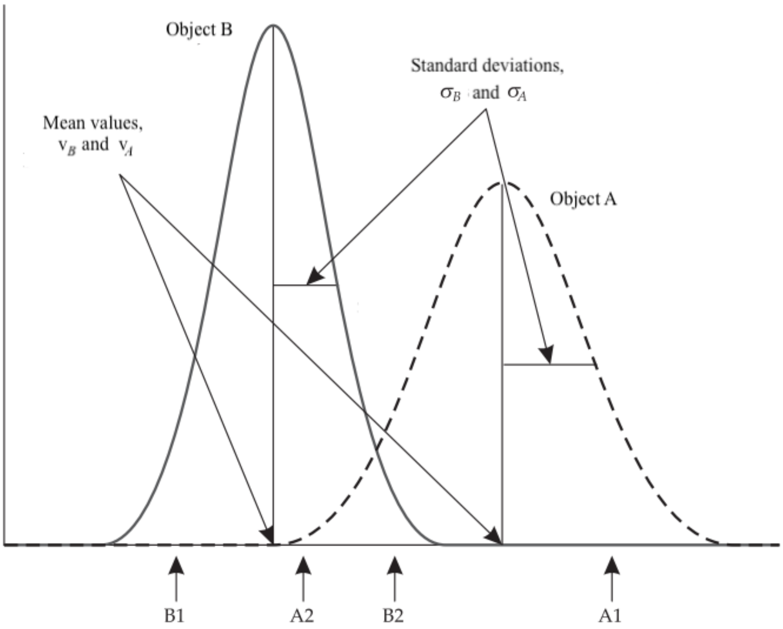


*Figure 3.1. Illustration of two overlapping distributions for texts A and B. Labels on the X-axis correspond to encounters of a given text. Adapted from Bramley (2007).*

When a judge compares two texts, Thurstone imagined that it is the values resulting from the discriminal processes that are compared. Algebraically, if texts $A$ and $B$ are compared, the judge will assert $A$ beats $B$, if $A_i > B_i$. Notice that in Figure 3.1, while the discriminal dispersion for $A$ is further along the merit continuum than $B$, the overlapping distributions mean that it is possible, based on some encounters of the two texts, for a judge to assert '$B$ beats $A$'.

To evaluate the likelihood of the two outcomes, we consider the distribution given by the difference of the discriminal dispersions, with standard deviations $\sigma_A$ and $\sigma_B$. Call this new distribution the paired discriminal dispersion, with standard deviation $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2 - 2R_{AB}\sigma_A\sigma_B}$ where $R_{AB}$ is the correlation between discriminal dispersions[3]. The new distribution is also normal, with mean $v_{AB}$, the difference $v_B - v_A$ (see Figure 3.2).

As is shown in Figure 3.2, for a given comparison, the probability that $A$ beats $B$ is the proportion of the distribution where $A_i > B_i$. This is the area to the right of zero under the curve with mean $v_A - v_B$, and is determined by the $z$-score of zero in the distribution $\mathcal{N}(\sigma_{AB}, v_A - v_B)$. This is used to state a

[3] This can be proven by considering the variance associated with $\sigma_A$ and $\sigma_B$.

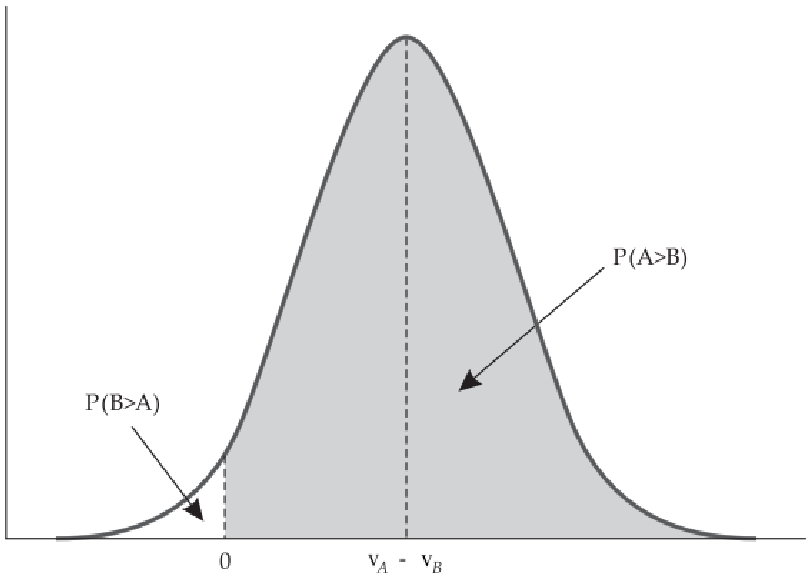


*Figure 3.2. An illustration of the distribution centered at $v_{AB}$ used to estimate the likelihood of A beats B (the shaded area to the right of zero) and B beats A (to the left of zero). Adapted from Bramley (2007).*

general form of Thurstone's law of comparative judgment as

$$X_{AB} = \frac{v_A - v_B}{\sigma_{AB}},$$

where $X_{AB}$ is precisely the $z$-score of zero.

The most general version of the law of comparative judgment applies only to comparisons of a single pair of objects. To generate a model more useful for practical application, Thurstone proposed a series of five cases, each imposing stricter assumptions than the last. The fifth and final case assumes that every object has the same discriminal dispersion, call it $\sigma$, and that all dispersions are uncorrelated, i.e. $R_{AB} = 0$. This results in $\sigma_{AB} = \sqrt{2}\sigma$ and allows the simplification $X_{AB} = \frac{v_A - v_B}{\sqrt{2}\sigma}$. The denominator here is constant and can be considered an arbitrary unit of measurement so, without losing information, we can equivalently state the law of comparative judgment as

$$X_{AB} = v_A - v_B.$$

Our unit of measurement imbues a particular meaning on the scores produced in the eventual model, allowing scores to be interpreted as standard deviations from the mean.

From Thurstone's assumptions, we can approximate the likelihood that '$A$ beats $B$' by considering the area to the right of zero, under the standard normal curve centred at $v_a - v_b$. This is given by

$$P(A > B) = \frac{1}{\sqrt{2\pi}\,\sigma_{AB}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{[t-(v_A - v_B)]^2}{\sigma_{AB}^2}\right) dt.$$

## From a model to a measure

Recall that our goal is to estimate the relative merit of mathematical texts, based on a set of binary pairwise comparisons. So far, we have a way of mathematising a single comparison, $A$ vs $B$, and a probability model for estimating the likelihood of $A$ beating $B$ and vice versa. Before we can begin the process of assigning scores to texts, we have two problems to solve. The first is that the integral of the normal distribution function famously has no analytical solution. The second is that the above function is dependent on $v_A$ and $v_B$.

The first problem was solved by Andrich (1978), who proposed replacing the normal distribution with a logistic one with near-identical outputs:

$$P(A > B) = \frac{e^{\rho(v_A - v_B)}}{1 + e^{\rho(v_A - v_B)}}$$

By setting $\rho$, an arbitrary scaling parameter, to $1.7/\sigma$, Andrich's new model generates near-identical outputs for the two distributions (see Figure 3.3). However, the values generated by Thurstone's model have no particular importance, so for simplicity it is standard to set $\rho = 1$, resulting in the simpler logistic model,

$$P(A > B) = \frac{e^{(v_A - v_B)}}{1 + e^{(v_A - v_B)}}.$$

The value of Andrich's new model was an improvement in computability. We now have an easily solvable expression for the probability $P(A > B)$. However, we still have the problem of dependence on the unobservable $v_A$ and $v_B$. This was solved by Bradley and Terry (1952)[4], giving their names to the model of comparative judgment used in modern education[5].

[4] Bradley and Terry's original model was more general than Thurstone's, based on a set of less stringent assumptions. In particular, only Thurstone assumed an equivalence across discriminal dispersion. An in-depth discussion of the consequences of this assumption can be found in Bramley (2007), along with a justification for the numerical and theoretical equivalence between the two.

[5] Luce (1959) presented very similar work and on occasion, this model is referred to as the 'Bradley-Terry-Luce' model (E.g. Verhavert et al., 2018). In line with the majority of the comparative judgment literature, I refer only to the Bradley-Terry model from here on.

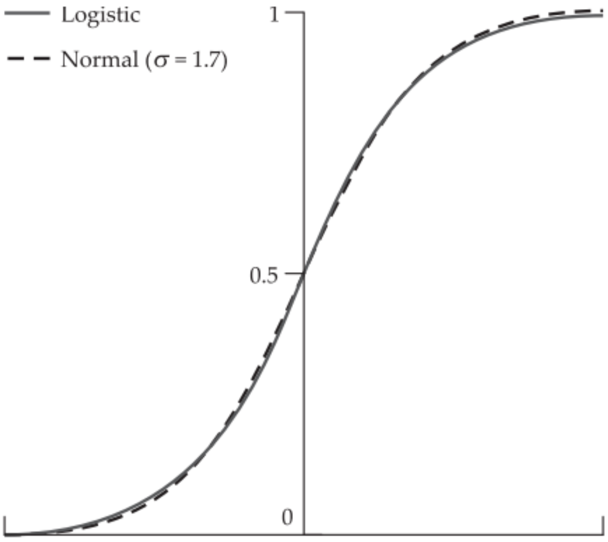


*Figure 3.3. A visual comparison of the logistic (Andrich) and Normal (Thurstone) models. Adapted from Bramley (2007).*

Using the probability expression above, the *Bradley-Terry* model takes in a set of $N$ pairwise comparisons on $M$ distinct texts, and outputs a set of numerical values, $v_i$, estimating the perceived merit of each text where $i \in \{1, 2, \ldots, M\}$. Notice that we recycle the notation from earlier, saying that $v_i$ estimates the merit of text $i$. This is a direct analogy whereby the $v_i$ are in fact estimates of the modal discriminals (the peak of the discriminal dispersion determined by the set of discriminal processes).

In summary, the Bradley-Terry model uses a Maximum Likelihood procedure (Rasch, 1960) to estimate the quality of a text, based on the number of comparisons won by that text. By comparing the number of comparisons a text actually wins with the number we expect it to win, we can iteratively improve our estimate of merit for each text.

To this end, we start by computing a raw score for each text. For this, we distill our $N$ decisions into numerical values by saying $D_{AB} = 1$ when $A$ beats $B$, and $D_{AB} = 0$ otherwise. Notice $D_{AB} = 0$ either when $B$ beats $A$ or when the texts $A$ and $B$ are not compared. We determine the raw score, $R_A$, as the number of comparisons won by $A$:

$$R_A = \sum_{i=1}^{M} D_{Ai}.$$

This raw score is, in itself, an estimate for the merit of each text. However, the raw score is not sensitive to the merit of the comparison set for each text. It is rarely possible to compare each text with every other text. A poor text could receive a high raw score by being compared only with other poor texts[6]. With this in mind, we want to generate a more nuanced estimate for the merit of each script.

To this end, we estimate the raw scores by replacing the binary values from $D_{AB}$ with the probability, $P(A > B)$. This gives a new estimate,

$$E(R_A) = \sum_{i=1}^{M} P(A > i) = \sum_{i=1}^{M} \left[ \frac{e^{(v_A - v_i)}}{1 + e^{(v_A - v_i)}} \right].$$

We have now returned to an earlier problem, where $E(R_A)$ is a function of the $v_i$, the very values we are eventually attempting to estimate. However, we now have all the tools in place to determine an iterative expression for the quality of $v_i$ using the Newton-Raphson method as follows:

$$v'_A = v_A + \frac{R_A - E(R_A)}{\sum\limits_{i=1}^{M} [P(A > i)][1 - P(A > i)]}.$$

At each iteration, we improve the estimate for each text based on the most recent $v_i$ for each script. All that remains is to determine an initial state for the iterative process. For this, we have a ready-made set of candidates in the raw scores, $R_i$. In this way, write $R_i = v_i^0$ where the superscript indicates the number of iterations of the Newton-Raphson method used. Notice that $v_i^k$, the current estimate of the merit of text $i$ at the $k^{\text{th}}$ iteration, is based on the $v_j^k$ for $j < i$ and $v_j^{k-1}$ where $j \geq i$.

This process generates an increasingly accurate score for each script at each iteration, and probably stabilises after approximately $k = 3$ iterations (Pollitt, 2012a). In this thesis, all implementations of the Bradley-Terry model are executed using the *btm()* function from R's *sirt* package, with a convergence criterion, $\varepsilon < 10^{-4}$. The maximum number of iterations was left at the default value, 100, and was not reached in any analysis. The *btm()* function is among the most robust tools for implementing comparative judgment and is consistent with best practice from the Rasch modelling literature (Verhavert, 2018).

[6] In principle, one could imagine a dataset with every text compared with every other exactly once and this problem disappears. However, such a dataset is impractical to generate.

### 3.5.2 Reliability

Having discussed the theoretical models and numerical computations used to estimate the quality of each text, I now turn to the reliability of these estimates. In particular, I discuss three standard measures of reliability, and potential measurement problems stemming from the use of each.

**Scale Separation Reliability (SSR)**

In comparative judgment research, by far the most common measure of reliability is Scale Separation Reliability (SSR), used by all studies cited in the literature review in Chapter 2. In literal terms, SSR measures how well the assessment separates the texts[7] and takes its name from its origins in Rasch analysis. Andrich (1978) showed that SSR can be interpreted similarly to Cronbach's alpha, and is thus often interpreted as a measure of internal consistency, with the same thresholds for success ($\alpha > .7$ as acceptable).

To compute SSR, we first compute a Separation Coefficient, $G = \mathrm{sd}_v/\mathrm{rmse}$, where $\mathrm{sd}_v$ is the standard deviation of the estimates $v_i$, and rmse is the root mean square of the estimation error[8]. This is then converted into Scale Separation Reliability,

$$\mathrm{SSR} = \frac{G^2}{1+G^2}.$$

While understood as a robust measure of internal consistency, SSR is sensitive to over-estimation based on the size of the data and the type of comparative judgment algorithm used (Jones et al., 2019; Verhavert, 2018).

In response to the potentially prohibitive volume of data required to generate reliable scores, researchers have sought to adapt the standard algorithm to generate pairings in which more 'information' is generated by each judgment Pollitt (2012b). Adaptive comparative judgment reportedly generates stable scores with fewer judgments than the non-adaptive approach (*ibid*). However, it also artificially inflates SSR, leading to a problematic basis upon which to conduct reliability research (Bramley, 2015; Bramley and Vitello, 2019). With this in mind, I use only non-adaptive comparative judgment in the research presented in later chapters.

SSR increases with the number of judgments, meaning that one gets a higher estimation of reliability simply by collecting more judgments. From a practitioner's perspective, this is arguably an asset, providing an indication of the

[7] The name was first introduced by Bramley (2007) following the observation that several authors were reporting the same measure by different names.

[8] The estimation error for each $v_i$ is computed using the inverse of Fisher's information Matrix (Hunter, 2004), equivalent to the covariance matrix.

minimal input necessary to produce a stable output. However, for research purposes, this is problematic particularly when the reliability of a measure is in question. This sensitivity also limits the meaningfulness of comparisons across studies with varying numbers of judgments.

### Inter-rater reliability

To evaluate inter-rater reliability, I use the split-half method introduced to the comparative judgment literature by Bisson et al. (2016). To produce this measure, judges are split, post-judging into two randomly generated groups and scores are recalculated for each group. Reliability is estimated by computing the Pearson Product-Moment correlation coefficient between the two groups. This procedure is repeated 100 times and the median correlation coefficient generates a measure of inter-rater reliability. Knowing that reliability increases with data size, this split-half process usually generates an under-estimate of reliability as a result of only using half the decisions in each isolated calculation. As a result, researchers are compelled to collect more data to generate the same conclusions than if they used only SSR.

This stricter measure of reliability is not as sensitive to the number of judgments and, despite the necessity for more data, has gained popularity in recent literature (Jones and Karadeniz, 2016; Jones et al., 2019; Bisson et al., 2016; Verhavert et al., 2018). In a meta-analysis of comparative judgment-based research, Verhavert (2018) demonstrated that this split-half measure is significantly correlated with SSR and can therefore also be meaningfully interpreted as a measure of internal consistency. This suggests that the more resource-intensive approach to reliability may be unnecessary. Heldsinger and Humphry (2013) and Jones and Inglis (2015) also reported versions of inter-rater reliability based on correlations between pairwise comparisons between judges. This relies on the unjustified assumption that the scores generated by decisions from an individual judge are reliable.

Given the potentially problematic elements of the literature stemming from those reporting only SSR (Bramley and Vitello, 2019; Jones et al., 2019), and following best practice laid out in the same articles, I present both SSR and split-half reliability in all cases. In a meta-study of 49 comparative judgment studies, Verhavert (2018) found that in general, one requires 12 judgments per script to expect to reach an acceptable threshold, SSR $> 0.7$. With split-half reliability in mind, I aimed to collect 20 judgments per script, although this was not attained in all cases. This is more than enough to evaluate SSR and facilitates inter-rater reliability analysis based on approximately 10 judgments per script. In light of the discussion above, I expected to find that SSR would be greater than split-half reliability in all cases and interpreted my results with this in mind.

### Judge and script misfits

A third possible approach to reliability is a measure of an item's fit to the model (Pollitt, 2012a). For every pairwise comparison, it is possible to deduce a measure of 'surprise' (Pollitt, 2012a, p. 164) inherent in that comparison. The degree of surprise, or fit, is quantified by the difference between the expected and observed values. By considering the surprise inherent in decisions made by a given judge, one can produce a measure of the misfit for that judge.

The role and use of misfit in the literature has been inconsistent, raising questions about its value for education research. I position misfit as related to reliability as it is a measure of the difference between judges and can hence be viewed as a proxy for inter-rater reliability. However, its standard usage, set out by Pollitt (2012a), is one regarding quality control and is hence more closely akin to external validity. While it is reasonable to evaluate the quality of a given dataset by investigating the number of judges (and scripts) behaving unexpectedly, two problems arise when using this measure to consider excluding data. These stem from the tension between misfit as a measure of reliability or validity.

The first is a recursion problem. After excluding the misfit data, one presumably computes a new model and checks for misfit data again. It is likely that new data will now appear as misfitting. This problem can be solved with pre-registered analysis. However, in doing so, the misfit measure becomes a tool to improve the quality of the data, but loses its power to evaluate reliability and validity.

Similarly, it is unclear what researchers should do with responses identified as misfits, but that do not appear qualitatively unusual. In education, comparative judgment is often used on the premise that identifying the quality of scripts is difficult in isolation. A researcher can 'examine' a misfit script and qualitatively consider its place in the dataset, but this subjective approach appears to somewhat undermine the quantitatively driven method.

I argue that misfit is not a productive tool for research purposes and hence do not report these values in the empirical work. Consistent with recent literature (Jones et al., 2015; Hunter and Jones, 2018; Heldsinger and Humphry, 2010; Bisson et al., 2016), I informally explored misfits in each of the empirical studies presented in this thesis. In the absence of substantive findings, and in light of the issues discussed here, these informal explorations are not reported.

This discussion of misfit should be viewed as a peripheral topic and is included here for completeness only. For this reason, the technical details of misfit calculations are presented in Appendix A, alongside a more substantive argument for its exclusion in the empirical chapters of this thesis.

## 3.6 Ethics

Ethical approval for this research was obtained from two sources. For the data presented in Chapter 8, ethics approval was granted by the *Rutgers University, Arts and Sciences Institutional Review Board* (IRB) on February 19, 2018. All other data collection was approved via the Ethical Clearance Checklist submitted to the *Loughborough University (Human Participants) Sub-Committee* on October 18, 2016. This checklist was submitted in accordance with the Generical Protocol established by the Mathematics Education Centre (Ref: G09-P1).

# Chapter 4

# Proof conceptions I: Comparing students and mathematicians

In this chapter, I present the first of two studies on the Conceptions Task, which asks respondents to explain what mathematicians mean by proof. This study has three aims corresponding to the research questions 1, 2a and 3a on page 34: 1) to record the conceptions of proof held by students and mathematicians, 2) to systematically investigate the nature of conceptions most valued by mathematicians as judges, and 3) to examine the reliability and validity of the conceptions scores. This study is reported in four phases, with Phase 3 providing the most substantive contribution.

Phase 1 was a pilot phase involving undergraduate students' conceptions of proof, judged by graduate students of mathematics. This phase provided two insights important for the phases to follow. The first was the reliability of the judgments, confirming that the scores yielded the statistical properties required to justify further analysis. Second, by recruiting graduate-level judges, this contributed to later analysis contrasting the scores resulting from judging cohorts with different academic backgrounds. I also computed a series of correlational analyses, comparing the Conceptions Task scores with established measures of proof comprehension and more general mathematical performance. I investigated the possible relationship between abstract understanding of proof (captured by the Conceptions Task) and concrete understanding based on more direct measures of mathematical performance.

Phase 2 was a repeat of Phase 1, replacing the graduate-student judges with research-active mathematicians. This phase investigated the possibility that graduate students' judgments would result in scores different from those generated by more qualified experts. Phase 2 served as a secondary pilot, justifying the larger investigation presented in Phase 3.

In Phase 3, a new set of research-active mathematicians was recruited, this time asked to act as both respondents and judges. Mathematicians' responses were judged in the same pool as the undergraduate responses from the earlier phases. This led to three separate strands of analysis. First, scores for undergraduates' and mathematicians' responses were compared to ascertain whether the conceptions scores were related to mathematical expertise. Under the assumption that mathematicians should out-perform undergraduates on any task in which mathematical expertise is measured, this is interpreted as evidence for the validity of the Conceptions Task scores as a measure of general mathematical expertise. Second, I present a content analysis, conducted to better understand the nature of the conceptions held by members of both groups, and the differences between them. Finally, by combining the coding-based content analysis with statistical modelling, I identified the types of responses most rewarded by the comparative judgment process. This provided further insight into the conceptions of proof held by mathematicians, as well as indicating content validity based on references back to the theoretical literature on proof itself.

Finally, Phase 4 involved non-expert judges, recruited to explore the divergent validity of the scores produced. Phases 1, 2 and 3 provided insights into students' and mathematicians' conception. However, as with much research conducted using comparative judgment, they leave open the possibility that the responses were rewarded for confounding non-mathematical features such as linguistic skill or grammatical accuracy. Non-expert judges have a limited capacity to make judgments based on mathematical expertise. Hence, by comparing the judgments of non-experts and experts, I generate insight into the importance of subject-specific knowledge in the judging process.

In sum, the four phases of this study combine to provide a detailed understanding of the reliability and validity of evaluating conceptions of proof using this comparative judgment-based approach.

## 4.1 Phase 1: Methods

In this pilot phase, undergraduates' responses to the Conceptions Task were judged by graduate students of mathematics.

### 4.1.1 Materials

The Conceptions Task asked respondents to ‘explain what mathematicians mean by proof in 40 words or fewer’. This appeared as the third of three tasks in a booklet also containing a proof of the uncountability of the unit interval, a multiple-choice Proof Comprehension Test (evaluating students’ understanding of the given proof), and the Summary Task (evaluating students’ ability to summarise the given proof), all of which are discussed in Chapter 6.

### 4.1.2 Participants

One hundred and sixty-one undergraduate students from the same British university participated in this study. Eighteen declined to have their data used for research purposes, leaving a total of 143 participants for analysis. All participants were enrolled in an introductory module on Real Analysis, compulsory for all students majoring in pure mathematics. This module covers fundamental concepts related to sequences, series and epsilon-$N$ definitions in analysis, and is taken by students in their first or second year of study.

### 4.1.3 Procedure

Data collection took place in a week-eight lecture. Participants were given 40 minutes to complete the task booklet, and were advised to dedicate 10 minutes to the Conceptions Task.

All responses were typeset in an identical format to remove the potential influence of handwriting.

### 4.1.4 Comparative judgment

Eleven graduate students performed 142 or 143 judgments each[1], for a total of 1572 with a median of 11.4 seconds per judgment. Each response received between 20 and 22 judgments. Judges were compensated for their time based on an assumed 20 seconds per judgment.

The Conceptions Task scores had a mean 0.00 ($\sigma = 2.01$). The comparative judgment algorithm discussed in Chapter 3 is based on a $z$-score calculation so it is common to find a mean close to but not precisely zero. This standard deviation is in line with previous studies in this thesis, and with others having reported similar statistics (Hunter and Jones, 2018).

[1] Although one judge did not complete the full complement of 143 requested, the shortfall does not substantively impact the analysis.

### 4.1.5 Data analysis

Phase 1 focused on establishing preliminary reliability, evaluated using Scale Separation Reliability (SSR) and inter-rater reliability, both discussed in Section 3.5.2. I then computed a series of statistical analyses, comparing scores from the Conceptions Task with those from a Proof Comprehension Test, students' final scores on the module from which they were recruited, and students' performance on the Summary Task.

## 4.2 Phase 1: Results

### 4.2.1 Example responses

I first include a series of verbatim examples to orient the reader to the types of responses elicited by the Conceptions Task, asking participants to 'explain what mathematicians mean by proof in 40 words or fewer'[2].

**Top five responses**

- A reasoning that shows that a theorem is true or false, using theorems and principles that are already deemed true.
- A proof is a chain of logical implications that starts from axioms or from already-proved results, and show that a new claim is necessarily true if we regard those axioms or those earlier results as true.
- A way of definitively arguing that a theorem (or similar) is correct, in a way that means there are no logical gaps and everyone would agree with its conclusions.
- Proof is when you give evidence to how something is. It is a detailed step by step process to show how you get something with facts that you already know. They are many types of proof by exhaustion, contradiction, counterexample which all gives a result of the claim being true.
- Proof is a logical argument in mathematics which uses previously proven theorems and ideas to build upon and generate new mathematics. Is it there to show whether something is true or not.

[2] This word-limit was not enforced, and was included only as an indicator of the expected length.

**Bottom five responses (excluding five blanks)**

- Using already learnt skills to…
- Definition – needs to be stated and then to prove that this definition is true.
- A proof using 40 words using equations etc. or less.
- Proving things…
- Proving something in less than 40 words.

### 4.2.2 Reliability of conceptions scores

Scale Separation Reliability was found to be acceptable, SSR = .87. Inter-rater reliability gave $r = .74$, based on 100 iterations of Bisson et al.'s (2016) split-half method discussed in Section 3.5.2. This was also deemed acceptable.

Given the absence of a consensus on the nature of proof itself, it is notable that this comparative judgment-based approach elicited apparent consensus on the topic. Interpretations and consequences of this finding are discussed later in this chapter.

### 4.2.3 Module scores

Final module scores ranged from 33% to 97%, with a mean of 56% ($\sigma = 14$). These were based on a weighted aggregate of coursework (25%) and final examination (75%).

### 4.2.4 Proof Comprehension Test

The Proof Comprehension Test was a multiple-choice assessment of students' local understanding of the uncountability proof, featured in the task booklet. This test has 12 items, each with a correct answer and three distracters. Scores ranged from 1 to 12, with an average of 4.2 ($\sigma = 2.4$). These scores were unexpectedly low, indicative of the difficulty of the test for the participants involved. However, performance was significantly above chance ($M = 3$ with four options per item), with $t(133) = 6.35, p < .001$.

The test also yielded low internal consistency, with Cronbach's $\alpha = .53$, indicating potential problems with the test as a meaningful measure of proof comprehension. The statistical properties of these data are explored in more detail in Chapter 6, where the focus is more explicitly on proof comprehension.

For the present purpose of comparing Proof Comprehension Test scores with conceptions scores, I simply note that the relevant correlational analyses should be interpreted with caution. This is not of substantive concern, given it is not the primary motivation of this study.

### 4.2.5 The Summary Task

The Summary Task asked students to summarise the given proof, and was also scored using comparative judgment with mean 0.00 ($\sigma = 1.78$).

### 4.2.6 Criterion analyses

To investigate the extent to which the Conceptions Task aligned with more conventional measures of mathematical performance, I computed a series of Spearman correlations. Conceptions Task scores were not significantly related to the Summary Task scores ($\rho = .38$, $p = .038$), Proof Comprehension Test scores ($\rho = -.09$, $p = .291$) or final module scores ($r = .01$, $p = .882$). The Summary Task did yield a $p$-value below .05, but was considered non-significant under the Holm-Bonferroni correction.

Given that both the Summary and Conceptions Tasks were assessed using comparative judgment, it is likely that the near-significance of this relationship can be explained by the judges having rewarded similar, but non-mathematical properties of participants' responses in both cases (e.g. presentation or linguistic appeal). A similar observation was made by Jones and Inglis (2015) in the context of school students' problem-solving. This is especially likely in the present study, given that the same judges were used for both tasks. I investigate the possibility of judges rewarding non-mathematical features in Phase 4.

Multiple linear regression was used to investigate which of the mathematical measures, if any, best predicted conceptions scores. The model, $F(3, 130) = 2.50$, $p = .064$, was not significant, and the three measures explained only 5.4% of the variance. It has, therefore, been omitted.

## 4.3 Phase 1: Discussion

Phase 1 demonstrated that scores on the Conceptions Task were reliable, indicating that students' conceptions of proof can be measured using this comparative judgment-based approach. The high reliability indicates that, although mathematicians do not agree on the nature of proof (Balacheff, 2008; Weber et al., 2014a), there is at least some consensus on what they want students to say in this comparative judgment-based context. This could be considered surprising given the lack of consensus on proof in the literature and could have noteworthy implications for those wishing to quantify conceptions or beliefs in other realms in which experts do not reach consensus.

Having investigated the relationship between this new measure of proof conceptions and other measures of proof comprehension, I found no evidence for a relationship between conceptions scores and performance on the Proof Comprehension Test, or final module scores. These findings are consistent with the findings of Stylianou et al. (2015), who also found limited evidence for a relationship between students' beliefs about proof and their performance on proof-related tasks.

In Phase 2, research-active mathematicians were recruited as judges in an attempt to understand whether the absence of these relationships was a function of judging expertise.

## 4.4 Phase 2: Methods

In this phase, research-active mathematicians were recruited to judge the proof conceptions collected in Phase 1.

Mathematicians recruited in this phase were also asked to provide their own response to the Conceptions Task, before performing their judgments. These responses are then judged in Phases 3 and 4.

### 4.4.1 Materials

An e-version of the Conceptions Task was produced using OnlineSurveys.com. After participants had consented to the study and given their response to the Conceptions Task, they were redirected to nomoremarking.com to make their judgments.

### 4.4.2 Procedure

Research-active mathematicians were recruited via email and in-person at the completion of two academic presentations at two British universities. Those invited in-person completed a physical copy of the task sheet, extracted from the task booklet in Phase 1, and were emailed a link to the judging platform.

In all cases, judges first responded to the Conceptions Task before completing their judgments. Judges were asked to complete between 20 and 100 judgments. The minimum was given only to encourage judges not to perform a trivial number of judgments and was not enforced. The software was set to allow no more than 100 judgments per judge.

All responses were typeset in an identical format to remove the potential influence of handwriting.

### 4.4.3 Participants

Forty responses were received. Five were excluded as they did not identify as research-active mathematicians, and a further seven completed the Conceptions Task but did not perform any judgments. This left a total of 28 eligible judges, 23 from email recruitment and 5 from in-person invitations.

### 4.4.4 Comparative judgment

The 28 eligible judges completed a total of 1693 judgments[3] on students' responses. Each judge performed between 2 to 100, with a median of 44.5. Each response received between 20 and 27 judgments with a median of 24 and the median time per judgment was 12.3 seconds. Judges were not compensated for their time.

In this phase, the Conceptions Task scores had mean 0.00 ($\sigma = 1.67$).

### 4.4.5 Data analysis

To assess the degree of agreement between research-active mathematicians and the graduate students, I conducted a Pearson correlation comparing scores generated by the two judging cohorts. Statistical analyses, identical to those in Phase 1, were then conducted using the scores generated from mathematicians' judgments.

## 4.5 Phase 2: Results

### 4.5.1 Reliability

When judged by research-active mathematicians, internal consistency was acceptable, SSR = .87, as it was for the graduate students. Inter-rater reliability, $r = .74$, was also acceptable[4].

[3] Two respondents participated in the judging portion of the study but did not complete the Conceptions Task themselves. This has no direct bearing on the current phase but will be relevant in Phase 3.

[4] Reliability measures in Phase 1 and 2 were coincidental to two significant figures. Both measures diverge in the third significant figure.

These values confirm the reliability findings from Phase 1 and, by their similarity, contradict the conjecture that mathematicians would be better judges in this context.

### 4.5.2 Comparing graduate students' and mathematicians' judgments

The correlation between conceptions scores generated by the two judging cohorts, $r = .79, p < .001$, further indicated that both cohorts of judges behaved similarly.

### 4.5.3 Correlational analysis based on mathematicians' judgments

For completeness, I also ran the same correlational analysis from Phase 1 here. No significant relationship was found between conceptions and summary scores ($\rho = .32$, $p = .385$), Proof Comprehension Test ($\rho = .30$, $p = .255$) or module scores ($\rho = .34$, $p = .098$).

## 4.6 Phase 2: Discussion

There was no evidence to suggest that mathematicians behaved differently from graduate students when judging the Conceptions Task. The analysis presented here confirms the conclusion on reliability from Phase 1, providing further evidence that conceptions of proof can be reliably evaluated using a comparative judgment-based approach. Moreover, these findings are also in line with Stylianou et al.'s (2015) conclusions regarding the independence of students' proof conceptions and their mathematical performance.

The next phase constitutes the major contribution of this chapter, considering the validity of the conceptions scores via mixed methods comparisons between students' and mathematicians' performance on the task.

## 4.7 Phase 3: Methods

In Phase 3, mathematicians' responses to the Conceptions Task were judged alongside the undergraduates' responses featured in Phases 1 and 2. Scores for the mathematicians' conceptions were compared with those assigned to the undergraduate responses in an attempt to understand the extent to which mathematical expertise is rewarded in this task. Based on the assumption that mathematicians should out-perform undergraduates on a task requiring mathematical expertise, this comparison is interpreted as an indicator of divergent validity.

This phase also features a content analysis aimed at understanding the differences between mathematicians' and students' responses, as well as providing some insight into the types of responses most valued by the judges.

### 4.7.1 Materials

No new materials were introduced in this phase.

### 4.7.2 Procedure

In this phase, mathematicians were recruited via email only, and were asked to both complete the Conceptions Task, then judge responses from others. Their responses were added to the judging pool during active data collection, resulting in newly collected responses initially receiving fewer judgments than those collected earlier. This allowed data collection to happen more quickly than if responses and judgments had been collected in two stages.

However, this approach did present the problem that responses most recently collected will have received fewer judgments. This was overcome with a two-part solution. First, the software for pairing responses favours, where possible, those having received fewer judgments. Second, an initial deadline for data collection was determined to be three weeks from the beginning of data collection. After this date, participants were sought but responses to the task would not be added to the judging pool. For consistency, judges recruited after the three-week deadline were also asked to complete the Conceptions Task. However, these responses were not used as they could not be judged without recruiting more judges, extending the study ad infinitum.

Fifteen judges were recruited before the three-week deadline and had their responses included in the judging pool[5]. To reach the required 20 judgments per response, data collection continued for a further five weeks, during which time a further 14 judges were recruited.

### 4.7.3 Participants

One hundred and thirty of the original students' responses to the Conceptions Tasks were reused here. Thirteen blank responses were excluded to optimise

[5] These 15 judges completed a small number of judgments on their own responses. However, given the relative size of the dataset, this was not deemed problematic.

the efficiency of the judgments collected. Thirty mathematicians' responses, collected in Phase 2, were also judged here, as well as the 15 responses collected exclusively for this phase. In total, Phase 3 comprised 175 proof conceptions, evaluated by 29 judges.

### 4.7.4 Comparative judgment

In total, 1941 judgments were collected, with each of the 29 judges completing between 11 and 100 judgments. The median number of judgments per judge was 86. Each response received between 20 and 27 judgments, the median number of judgments per response was 22, and the median time spent on each judgment was 10.6 seconds. As in Phase 2, judges were not compensated for their time.

### 4.7.5 Data analysis

After examining reliability, I first compared conceptions scores for mathematicians' and students' responses using a two-sample $t$-test. This was followed by a content analysis consistent with the principles of thematic analysis (Braun and Clarke, 2006). A series of chi-squared tests were conducted to identify code-by-code differences between cohorts. Finally, a regression analysis on codes from the content analysis then explored the specific aspects of responses valued most by judges.

In this phase, Conceptions Task scores had mean 0.00 ($\sigma = 1.67$).

## 4.8 Phase 3: Results

### 4.8.1 Reliability

In this phase, Scale Separation Reliability was SSR = .83, while inter-rater reliability gave $r = .68$, based on 100 iterations of the split-half method discussed in Section 3.5.2. Again, both are deemed acceptable.

### 4.8.2 Comparing mathematicians' and students' Conceptions Task scores

On average, mathematicians received significantly better scores for their conceptions of proof ($N = 45, M = 1.23, \sigma = 1.35$), than undergraduates ($N = 130$, $M = -0.43$, $\sigma = 1.34$) (see Figure 4.1). This difference was significant, $t(88.75) = 7.95$, $p < .001$, with an effect size of $d = 1.33$.

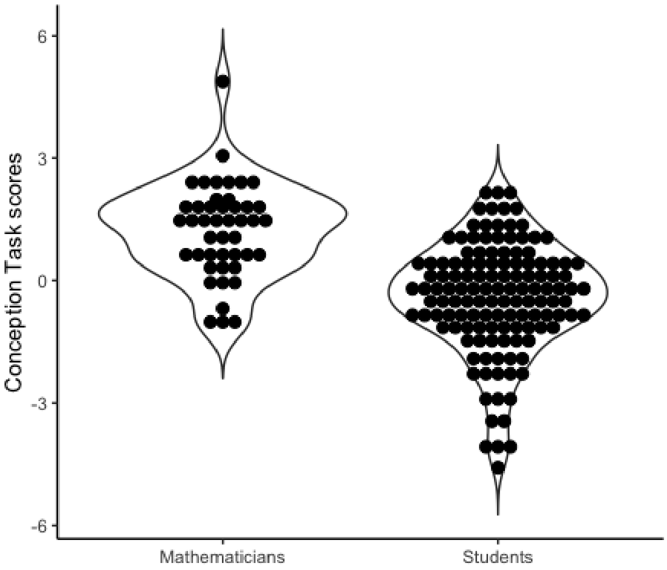


*Figure 4.1. Comparison of scores assigned to the proof conceptions of undergraduates and mathematicians.*

It is also worth noting that the top five responses were all from research-active mathematicians, while the lowest-scoring response from a mathematician was 144th of 175. Here, I show the top five responses from mathematicians:

- A logical derivation of a mathematical statement based on statements that are already known or assumed to be true.
- A proof is a checkable record of reasoning establishing a fact from agreed, more basic assumptions.
- A proof is a step-by-step variable argument, proceeding from some assumptions to a desired conclusion using only previously proved statements or accepted axioms.
- A comprehensive logical argument that a statement is true, based on clearly formulated assumptions and following generally accepted lines of reasoning and level of detail.
- A logically coherent argument establishing the truth of an assertion from a known and agreed base.

The bottom mathematicians' response scored 144th of 175:

- A mathematical proof is like algorithm to solve problems in mathematics. It contains statements that ordered logically depending on definitions and some known theorems.

### 4.8.3 Content analysis

**Developing a coding scheme**

To analyse the content of all 175 responses, I developed a coding scheme with two fellow researchers: a graduate student colleague and an academic supervisor. This analysis facilitates a detailed comparison of the differences between cohorts, as well as providing an understanding of the nature of responses most valued by judges.

In examining 10 undergraduates' proof conceptions, the graduate student researcher and I identified common themes and phrases. This preliminary content analysis yielded eight themes. We then independently applied the existing scheme to 10 further responses, noting possible edits to the list of codes including additions and mergers. Discrepancies were discussed and a new scheme agreed. This process was repeated with a third set of 10 responses, resulting in an 11-code scheme applied to the full dataset.

At this stage, an academic supervisor was brought in to the analysis team, replacing the graduate student who was no longer available. We used the 11-code scheme to analyse the 130 undergraduates' responses. Based on this full analysis, a further three themes were identified. As a result, the 130 responses were coded by both researchers again, this time using the 14-code scheme. Finally, the 45 mathematicians' responses were also coded by both researchers, prompting the addition of one further code. The 130 undergraduate responses were then rechecked for evidence of this 15th code.

Each code was considered as a binary evaluation for a given response, meaning that a response was assigned each code at most once. The resulting 15 codes are shown in Table 4.1, together with their respective frequencies and a comparison of the differences between students' and mathematicians' responses.

To evaluate inter-coder reliability, I examined pooled Cohen's Kappa, $\kappa = 0.79$, indicating acceptable inter-coder reliability. Pooled $\kappa$ between .6 and .8 indicates 'substantial' agreement (De Vries et al., 2008, p. 278). Given the discussions throughout the coding process, this $\kappa$ is probably an over-estimate of true inter-coder reliability. However, 0.79 is high enough to suggest that the reliability of this process is acceptable, even if slightly over-estimated.

**Comparing students' and mathematicians' responses**

In comparing the responses of students and mathematicians, an independent chi-squared test was run on each of the 15 codes, see Table 4.1. There were significant differences between students' and mathematicians' responses in five

*Table 4.1*

*Code scheme for responses to the Conceptions Task.*

| Code | Description | Experts | UGs | $\chi^2$ | $p$ |
|---|---|---|---|---|---|
| Argume-<br>ntation | Reference to an 'argument', 'ch-ain of reasoning' or 'derivation'. | 80% | 21% | 50.90 | <.001* |
| Object | Naming the object to be proved, e.g. 'theorem, statement, result'. | 80% | 82% | 0.05 | .820 |
| Certainty | Reference to 'truth' or 'correctness'. | 44% | 76% | 15.44 | <.001* |
| Established knowledge | Reference to 'agreed assumptions' or 'shared knowledge'. | 38% | 29% | 1.33 | .287 |
| Conviction | Reference to the readers' increased conviction in the statement. | 22% | 2% | 22.39 | <.001* |
| Conditions | Reference to the domain of applicability for a statement. | 20% | 25% | 0.40 | .529 |
| Explanation | Reference to 'how' or 'why' the statement is true. | 16% | 23% | 1.33 | .287 |
| Verification | E.g. 'confirms', 'validates', 'checks', 'justifies', or 'shows'. | 16% | 9% | 1.38 | .240 |
| Axiom | Use of the term 'axiom'. | 13% | 8% | 0.90 | .341 |
| Deconst-<br>ruction | Reference to 'breaking down' the theorem into familiar truths. | 7% | 5% | 0.30 | .749 |
| Discovery | Reference to proving something 'not already known'. | 7% | 9% | 0.28 | .596 |
| Incontr-<br>overtibility | Reference to 'undoubted', 'cannot be argued with'. | 7% | 13% | 1.36 | .244 |
| Empiricism | Reference to empirical evidence. | 4% | 8% | 1.36 | .376 |
| Falsification | Reference to disproving. | 2% | 27% | 12.48 | <.001* |
| Generality | Reference to 'all cases'. | 0% | 18% | 9.37 | .002* |

*Note.* Experts = research-active mathematicians, UGs = undergraduate students. Codes ordered by frequency in expert responses. Significance determined based on the Holm-Bonferroni method with initial $\alpha = .05$.

codes. Mathematicians were significantly more likely to refer to *argumentation* and *conviction*, while students were more likely to refer to *falsification*, *certainty* and *generality*. I first discuss students' emphasis on falsification and generality, as I believe these to be consequences of the immediate educational environment from which they were recruited. I then discuss the more epistemologically interesting codes: *certainty, argumentation* and *conviction.*

The students were from a Real Analysis module with two features that may have promoted their emphasis on *falsification* and *generality*. These features are post hoc justifications for the analysis already presented, and are based on informal discussions with the module leader. First, 'true or false' tasks were a common feature of formative and summative assessment, possibly promoting a connection between proof and falsification. Second, emphasis on quantifiers was a common feature of lectures, probably leading students to refer to generality in their explanations of proof. The emphasis on quantifiers and generality might be similar in other Real Analysis courses, although the UK context meant that this module is taught early in the degree programme, so may be less explicit in contexts where Real Analysis is taught later. The extensive use of 'true or false' tasks is probably an unusual preference of the particular lecturer and is therefore not to be expected in different contexts.

The notion of proof as providing *certainty* was also significantly more frequent in students' responses. While certainty featured in Czocher and Weber's properties of proof via the notion of truth, other authors have contested the Platonic notion of pairing certainty (or truth) and mathematics (Marcus and McEvoy, 2016). Certainty can also be viewed as more consistent with the day-to-day experience of students via the definition-theorem-proof structure of much undergraduate mathematics education (Moore, 1994), where proofs are often presented as bearing authority (Harel and Sowder, 1998). While showing a significant difference between students' and mathematicians' responses, it is worth noting that certainty was also the third most frequently applied code for the mathematicians' responses. This suggests that although the majority of mathematicians do not prioritise this conception, certainty is not necessarily an indicator of a lack of sophistication. As discussed later, it is unclear how much weight should be placed on this conclusion given that a non-trivial number of mathematicians can be argued to have a relatively poor understanding of the philosophy of mathematics.

On the other hand, the notions that proofs involve *argumentation* and provide *conviction*, both more common in mathematicians' responses, suggest a socially constructed view of mathematics in which proofs are written for an audience. This is consistent with the writing on proof from both Aberdein (2009) and Czocher and Weber (in press). Aberdein (2009) claimed that the majority of mathematics is not written in formal logic and that the majority of mathematical activity is best understood as a 'species of argument' (p. 1). Similarly, Czocher and Weber's list of properties contributing to their cluster definition began with proof as a 'convincing justification that will remove all doubt that a theorem is true for a knowledgeable mathematician' (p. 20).

### Features most valued by mathematician judges

Having compared the responses of students and mathematicians, I now focus on the content most rewarded by the mathematician judges. Here, I present a series of quantitative analyses, starting with the Spearman rank-order correlations between each code and the conceptions scores. This is followed by a regression analysis predicting conceptions scores using all 15 codes. While the ratio of predictor variables to data-points is beyond the bound recommended by Field et al. (2012), this is the only available regression modelling approach given the absence of robust theoretical reasons to include one code over another.

Table 4.2 shows the correlation coefficients comparing conceptions scores with each code, followed by the forced entry regression model. The regression model, $F(15, 159) = 6.46$, $p < .001$, $R^2 = .38$, explains 38% of the variance.

Argumentation was identified as the most important code, both in the independent correlational analyses and in the forced-entry regression. Proof as argumentation yielded a significant relationship, $\rho = 0.48, p < .001$ with conceptions scores, and was responsible for 23% of the variance in the conceptions scores. This is consistent with the chi-squared analysis above, confirming that argumentation was the most important aspect of proof to the mathematicians, both as judges and in their responses to the task.

Other significant codes included *object*, *established knowledge* and *incontrovertibility.* The object code appears to reflect the clarity of responses, rather than something epistemologically meaningful. Given that more than 80% of all responses featured a reference to the object of interest, it is likely that those did not suffer from an absence of clarity or specificity. Established knowledge and incontrovertibility are both consistent with the cluster conception of proof promoted by Czocher and Weber (in press). Although not featuring heavily in the mathematicians' explicitly stated conceptions, it is unsurprising that mathematicians would deem such conceptions important when presented.

## 4.9 Phase 3: Discussion

This phase featured mathematicians' proof conceptions being judged alongside the original undergraduate responses. Two substantive findings were presented. First, mathematicians significantly outperformed undergraduates on the Conceptions Task. This is evidence of a relationship between proof conceptions and mathematical expertise and, by extension, evidence for the convergent validity

*Table 4.2*

*Regression modelling of Conceptions Task scores.*

| Code | Coefficients | | Regression model | | | |
|---|---|---|---|---|---|---|
| | $\rho$ | $p$ | $B$ | SE | $\beta$ | $p$ |
| Argumentation | 0.48 | <.001* | 1.47 | 0.21 | 7.07 | <.001* |
| Object | 0.34 | .059 | 0.64 | 0.25 | 2.53 | .013* |
| Certainty | 0.00 | .981 | 0.38 | 0.23 | 0.78 | .436 |
| Established knowledge | 0.21 | .005 | 0.57 | 0.23 | 2.47 | .015* |
| Conviction | 0.34 | .073 | 0.43 | 0.41 | 1.06 | .292 |
| Conditions | 0.00 | .959 | 0.20 | 0.23 | 0.88 | .380 |
| Explanation | -0.31 | .341 | -0.35 | 0.24 | -1.48 | .340 |
| Verification | 0.30 | .378 | 0.39 | 0.31 | 1.24 | .218 |
| Axiom | 0.21 | .006 | 0.49 | 0.35 | 1.40 | .364 |
| Deconstruction | -0.02 | .812 | 0.07 | 0.41 | 0.38 | .856 |
| Discovery | 0.32 | .323 | 0.49 | 0.35 | 1.41 | .362 |
| Incontrovertibility | 0.37 | .095 | 0.63 | 0.30 | 2.31 | .036* |
| Empiricism | -0.30 | .207 | -0.39 | 0.36 | -1.07 | .285 |
| Falsification | -0.02 | .788 | 0.00 | 0.25 | 0.01 | .993 |
| Generality | -0.32 | .315 | -0.05 | 0.29 | -0.39 | .850 |

*Note.* Forced-entry multiple regression model predicting Conceptions Task scores with coded content analysis. The 15-code model, $F(15, 159) = 6.46$, $p < .001$, explains 38% of the variance. Significance was determined using the Holm-Bonferroni method with initial $\alpha = .05$.

of the resulting scores.

After finding a significant difference between groups, a content analysis showed that mathematicians prioritised argumentation and conviction while students focused on falsification, certainty and generality. This was consistent with aspects of the literature on mathematicians' and students' conceptions of proof, and further indicates validity in the sense that the conceptions scores captured these differences.

On the other hand, it remains possible that the above findings were, to some extent, functions of non-mathematical features. In particular, it is possible that our mathematician judges rewarded the conceptions of their peers based on their familiarity in content and language choice. Given that mathematicians are assumed to be experts on this topic, I viewed this possibility as a necessary limitation of this phase, as designed.

This warrants the fourth and final phase of this study, exploring the role of non-mathematical features using non-expert judges.

## 4.10 Phase 4: Methods

This phase featured non-experts judging the same responses from Phase 3, borrowing a method from Jones and Alcock (2014). Judges without mathematical training were assumed not to make judgments based on mathematical content knowledge, but rather on non-mathematical features such as grammar, syntax and readability. By comparing the resulting scores with those from the mathematicians' judgments, I generate an understanding of the role of mathematical features in the judging process.

### 4.10.1 Materials

No new materials were introduced in this phase.

### 4.10.2 Procedure

Judges were contacted via email using contacts from previous research and were invited to participate through a link to OnlineSurveys.com, as in Phases 2 and 3.

### 4.10.3 Participants

Ten non-expert judges were recruited to perform the necessary judgments in this study. Eight were post-graduate students recruited from the same English university. The remaining two were working professionals deemed expert in the English language. All judges in this study were deemed non-expert in mathematics having not completed a mathematics qualification beyond GSCE mathematics or an international equivalent (year 11). The inclusion criteria were selected to ensure no judge had any formal educational exposure to mathematical proof.

### 4.10.4 Comparative judgment

Each of the 10 judges performed between 172 and 175 judgments, resulting in a total of 1740 judgments. Each response received between 20 and 23 judgments. The median time per judgment was 14.9 seconds. Judges were compensated for their time, based on an assumed rate of 20 seconds per judgment.

In this phase, Conceptions Task scores had mean 0.00 ($\sigma = 0.93$).

### 4.10.5 Data analysis

Reliability was investigated using SSR and inter-rater reliability. I then compared the conceptions scores generated by the two judging cohorts using a Pearson correlation. Finally, I conducted a two-sample *t*-test comparing students' and mathematicians' conceptions scores, based on the non-experts' judgments. Perceived differences between mathematicians and students were then compared across judging cohorts.

## 4.11 Phase 4: Results

### 4.11.1 Reliability

Internal consistency, SSR = .66, was lower than in previous phases but still considered acceptable. On the other hand, inter-rater reliability was low, $r =$ .39, based on 100 iterations of the split-half method discussed in Section 3.5.2.

This analysis indicates that the non-expert judges did not judge the responses reliably.

### 4.11.2 Comparing outputs between expert and non-expert judges

When comparing the two models (scores assigned to each response by the two judging pools), I found a correlation, $r = .54, p = .007$ (see Figure 4.2). This correlation coefficient is lower than that reported by Jones and Alcock (2014) in the context of mathematicians and non-experts judging first-year calculus work, $r(168) = .64$, although the difference is not significant, $Z = -1.41, p = .359$. Moreover, $r = .54$ is noticeably lower than the inter-rater reliability for expert judgments, $r = .68$; this difference is significant, $Z = 2.06, p = .040$, further indicating that non-expert judgments were less reliable than mathematicians.

### 4.11.3 Comparing mathematicians' and students' Conceptions Task scores

As in Section 4.8.2, based on experts' judgments, here I compare the scores assigned, using non-experts' judgments, to mathematicians' and students' responses to the Conceptions Task.

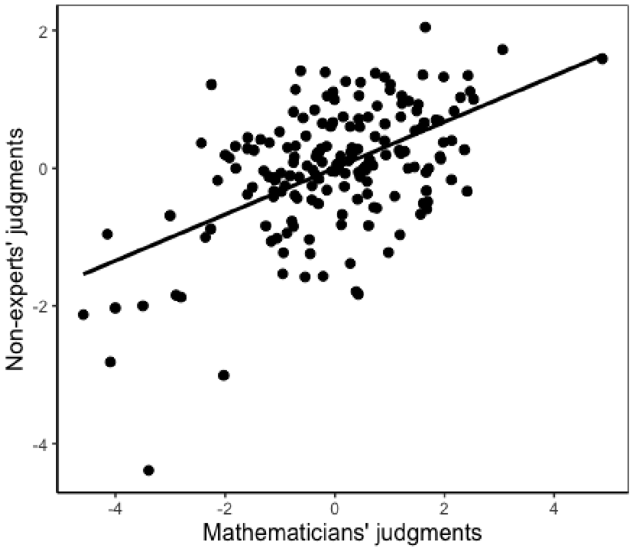


*Figure 4.2. Scatter plot comparing conceptions scores generated by non-experts and research-active mathematicians for the proof conceptions given by undergraduates and mathematicians.*

Mathematicians were again assigned significantly higher scores ($N = 45, M = 0.23, \sigma = 0.76$), than undergraduates ($N = 130, M = -0.08, \sigma = 0.97$). The difference was significant, $t(97.58) = 2.36, p = .029$, representing an effect size of $d = 0.35$. However, this effect size is noticably smaller than the $d = 1.33$ found when using mathematicians' judgments.

## 4.12 Phase 4: Discussion

The judges recruited for this phase were non-experts in mathematics. As such, I assumed that their judgments would be based on non-mathematical aspects of the proof conceptions.

Non-experts were, as expected, worse than mathematicians at judging the relative quality of proof conceptions. Concerning reliability, this is most strongly indicated by the low inter-rater reliability. This apparent lack of agreement demonstrates a higher degree of randomness in the non-experts' judgments. Further, the mathematicians' judgments captured a larger difference between mathematicians' and undergraduates' conceptions scores. Accepting that such a difference should exist, this comparison serves as evidence that the non-experts were poorer judges in this context.

On the other hand, non-expert Scale Separation Reliability was acceptable, and a significant difference between mathematicians and students was shown by their judgments. This probably reflects the notion that non-mathematical features are also relevant to the decision-making process and that, as should be expected from such a qualitative task, a combination of mathematical and non-mathematical factors influence judgments (Jones and Inglis, 2015).

## 4.13 Discussion

Here, I address each of the three research questions highlighted at the beginning of this chapter, before briefly addressing the implications of these findings.

### 4.13.1 Research question 1: What do students and mathematicians write when explicitly asked about their conceptions of proof?

From the content analysis of responses to the Conceptions Task (Table 4.1), I conclude that certainty was most central to students' understanding of proof, referenced in 72% of students' responses. Other frequently highlighted features include references to established knowledge (29%), falsification (27%), the domain of applicability of the relevant theorem (coded under *conditions*, 25%), and the notion of proof as providing explanation (23%).

On the other hand, mathematicians responses most frequently referenced argumentation (80%). Other important features included established knowledge (38%), conviction (22%) and the domain of applicability (20%).

### 4.13.2 Research question 2a: What do mathematicians most value when evaluating the written proof conceptions of others?

In identifying responses most important to the mathematician judges, statistical modelling showed that mathematicians rewarded summaries referencing argumentation (Table 4.2). Other codes identified as significant predictors of Conceptions Task scores included references to the object of the proof, established knowledge and incontrovertibility.

### 4.13.3 Research Question 3a: Do written proof conceptions, scored using comparative judgment, generate a reliable and valid output?

#### Reliability

In all phases involving expert judges, reliability was found to be acceptable. This provides initial evidence that, although mathematicians tend not to agree on many epistemological aspects of proof, there is at least some meaningful consensus when it comes to evaluating the written conceptions of others. I return to this topic in the following chapter, after further evidence has been presented.

#### Validity

This chapter features three distinct pieces of evidence suggesting that comparative judgment-based scores produce meaningful estimates for the quality of proof conceptions.

First, mathematicians out-performed undergraduates, suggesting that mathematical expertise is related to performance on the Conceptions Task. I also found qualitative evidence suggesting that this difference was the result of content-based differences consistent with aspects of the literature on students' and mathematicians' experiences with proof. Finally, by recruiting non-experts to repeat the judging process, I found further validity evidence for the task in the form of the poor performance of non-expert judges, suggesting that the original judgments were probably based on inherently mathematical observations.

On the other hand, I found no evidence suggesting that quality of proof conceptions were related to performance on the proof comprehension tasks available. I conclude from this that understanding of proof is a multi-faceted endeavour and that understanding of the nature of proof (as captured by the Conceptions Task) may be quantitatively distinct from other proof comprehension-related activities such as reading and constructing specific proofs.

### 4.13.4 Implications

The findings in this chapter open new avenues for how one may quantify individuals' conceptions of mathematical entities. In this work, I focused on written conceptions of proof, identifying those most valued by mathematician judges. By investigating the reliability and validity of this comparative judgment-based approach, I offer an important new understanding of this comparative judgment-based approach and its utility in contexts beyond the direct assessment of students' work. While the particular focus in this work was on proof, it seems this approach has utility in understanding other aspects of personal epistemology and their impact on behaviour or performance.

I acknowledge that, in isolation, this approach to quantifying subjective responses misses large amounts of the nuance inherent in research on conceptions and beliefs. However, it seems there is a place for such an approach in this area alongside other approaches that might better capture the nuance but are less amenable to quantification. While the nature of this work is inherently exploratory, I believe I have provided meaningful evidence for the reliability and validity of this particular comparative judgment-based approach, centred on the Conceptions Task.

### Next chapter

This study focused on a first implementation of the Conceptions Task at one British university. The following chapter reports a longitudinal study evaluating students' proof conceptions using comparative judgment. Data were collected at either end of an undergraduate mathematics modules at two universities in the United States, providing an insight into the predictive validity of the Conceptions Task. The next chapter concludes with summative remarks considering the two conceptions-focused studies in tandem.

# Chapter 5

# Proof conceptions II: A longitudinal study

This chapter presents the second of two studies focused on individuals' conceptions of proof. This study uses a repeated measures design to evaluate students' conceptions of proof at the beginning and end of an Introduction to Proof module for undergraduate mathematics students. In Chapter 4, I reported that mathematicians outperformed undergraduates on the Conceptions Task, and that there was a qualitative difference in the content of their responses. In this study, I focus exclusively on students, attempting to capture the development of their conceptions over time. This study provides further evidence suggesting that conceptions of proof can be meaningfully evaluated using this comparative judgment-based approach.

This study is based on data collected in collaboration with Dr Kristen Lew. Dr Lew collected the students' responses from two US universities in 2015/16. As part of my doctoral research, I collected judgment data and independently conducted all analysis presented in this chapter.

## 5.1 Methods

### 5.1.1 Materials

The data presented in this study are based on responses to two differently formatted versions of the Conceptions Task. These differences are the result of my opportunistic analysis of an existing dataset. The consequences of these imperfections are discussed throughout this section.

In week one of the module, students were given the following prompt:
*a) What do you think mathematical proof is?*
*b) What are the most important attributes of a mathematical proof? (Identify/describe 2-4 important attributes)*

In week 15, students received a different, albeit similar prompt:
*What do you think mathematical proof is? What are the most important attributes of a mathematical proof? (Identify/describe 2-4 important attributes)*

These prompts were considered similar enough to be treated as the same task in this study. However, several steps were necessary to mitigate any difference in the appearance of students' responses and are discussed below.

### 5.1.2 Procedure

At both universities, data collection took place in the final 10 minutes of a standard lecture in weeks one and 15. All responses were then prepared and uploaded to nomoremarking.com to be judged.

The hand-written responses were scanned and edited to show only the prompt from week 15 (see Figures 5.1a and 5.1b). Any response structured with 'a)' and 'b)' was edited to appear to be written in bullet-point form. These edits were made to minimise the impact on judges' behaviour. In particular, it was necessary to minimise the possibility that judges would identify the two-cohort structure of the responses and, consciously or otherwise, develop group-based prejudices impacting their decision-making.

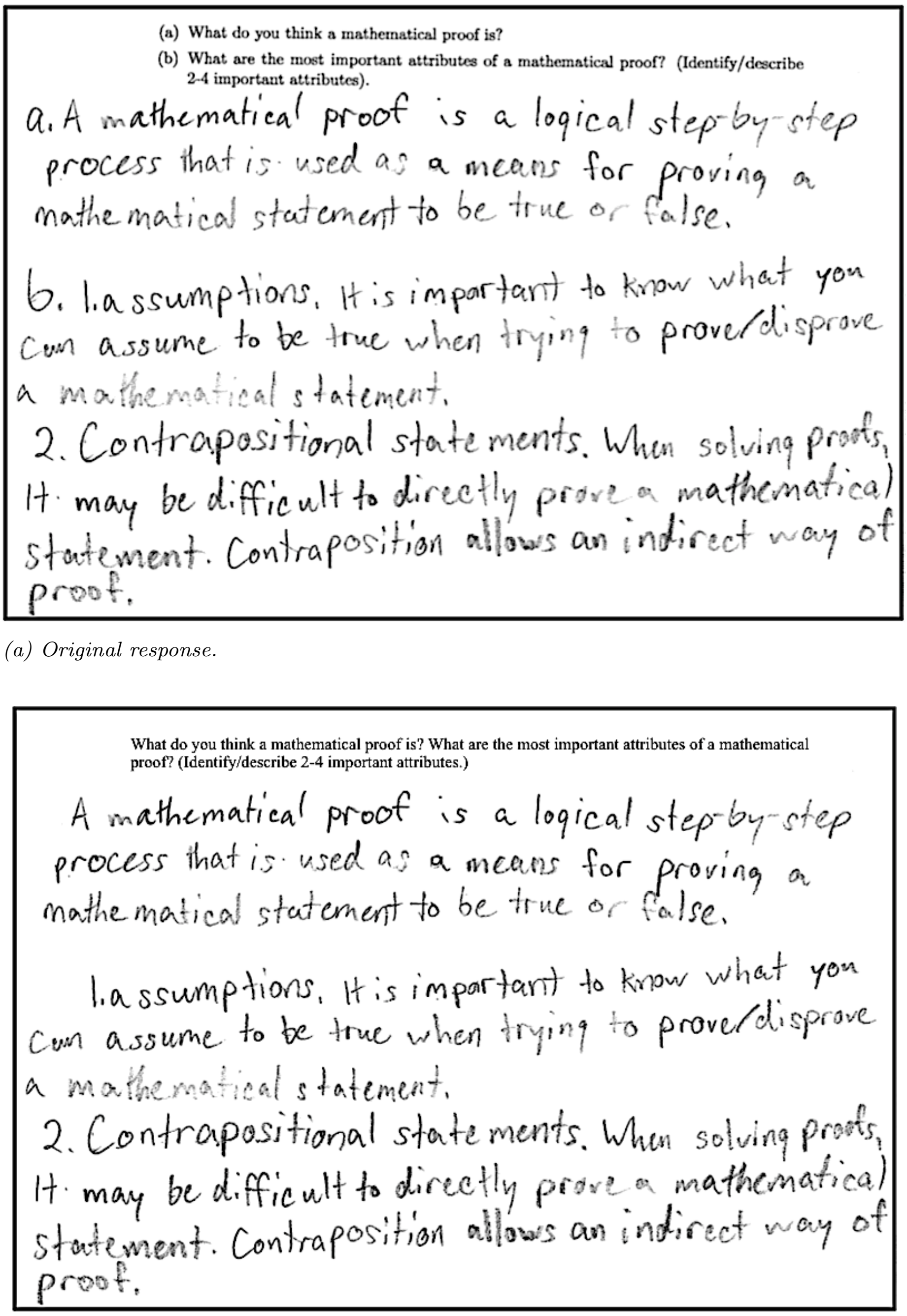
### 5.1.3 Participants

Forty-two students from two US universities participated in this study; 26 from University A and 16 from University B. At their respective institutions, all students were enrolled in the same section of the Introduction to Proof module.

### 5.1.4 Comparative judgment

Fifteen PhD students of mathematics from two different US universities judged the students' responses. Recruitment was conducted via email using professional contacts. Participation was strictly voluntary, with judges invited to complete up to 100 judgments each. No compensation was offered. In total, the 15 judges completed 870 judgments with each script receiving between 20 and 23 judgments. The median time taken was 24.4 seconds for each judgment.

The Conceptions Task scores had mean 0.00 ($\sigma = 1.67$).



(a) What do you think a mathematical proof is?

(b) What are the most important attributes of a mathematical proof? (Identify/describe 2-4 important attributes).

a. A mathematical proof is a logical step-by-step process that is used as a means for proving a mathematical statement to be true or false.

b. 1. assumptions. It is important to know what you can assume to be true when trying to prove/disprove a mathematical statement.

2. Contrapositional statements. When solving proofs, It may be difficult to directly prove a mathematical statement. Contraposition allows an indirect way of proof.

*(a) Original response.*

What do you think a mathematical proof is? What are the most important attributes of a mathematical proof? (Identify/describe 2-4 important attributes.)

A mathematical proof is a logical step-by-step process that is used as a means for proving a mathematical statement to be true or false.

1. assumptions. It is important to know what you can assume to be true when trying to prove/disprove a mathematical statement.

2. Contrapositional statements. When solving proofs, It may be difficult to directly prove a mathematical statement. Contraposition allows an indirect way of proof.

*(b) Edited response.*

*Figure 5.1. Example Conceptions Task responses, demonstrating the edits necessary to produce an apparently homogeneous dataset.*

### 5.1.5 Data analysis

Scale Separation and inter-rater reliabilities were first examined, as discussed in Section 3.5.2. I then conducted two $t$-tests examining differences across the two universities. I then computed a paired-samples $t$-test to investigate the key hypothesis of this study regarding the capacity of the Conceptions Tasks scores to detect the expected improvement over time. Finally, a content analysis was conducted in a manner consistent with the principles of thematic analysis (Braun and Clarke, 2006). As in the previous chapter, this content analysis aimed to generate a more holistic understanding of the nature of students' responses and the differences between responses from weeks one and 15. Finally, results from the present study were compared with the analysis from Chapter 4, to develop an understanding of the similarities and differences across contexts.

## 5.2 Results

### 5.2.1 Reliability

Internal consistency, measured using Scale Separation Reliability, was SSR = .85, while inter-rater reliability was $r = .69$ based on 100 iterations of Bisson et al.'s (2016) split-half method. Both measures were deemed acceptable.

### 5.2.2 Comparing performance between universities

To check for systematic differences between scores from students at different universities, I conducted two independent-samples $t$-tests.

First, I considered only the week one responses. No significant difference was found between responses from University A ($N = 16, M = -0.98, \sigma = 1.41$) and University B ($N = 26, M = -0.57, \sigma = 1.45$), $t(40) = 0.89, p = .381$.

I then compared responses from week 15. Again, no significant difference was found between responses from University A ($N = 16, M = 1.35, \sigma = 1.38$) and University B ($N = 26, M = 0.34, \sigma = 1.60$), $t(40) = -2.09, p = .043$.

While the difference in week 15 responses was not significant under a Holm-Bonferroni correction, this test yielded a $p$-value below .05, providing at least some evidence that there were important differences in Conceptions Task scores attributable to the university from which they were recruited.

To further investigate the necessity for a multi-leveled approached, I computed a chi-squared goodness of fit test, comparing the linear model allowing intercepts to vary by university, with the model given no hierarchy (Field et al., 2012). There was no significant difference between the models, $\chi^2(1) = 0.00, p > .999$. Hence, I conclude that there is no need to consider a multi-leveled approach.

Given the absence of evidence to the contrary, the proceeding analysis treats all participants as belonging to the same homogeneous population.

### 5.2.3 Predictive validity analysis

To evaluate the expected change in students' conceptions of proof, I conducted a paired-samples $t$-test to examine the difference between their responses to the Conceptions Task at the beginning and end of their respective modules. On average, responses from the final week received better scores ($N = 42, M = .73, \sigma = 1.58$), than responses from week one ($N = 42, M = -.73, \sigma = 1.43$). This difference was significant, $t(41) = 5.80, p < .001$, representing an effect size of $d = 0.96$ (see Figure 5.2).

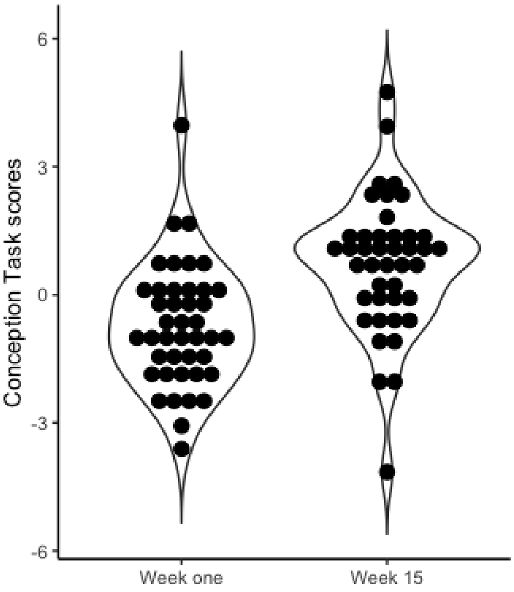


*Figure 5.2. Dot plot comparing Conceptions Task scores from responses collected in weeks one and 15.*

This is substantive evidence that the comparative judgment-based conceptions scores reflected the expected change in students' conceptions across the module, and is also evidence of the predictive validity of the Conceptions Task as a meaningful measure of philosophical awareness in this context.

### 5.2.4 Content analysis

Following Chapter 4, I present here a content analysis investigating the nature of students' responses, attempting to develop a more in-depth understanding of the differences between responses from the beginning and end of the module. The coding in this analysis was conducted by myself and the same academic advisor involved in earlier analysis.

#### Adapting a previous scheme

We first applied the existing 15-code scheme from the previous chapter to the current set of 84 responses. In doing so, we found that a further four codes were required to adequately capture the dataset at hand: *clarity*, *format*, *procedure*, and *audience*. Each response was then checked for evidence of the new codes.

To examine inter-coder reliability for the two coders, I calculated pooled Cohen's Kappa, $\kappa = 0.69$, with 91.2% code-by-code agreement. As discussed in the previous chapter, this is probably an over-estimate of true inter-coder reliability, given the volume of discussion between coders during the coding process. According to De Vries et al. (2008), anything above .6 is indicative of substantive agreement, so inter-coder reliability was deemed acceptable, despite this potential over-estimate. The full 19-code scheme used in this analysis can be found in Table 5.1.

#### Students' conceptions of proof

From the raw frequencies, most students referenced the 'object to be proved' in both weeks one and 15. This is not epistemologically interesting and showed no statistical relationships of note. Hence, the object code is not discussed further. The next most frequently assigned code was that referring to certainty, present in more students' responses than any other aspect of proof in weeks 1 and 15. In contrast to the previous study, we also see that many students referred to argumentation at both ends of the module.

#### Identifying changes in students' responses

Earlier I observed that students' performance improved over time. I now consider the content-specific changes underlying this change in Conceptions Task scores. It is interesting to note that the prevalence of 16 out of 19 codes increased from the first week to the last (although only two reached statistical significance under the Holm-Bonferroni correction). While a testing effect probably has a role to play in this, I also conjecture this to be a function of students

*Table 5.1*

*Revised code scheme for Conceptions Task responses.*

| Code | Description | Week 1 | Week 15 | $\chi^2$ | $p$ |
|---|---|---|---|---|---|
| Argume-ntation | Reference to an 'argument', 'ch-ain of reasoning' or 'derivation'. | 31% | 52% | 3.97 | .046 |
| Object | Naming the object to be proved, e.g. 'theorem, statement, result'. | 60% | 79% | 3.56 | .059 |
| Certainty | Reference to 'truth' or 'correctness'. | 38% | 67% | 6.87 | .009 |
| Established knowledge | Reference to 'agreed assumptions' or 'shared knowledge'. | 24% | 33% | 0.93 | .334 |
| Conviction | Reference to the readers' increased conviction in the statement. | 0% | 2% | 1.01 | .314 |
| Conditions | Reference to the domain of applicability for a statement. | 5% | 33% | 11.32 | <.001* |
| Explanation | Reference to 'how' or 'why' the statement is true. | 21% | 29% | 0.57 | .450 |
| Verification | E.g. 'confirms', 'validates', 'checks', 'justifies', or 'shows'. | 2% | 14% | 0.00 | .999 |
| Axiom | Use of the term 'axiom'. | 2% | 0% | 1.01 | .314 |
| Deconst-ruction | Reference to 'breaking down' the theorem into familiar truths. | 12% | 12% | 0.00 | .999 |
| Discovery | Reference to proving something 'not already known'. | 19% | 7% | 2.62 | .306 |
| Incontr-overtibility | Reference to 'undoubted', 'cannot be argued with'. | 12% | 14% | 0.30 | .746 |
| Empiricism | Reference to empirical evidence. | 7% | 10% | 0.36 | .693 |
| Falsification | Reference to disproving. | 17% | 36% | 3.94 | .047 |
| Generality | Reference to 'all cases'. | 2% | 14% | 3.90 | .048 |
| Clarity | Reference to clarity, brevity or concision | 5% | 38% | 13.86 | <.001* |
| Format | Reference to 'Beginning, middle,. end' or 'box/QED'. | 12% | 36% | 6.56 | .010 |
| Procedure | Reference to method', 'algorithm' or 'process'. | 24% | 17% | 0.66 | .415 |
| Audience | Reference to audience or reader. E.g. 'easily understood'. | 10% | 30% | 4.94 | .026 |

*Note.* Significance indicators are based on the Holm-Bonferroni method with $\alpha = .05$. The final four codes, added for this analysis, did not appear in the previous chapter.

having more confidence with proof at the end of the module, and hence being more willing to share their ideas about proof. Moreover, the largest decrease in frequency came in references to proof as procedures, consistent with the notion that students' conceptions of proof have improved over time. However, this difference was not significant.

This chi-squared analysis, shown in Table 5.1, identified two codes with a significant difference in prevalence from week one to week 15: *conditions* and *clarity*. Both of these changes can be at least partially explained from informal discussions with the module's lecturer. Assessment in this module was driven by many 'state and prove' questions, in which the students were required to precisely state a definition and/or theorem, before providing a proof. In these introductory modules on proof, the lecturer reported that students would frequently misstate given theorems or definitions, neglecting to include details such as the domain of applicability. These errors led the lecturer to consistently emphasise the role of *conditions* in her feedback to students; a feature of her teaching reflected in students' responses to the Conceptions Task. In our informal discussions before this content analysis was conducted, nothing explaining the *clarity* code was discussed. However, in subsequent discussions following this chi-squared analysis, the lecturer recalled discussing notions of elegance with students in their proving practices. She reported having seen many surprising and indirect proof attempts, leading to in-class discussions focused on simplicity and elegance in proving theorems. I believe that it was these discussions that led to the significant increase in students referring to *clarity* in their responses to the Conceptions Task.

### Identifying mathematician judges' priorities

Here, I first report Spearman's rank-order correlations between each code and scores on the Conceptions Task. This is followed by a regression model predicting scores using codes identified as significant in the correlational analysis, see Table 5.2.

All three codes were significant contributors to the 35% variance explained by this model, $F(3, 80) = 14.46, p < .001$. *Argumentation* was also identified as significant in Chapter 4 where it was argued to be consistent with the philosophy of mathematical practice literature, as discussed in Aberdein (2009). This finding provides further support to the claim the argumentation is an important aspect of proof. *Certainty* was not found to be significant in Chapter 4 and was argued to be consistent with more naive student experience-focused conceptions of proof. As such, it is surprising to find it rewarded in the current study.

*Table 5.2*

*Regression modelling of Conceptions Task scores.*

| Code | Coefficients | | Regression model | | | |
|---|---|---|---|---|---|---|
| | $\rho$ | $p$ | $B$ | SE | $\beta$ | $p$ |
| Argumentation | 0.34 | .001* | 1.08 | 0.30 | 3.56 | <.001* |
| Object | 0.29 | .008 | | | | |
| Certainty | 0.37 | <.001* | 1.03 | 0.31 | 3.37 | .001* |
| Established knowledge | 0.36 | .354 | | | | |
| Conviction | 0.26 | .016 | | | | |
| Conditions | 0.39 | .081 | | | | |
| Explanation | 0.25 | .019 | | | | |
| Verification | 0.06 | .567 | | | | |
| Axiom | 0.04 | .727 | | | | |
| Deconstruction | -0.32 | .259 | | | | |
| Discovery | -0.16 | .346 | | | | |
| Incontrovertibility | 0.35 | .386 | | | | |
| Empiricism | -0.04 | .748 | | | | |
| Falsification | 0.35 | .372 | | | | |
| Generality | 0.38 | .302 | | | | |
| Clarity | 0.40 | <.001* | 1.26 | 0.37 | 3.38 | .001* |
| Format | 0.20 | .072 | | | | |
| Procedure | -0.01 | .929 | | | | |
| Audience | 0.25 | .021 | | | | |

*Note.* Codes with a significant Spearman coefficient were entered into a forced entry regression model predicting Conceptions Task scores. Significance was determined using the Holm-Bonferroni method with initial $\alpha = .05$. The resulting three-code model, $F(3, 80) = 14.46, p < .001$, explains 35% of the variance.

Finally, I note the absence of the *object* and *established knowledge* codes in the final regression presented here. Both codes were identified as significant predictors in Chapter 4. Along with the *certainty* code that was also significant in one out of two studies, these findings begin to suggest that the nature of the themes rewarded by this comparative judgment-based evaluation are dependent on the educational context. I discuss this further in Chapter 10.

## 5.3 Summary of longitudinal study results

Both measures of reliability gave acceptable results in the present study, providing further evidence that although mathematicians may not agree on proof itself, there is substantial agreement on what they want others to say about it. The primary focus of the present study was an analysis of the predictive validity of the Conceptions Task. In particular, I investigated the capacity of the scores to reflect the expected improvement in students' conceptions of proof over the course of an introduction to Proof module. Regarding predictive validity, the primary thrust of this study, I found a significant difference with an effective size of $d = .96$, representing evidence for the predictive validity of the scores in their capacity to reflect the expected improvement. Regarding the content of students' responses and their judgments, I found that students tended to focus on proof as arguments that provide certainty, while the judging mathematicians tended to reward responses containing references to argumentation, certainty and clarity. This is partially consistent with the findings from the previous chapter. Coupled with the alignment between these findings and the theoretical literature on proof, this provides further evidence that the Conceptions Task scores are meaningful, valid estimates of the quality of individuals' proof conceptions.

## 5.4 Discussion of research on the Conceptions Task

Chapters 4 and 5 presented two studies focused on understanding students' and mathematicians' conceptions of proof. Here, I summarise the findings related to each of three relevant research questions, before considering the theoretical and methodological implications of this work.

### 5.4.1 Research question 1: What do students and mathematicians write when explicitly asked about their conceptions of proof?

To answer research question 1, I draw on two types of evidence, each present in both studies. The first form of evidence is from content analyses of students' and mathematicians' written responses to the Conceptions Task. The second, applicable only to the mathematicians, is the identification of judges' priorities through statistical modelling with content-based codes as predictors of conceptions scores.

From the students' perspective, ignoring the epistemologically uninteresting *object* code, the most important aspect of proof was certainty, indicating a philosophical naivety which might be considered consistent with their apprentice status in the mathematics community. This was consistent across both studies, with at least two-thirds of both student cohorts referencing certainty in their responses. While further research is necessary to determine the generalisability of these findings, I conjecture that certainty is likely to be identified as an important aspect of proof by most undergraduate students of mathematics and that such a result would be reproducible in most similar research settings.

Other features prominent in students' responses included falsification, generality and the scope of the theorem at stake. Each of these features was identified as important in only one of the two studies, suggesting that they are idiosyncratic functions of the educational environment, rather than generalisable features that one should expect from any undergraduate student in response to the Conceptions Task.

From the mathematicians' perspective, the most important aspect of proof was argumentation. This was reflected both in the frequency of appeals to argumentation in mathematicians' responses, and in the regression modelling from both studies. Other characteristics of proof valued by mathematicians included certainty, incontrovertibility and appeals to established knowledge. These findings are consistent with the literature on proof, providing empirical backing to the theoretically driven writings of Aberdein (2009), on proof as argumentation, and Weber and Czocher (2019,) on proof as a cluster concept.

These findings have implications for the validity of Conceptions Task scores and for our understanding of proof itself. Both are discussed below, in answers to research questions 2a and 3a.

### 5.4.2 Research question 2a: On mathematicians' priorities in evaluating the conceptions of others

As discussed above, mathematicians most heavily rewarded responses containing reference to argumentation. This was reflected in the statistical modelling in both Chapters 4 and 5. Other important features included references to established knowledge, incontrovertibility, certainty and clarity, each with varying strengths of evidence and each featuring in only one of the two relevant studies.

Beyond argumentation, the variation between the two studies is worthy of attention. In the first study, participants were asked to 'explain what mathematicians mean by proof in 40 words or fewer'. In the second, participants were asked a two-part question 'What do you think mathematical proof is? What are the most important attributes of a mathematical proof?' In both cases, judges rewarded responses referencing argumentation but differed in the other aspects they rewarded. For the first task, asking for an explanation, judges rewarded references to incontrovertibility and established knowledge. I view these features as philosophically appropriate and consistent with the literature on proof as an argument produced to remove doubt in the truth of the theorem (Czocher and Weber, in press), by showing the theorem to be the 'logical consequence of axioms, assumptions and/or previously established claims' (*ibid*, p. 20).

For the second task, asking what mathematical proof is and for important attributes, judges rewarded references to certainty and clarity. References to clarity can also be interpreted through the cluster definition of Czocher and Weber (in press), who noted proofs should be transparent justifications, comprehensible by any sufficiently knowledge parties. However, appeals to certainty can be viewed as signs of philosophical naivety, given the literature arguing against the primarily pre-19$^{\text{th}}$ century of mathematics as the business of certainty and truth (Marcus and McEvoy, 2016). This prompts an important question about the content validity of these judgments, given that one of the features rewarded by judges is not well aligned with the philosophical literature on the topic. I address this question as part of the discussion on validity below.

### 5.4.3 Research question 3a: On the reliability and validity of the Conceptions Task

Regarding the reliability of the comparative judgment-based scores, all cases based on expert judgments demonstrated acceptable statistical reliability suggesting that there was sufficient consensus amongst the judging cohorts to generate reliable scores in multiple settings. Given the range of judges recruited, I conjecture that any acceptably qualified set of judges would produce similar scores. Above a threshold of expertise, potentially as low as a relevant tertiary degree, the scores produced do not appear sensitive to the qualifications of the judges.

In addressing validity, I consider both criterion (concurrent and predictive) and content validity. Regarding concurrent validity, Chapter 4 reported that mathematicians outperformed students on the Conceptions Task. Regarding predictive validity, Chapter 5 demonstrated that conceptions scores reflected the expected improvement over the course of a 15-week Introduction to Proof module. Both conclusions are interpreted as evidence for the validity of conceptions scores as measures of philosophical awareness.

Regarding content validity, I first conducted a coding-based content analysis of the responses collected in each study. I then used the resulting codes to statistically model the Conceptions Task scores. The findings were discussed in answers to questions 1 and 2a above. In short, this regression modelling identified argumentation as the most important feature, alongside established knowledge, incontrovertibility, clarity and certainty. I argued that all bar one of these features is consistent with the established literature, suggesting strong evidence for the content validity of the conceptions scores. ‘Certainty’ was the one feature identified as significant in the statistical modelling and inconsistent with the literature on proof. That said, the unfavourability of the view that proof provides certainty is a relatively recent phenomenon and, as is evidenced by mathematicians’ own responses to the Conceptions Task, still features as an important aspect of proof for many mathematicians.

A further consideration in understanding validity is the origin of students’ conceptions. I speculate that students most likely adopt their views of mathematical topics from the mathematicians by whom they are taught. Many mathematicians also featured certainty in their responses to the Conceptions Task. Hence from the perspective that the comparative judgment-based scores are intended to reflect the collective expertise of the judging cohort, it is to be expected that conceptions containing reference to certainty be rewarded, even if this is not perfectly aligned with the philosophical literature on the topic.

Finally, I considered the relationship between the conceptions scores and more traditional proof comprehension measures. To this end, I found that while the conceptions scores appear to be robust reflections of the quality of individuals’ conceptions of proof, they are statistically unrelated to more traditional measures of proof comprehension.

The relationship between conceptions and comprehension is discussed in Chapter 10, alongside summative remarks on the use of comparative judgment in the realm of proof.

### 5.4.4 Implications and conclusions

Beyond the unique contribution of documenting the written conceptions of students and mathematicians, I conclude that there is strong evidence that the Conceptions Task yielded reliable and valid scores, reflective of individuals’ philosophical awareness regarding proof. This has substantive implications for the literature in the subjective realm of conceptions and beliefs. Reliable and valid instruments on such topics are difficult to generate and other approaches necessarily involve coarse-grained analyses or lack the capacity for systematic quantitative comparison. Using comparative judgment, I have generated scores for responses to an open-ended task without using pre-determined criteria or a restrictive definition of proof.

While the evidence here is presented in only one domain, mathematical proof, it seems that such an approach would be profitable in evaluating individuals’ conceptions of other (mathematical) topics. Beyond the scope of the research presented here, I speculate that similar approaches could be used to evaluate more esoteric beliefs within and outside of mathematics. It may be possible either to develop meaningful scores for the quality of certain beliefs viewed as outside the realm of quantitative analysis, or to learn about the beliefs of a judging cohort via a mixed methods analysis of the responses they deem most valuable (Section 5.2.4) . I return to further methodological implications of this work in the final chapter.

This concludes the work on the Conceptions Task. I return to several topics discussed here in greater detail in the final chapter, after having presented a series of further studies focused on proof comprehension and the Summary Task.

### Next chapter

The following chapter is the first of four on the Summary Task, and focuses on the uncountability proof.

# Chapter 6

# Proof summaries I: The open unit interval is uncountable

This chapter, which begins Part Two, presents the first of five studies on the Summary Task. In this and subsequent chapters, I address two research questions:

> Research question 2b: What do mathematicians most value when evaluating students' proof summaries?
>
> Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

In addressing 2b, Chapters 6, 7 and 8 present a series of content analyses on students' summaries of three different proofs. On each occasion, I use statistical modelling to predict the Summary Task scores with content-based codes. The result of this statistical analysis is a list of mathematical features most valued by the mathematician judges. Finally, the interviews presented in Chapter 9 provide a qualitative perspective on the same question and are used to triangulate across the two perspectives. While providing a direct answer to research question 2b, these analyses also provide insight into the validity of the resulting scores via comparisons with the theoretical literature on proof comprehension assessment.

The reliability aspect of research question 3b is addressed using the two statistical measures discussed in Chapter 3: Scale Separation Reliability and inter-rater reliability. In addressing the validity of the Summary Task scores, I consider both criterion and content validity. Criterion validity analysis is based on statistical comparisons between the Summary Task scores and a series of established measures of proof comprehension and general mathematical expertise. These measures include Proof Comprehension Tests (from Mejia-Ramos et al., 2017), various modules from undergraduate mathematics and SAT scores. Content validity is evaluated in Chapters 6, 7 and 8, using the regression modelling from research question 2b. As above, Chapter 9's interview analysis also provides insight on the content validity of the resulting scores by directly considering the judges' spoken understandings of their own decision-making processes.

In this chapter, in particular, I focus on students' summaries of the *uncountability proof*, demonstrating the uncountability of the open unit interval. I establish preliminary evidence for the reliability of the Summary Task scores in this context, followed by quantitatively driven investigations of criterion and content validity.

The data collected for this study were collected alongside the data presented in Chapter 4. With the same set of student participants in both studies, I present ancillary analyses comparing students' performance on Conceptions and Summary Tasks.

## 6.1 Methods

### 6.1.1 Materials

A task booklet was generated containing a theorem and its proof (see Figure 6.1), two proof comprehension tasks and the Conceptions Task. The first was a multiple-choice Proof Comprehension Test from Mejia-Ramos et al. (2017). Permission has not been granted for the specific questions to be published, but the full test can be requested at pcrg.gse.rutgers.edu. The other is known as the Summary Task, asking students to 'summarise the proof in 40 words or fewer' (see Figure 6.2). This booklet ended with the Conceptions Task, discussed in Chapter 4.

### 6.1.2 Participants

One hundred and sixty-one undergraduate mathematics students from the same British university participated in this study. Eighteen declined to have their data used for research purposes leaving a total of 143 participants. All participants were enrolled in a compulsory introductory module on Real Analysis for first

**Theorem:** The open interval $(0,1)$ is uncountable.

**Proof:** The interval $(0,1)$ includes the subset $\left\{\frac{1}{2^k} : k \in \mathbb{N}\right\}$, which is infinite. Thus, $(0,1)$ is infinite.

Suppose $(0,1)$ is denumerable. Then, there is a function $f : \mathbb{N} \to (0,1)$ that is one-to-one and onto $(0,1)$. Now, we write the images of $f$, for each $n \in \mathbb{N}$, in their decimal form:

$$\begin{aligned}
f(1) &= 0.a_{11}a_{12}a_{13}a_{14}a_{15}...\\
f(2) &= 0.a_{21}a_{22}a_{23}a_{24}a_{25}...\\
f(3) &= 0.a_{31}a_{32}a_{33}a_{34}a_{35}...\\
f(4) &= 0.a_{41}a_{42}a_{43}a_{44}a_{45}...\\
&\vdots\\
f(n) &= 0.a_{n1}a_{n2}a_{n3}a_{n4}a_{n5}...\\
&\vdots
\end{aligned}$$

Since some elements of $(0,1)$ have two different decimal representations (one with an infinite string of 9's and another one with an infinite string of 0's), we do not use representations that contain an infinite string of 9's. That is, for all $n \in \mathbb{N}$ we represent $f(n) = 0.a_{n1}a_{n2}a_{n3}a_{n4}a_{n5}...$ in such a way that there is no $k$ such that for all $i > k$, $a_{ni} = 9$.

Now let $b$ be the number $b = 0.b_1b_2b_3b_4b_5...$, where $b_i = 5$ if $a_{ii} \neq 5$ and $b_i = 3$ if $a_{ii} = 5$. Because of the way $b$ has been constructed, we know that $b \in (0,1)$ and that $b$ has a unique decimal representation. However, for each natural number $n$, $b$ differs from $f(n)$ in the $n$th decimal place. Thus $b \neq f(n)$ for any $n \in \mathbb{N}$, which means $b$ does not belong to the range of $f$. Thus, $f$ is not onto $(0,1)$. This contradicts our assumptions. Therefore, $(0,1)$ is not denumerable. □

*Figure 6.1. The proof given to participants showing the uncountability of the unit interval. This proof was the basis for both the Summary Task and Proof Comprehension Test.*

Summarise the proof, given on the previous page, **in 40 words or fewer**.

Note: You are not being asked to reproduce the proof. The best responses will be those that succinctly communicate the most important aspects/ideas in the proof.

Write your summary in the box below:

*Figure 6.2. The Summary Task.*

and second-year students, covering fundamental concepts related to sequences, series and epsilon-$N$ definitions.

Participation was not connected to examination and was made voluntary by giving students the option to have their data excluded from any analysis. However, those present at the lecture were required to complete the booklet. Participants were told that general feedback on overall student performance would be given to the lecturer, based on the anonymised data.

### 6.1.3 Procedure

Data collection took place in a week-eight lecture, with content directly related to the task presented in the preceding lecture. Participants were given 40 minutes and advised to spend 20 minutes on the Proof Comprehension Test, and 10 minutes on both the Summary and Conceptions Tasks. For practical reasons neither the time allocation nor response order were monitored but there was no reason to believe a substantial number of participants ignored these instructions.

Module scores were also made available by the lecturer and were used as a secondary comparison measure against which to evaluate the concurrent validity of the Summary Task scores. The module scores probably capture a more general measure of mathematical success than the Proof Comprehension Test. However, it seems reasonable to expect a measure of proof comprehension to correlate with scores on an introductory Real Analysis module.

Consent for analysis of module scores was given by 134 of the 143 participants.

### 6.1.4 Comparative judgment

Eleven judges were recruited using contacts from previous similar studies. Seven were PhD students of mathematics at the same British university, three were current PhD students from a second British university and one was a recent PhD graduate from a third. All judges were deemed qualified to assess introductory analysis by virtue of their own course of study.

All judges were asked to read the relevant proof before judging and advised to keep it on hand throughout the process.

Each judge performed 143 pairwise comparisons, resulting in a total of 1573 judgments. Based on an informal pilot study and previous experience, judges were paid based on an expected average of 20 seconds per judgment. In this study, the median time per judgment was 21.6 seconds.

The Summary Task scores had mean 0.00 ($\sigma = 1.78$).

### 6.1.5 Data analysis

First, I evaluated reliability using Scale Separation Reliability and inter-rater reliability. I then considered criterion validity by comparing the Summary Task scores with students' scores on the Proof Comprehension Test and module scores. This is followed by a content analysis, consistent with the principles of thematic analysis (Braun and Clarke, 2006), aiming to provide a more holistic view of the proof summaries given by students. This content analysis formed the basis of a statistical analysis identifying the features of students' summaries most valued by the mathematician judges. Finally, I examined information density, compared with Summary Task scores and summary word-count, to understand the extent to which judges prioritised brevity in their decision-making.

## 6.2 Results

### 6.2.1 Example responses

To orient the reader to the types of responses given, I present the top three summaries as judged by the research-active mathematicians (see Figure 6.3).

### 6.2.2 Reliability of Summary Task scores

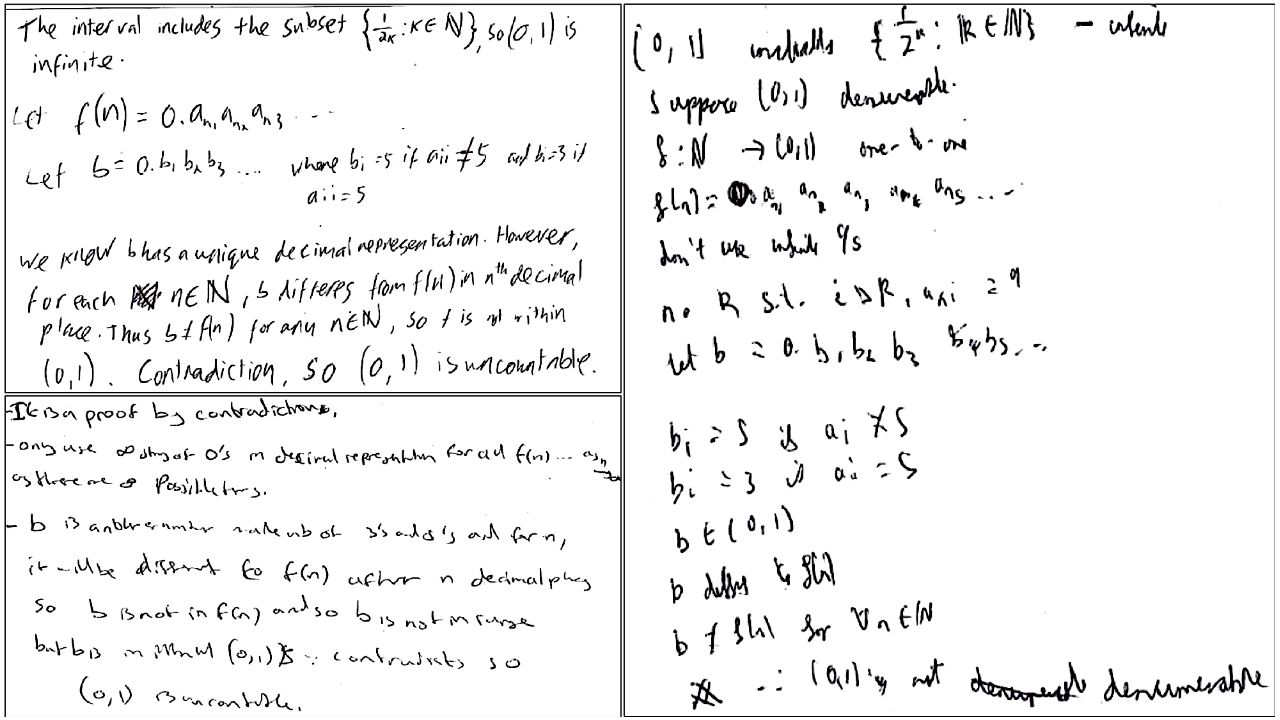
Reliability was examined in two ways. First, internal consistency was measured using Scale Separation Reliability and found to be acceptable, SSR = .86. Inter-rater reliability, measured using 100 iterations of the split-half technique discussed in Chapter 3 was also high, $r = .73$.

### 6.2.3 Introductory Real Analysis scores

Scores from the Introductory Real Analysis module ranged from 33% to 97%, with a mean of 56% ($\sigma = 14$). These were based on a weighted aggregate of students' coursework (25%) and final examination scores (75%) for the module from which they were recruited.

### 6.2.4 Proof Comprehension Test

For the multiple-choice Proof Comprehension Test, the internal consistency was measured by Cronbach's $\alpha = .53$. This was substantially lower than the $\alpha > .7$ reported in all trials reported by Mejia-Ramos et al. (2017), in which the same test was given to comparable students at a US university. I also note that



The interval includes the subset $\{\frac{1}{2k} : k \in \mathbb{N}\}$, so $(0,1)$ is infinite.

Let $f(n) = 0.a_{n_1}a_{n_2}a_{n_3}\ldots$

Let $b = 0.b_1b_2b_3\ldots$ where $b_i = 5$ if $a_{ii} \neq 5$ and $b_i = 3$ if $a_{ii} = 5$

We know b has a unique decimal representation. However, for each ~~[illegible]~~ $n \in \mathbb{N}$, b differes from $f(n)$ in $n^{th}$ decimal place. Thus $b \neq f(n)$ for any $n \in \mathbb{N}$, so f is not within $(0,1)$. Contradiction, so $(0,1)$ is uncountable.

- It is a proof by contradiction.
- only use ∞ strings of 0's in decimal representation for all $f(n)$ ... $a_{s_n}$ as there are ∞ possibilities.
- b is arbitrary number made up of 3's and 5's and for n, it will be different to $f(n)$ after n decimal places so b is not in $f(n)$ and so b is not in range but b is in interval $(0,1)$ ∴ contradicts so $(0,1)$ is uncountable.

$(0,1]$ includes $\{\frac{1}{2^k} : k \in \mathbb{N}\}$ – infinite

Suppose $(0,1)$ denumerable.

$f : \mathbb{N} \to (0,1)$ one-to-one

$f(n) = 0.a_{n_1}a_{n_2}a_{n_3}a_{n_4}a_{n_5}\ldots$

don't use infinite 9s

no k s.t. $i \geq k$, $a_{ki} = 9$

let $b = 0.b_1b_2b_3b_4b_5\ldots$

$b_i = 5$ if $a_i \neq 5$

$b_i = 3$ if $a_i = 5$

$b \in (0,1)$

b differs to $f(n)$

$b \neq f(n)$ for $\forall n \in \mathbb{N}$

$\#$ ∴ $(0,1)$ is not ~~denumerable~~ denumerable

Figure 6.3. The top three summaries of the uncountability, as determined by comparative judgment-based Summary Task scores.

students' scores out of 12 were low ($M = 4.19$, $\sigma = 2.16$). However, these scores were found to be significantly above the $M = 3$ one would expect if students answered questions randomly, $t(133) = 6.35, p < .001$. The low internal reliability is a notable limitation of the study and warrants further investigation before discussing the criterion validity analysis to follow.

In search of an explanation for the low Cronbach's alpha, a principal component analysis (PCA) was conducted on the 12-item test to investigate the possibility that the test measured multiple independent constructs. The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis KMO = 0.62 (mediocre but sufficient according to Field et al., 2012, p. 776). Bartlett's test of sphericity, $\chi^2(66) = 126.049, p < .001$, was significant, indicating correlations between items were sufficient to justify PCA. An initial analysis was run to obtain eigenvalues for each component in the data. The scree plot, shown in Figure 6.4, indicated that two components can be extracted, accounting from 29% of the variance. See Table 6.1.

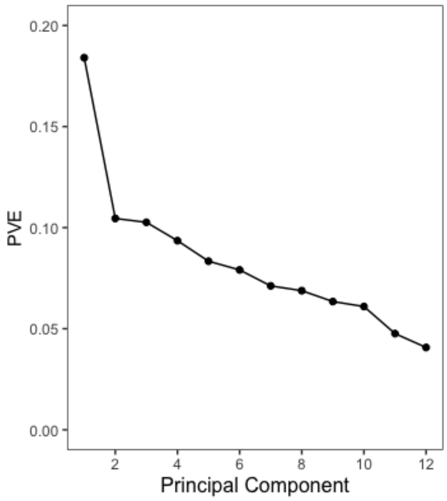


*Figure 6.4. Scree plot showing eigenvalues for the principal component analysis of the Proof Comprehension Test. Two components should be extracted.*

Component One, consisting of the seven questions with loadings above .4, had an internal reliability of $\alpha = .61$. This was an increase from the .53 found for the full 12-question test. While still below the standard .7 threshold for acceptable internal reliability, this suggests that questions 1, 5, 9, 11 and 20 were in some way problematic in this dataset. It should be noted that questions 2 and 6 both have non-trivial cross-loadings, but large enough primary loadings to be accepted as part of the first component. Component two comprised only questions 5 and 9 and did not appear theoretically meaningful.

*Table 6.1*

*Factor loadings for the PCA of the uncountability Proof Comprehension Test.*

| Question | Component One | Component Two |
|---|---|---|
| 1 | .34 | -.13 |
| 2 | **.61** | .27 |
| 5 | .04 | **.68** |
| 6 | **.57** | .25 |
| 9 | -.04 | **.65** |
| 11 | .10 | -.25 |
| 12 | **.53** | .16 |
| 15 | **.43** | -.20 |
| 16 | **.51** | -.06 |
| 18 | **.44** | -.28 |
| 19 | **.63** | -.14 |
| 20 | .27 | .04 |

*Note.* A bold item indicates a primary loading greater than .4. This 12-question test is a subset of the 20-question version presented then reduced in Mejia-Ramos et al. (2017). Original labellings have been retained to facilitate comparison with the original work.

Concerns regarding the reliability of the test were partially mitigated by the significant Pearson correlation with module scores, $r = .56, p < .001$, indicating that the test was not without meaning.

Given that only 29% of the variance was explained by the seven-item model, I present further analyses based on both the seven and 12-question versions of the test. I now turn attention to the Summary Task and the main body of analysis for this study.

### 6.2.5 Criterion validity

I first examined validity by comparing the Summary Task scores with established measures of proof comprehension and general mathematical performance.

The Summary Task and full 12-item Proof Comprehension Test offer the most important comparison, yielding a significant Spearman correlation, $\rho = .25, p < .001$ (see Figure 6.5). When using the seven-item version of the test (based on the PCA in Section 6.2.4), the correlation was not significantly different, $r = .28$, $p < .001$ ($Z = -0.27$, $p = .39$).

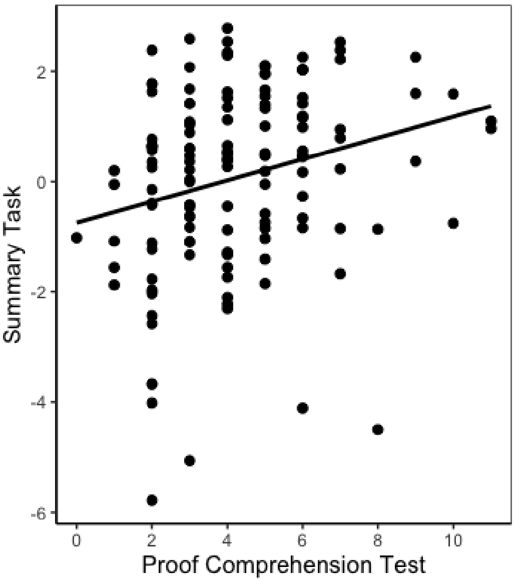


*Figure 6.5. Scatter plot comparison of performances on the Summary Task and 12-question Proof Comprehension Test.*

Summary Task scores were also significantly correlated with scores from the Introductory Real Analysis module, $r = .23$, $p < .001$.

From comparison with the Proof Comprehension Test, we learn that the Summary Task scores are probably indicative of local proof comprehension. Nevertheless, while the relationship was significant, the correlation coefficient was notably lower than in previous studies comparing comparative judgment-based scores and with established instruments. For example, Bisson et al. (2016) reported significant coefficients between .35 and .56 in similar investigations in secondary and tertiary mathematics. These included students' understanding of $p$-values in statistics, letters in algebra and derivatives in calculus.

On the other hand, the comparison between the Summary Task and the Introductory Real Analysis module suggests a more general domain of validity for the Summary Task scores. Further, this can be interpreted as evidence for the notion of proof comprehension as a singular entity, independent of the particular mathematical domain or proof in question.

I return to the relationship between Summary Task Scores and established measures in Chapters 7 and 8. In doing so, I develop a growing picture of the generality of the validity claims resulting from these quantitative comparisons.

### 6.2.6 Content analysis

I now turn attention to a more qualitatively oriented investigation of students' proof summaries. This section presents a content analysis addressing two issues not accessible through a strictly quantitative lens. First, I present a systematic analysis of the content students elected to include in their summaries. Second, I develop an understanding of the features most heavily rewarded by judges, and subsequently, the validity of the resulting scores based on these features.

#### Developing a code scheme

I developed the code scheme with a fellow researcher interested in proof comprehension[1]. The final scheme was the result of three iterative attempts at qualitatively describing the students' summaries. Each iteration was the result of an in-depth discussion between myself and one other researcher, focused on 10 proof summaries.

To generate the first iteration of the code scheme, we examined the original text (see Figure 6.1) alongside 10 summaries of the same proof from a pilot study that does not feature in this thesis. We elected not to begin with data from the main study in order to preserve the maximal number of responses for the final analysis.

Having found few student statements that could not be directly mapped to a discrete aspect of the proof, the first version of the coding scheme was simply a partitioning of the proof into 11 key ideas (codes).

In the second iteration, two researchers independently coded 10 summaries from the main dataset using the 11-code scheme, while highlighting any cases (pairs of codes and summaries) that appeared problematic to determine. Comparing these analyses led to a revised scheme, clarifying existing codes or dividing one code into several.

We were mindful to keep the scheme simple and opted only to increase the number of codes to capture substantive nuance. We also deemed it important to limit the necessity for value judgments on the quality of students' summaries.

The process of independently analysing 10 summaries prior to revising the scheme was repeated twice, now using responses from the primary dataset for this study. The result was the 15-code scheme presented in Table 6.2.

Having established a final scheme to use for the whole dataset, we turned to

[1] The development and implementation of the code scheme was conducted during an academic visit to a US university, alongside my academic host. As in earlier chapters, it was necessary to have multiple researchers involved in the coding process. The resulting analysis is solely my own.

*Table 6.2*

*Coding scheme for summaries of the uncountability proof.*

| Code | Frequency |
|---|---|
| Explicitly stated the interval $(0,1)$ is infinite. | 46% |
| Addressed a subset of $(0,1)$ (not necessarily explicitly naming $\{1/2^k : k \in \mathbb{N}\}$). | 48% |
| Explicitly related the infinitude of $(0,1)$ to an infinite subset. | 33% |
| Appealed to proof by contradiction (need not have featured associated wording, evidence of logical structure is sufficient). | 51% |
| Defined the function $f$ as a mapping $\mathbb{N} \to (0,1)$. | 46% |
| Described the function $f$ as injective. | 29% |
| Described the function $f$ as surjective. | 18% |
| Described the images of $f$ using decimal representation (any reference to decimal representations of $f(n)$'s is sufficient, reference to decimal representations of other values is not). | 28% |
| Appealed to 0's or 9's in reference to the decimal representations (accept references to elements of the range of $f$ or $(0,1)$, also accept ambiguity). | 37% |
| Addressed the constructed $b$ from the given proof in any way. | 70% |
| Constructed $b$ explicitly (sufficient to describe $b$ as a number differing from each $f(n)$ in the $n^{\text{th}}$ entry). | 41% |
| Explicitly stated that $b$ is not in the range of $f$ (or, that $b \neq f(n)$ for any $n$). | 50% |
| Explicitly stated that the constructed $b$ is in $(0,1)$. | 25% |
| Explicitly stated that $f$ is not surjective as a result of the surmised argument. | 17% |
| Included the term 'denumerable' anywhere. | 42% |

the remaining 123 summaries. Each researcher coded 75 summaries, leaving an intersection of 27 summaries to be used to evaluate inter-coder reliability. This intersection had a pooled Cohen's Kappa, $\kappa = .88$ with a 94.3% code-by-code agreement. The 23 instances of disagreement were discussed and found to be either coder errors or unique, unanticipated cases for which our scheme did not account. In the latter case, a decision was reached by attempting to maintain the clarity of each code, opting not to award any code to a clause for which we did not have an obvious code. These cases were rare enough not to warrant further revisions.

*Table 6.3*

*Regression modelling of the uncountability proof Summary Task scores.*

| Code | Coefficients | | Regression model | | | |
|---|---|---|---|---|---|---|
| | $\rho$ | $p$ | $B$ | SE | $\beta$ | $p$ |
| $(0,1)$ is infinite | 0.15 | .101 | | | | |
| $(0,1)$ has an infinite subset | 0.06 | .536 | | | | |
| Explicitly related codes 1 and 2 | 0.13 | .159 | | | | |
| Contradiction | 0.33 | <.001* | 0.70 | 0.26 | 0.19 | .008* |
| Defining $f$ | 0.38 | <.001* | 0.85 | 0.26 | 0.23 | .012* |
| $f$ is injective | 0.21 | .024 | | | | |
| $f$ is injective | 0.19 | .031 | | | | |
| Decimal representations | 0.37 | <.001* | 0.61 | 0.31 | 0.15 | .052 |
| 0's and 9's | 0.39 | <.001* | 0.31 | 0.29 | 0.08 | .293 |
| Introducing $b$ | 0.54 | <.001* | 0.93 | 0.39 | 0.23 | .024* |
| Constructed $b$ explicitly | 0.38 | <.001* | 0.51 | 0.30 | 0.14 | .086 |
| Stated $b \notin R(f)$ | 0.51 | <.001* | 0.30 | 0.34 | 0.08 | .387 |
| Stated $b \in (0,1)$ | 0.23 | .009 | | | | |
| $f$ is not surjective | 0.32 | <.001* | 0.71 | 0.34 | 0.15 | .043* |
| Denumerable | 0.22 | .017 | | | | |

*Note.* Codes with a significant Spearman coefficient were entered into a force-entry regression model as predictors of Summary Task scores. Significance was determined using the Holm-Bonferroni correction with $\alpha = .05$. The resulting eight-code model, $F(12, 110) = 9.65$, $p < .001$, explains 51% of the variance.

### Identifying important codes

Here, I present a regression analysis identifying the codes most rewarded by judges. I first examined Spearman correlations between each code and the Summary Task Scores, see Table 6.3. Codes that yielded a significant correlation were then entered into a forced-entry regression to identify those most predictive of Summary Task scores.

Eight of the 15 codes were significantly related to Summary Task scores when considered in isolation. Four of these eight were deemed significant predictors of Summary Task score in the force-entry regression model. This model, $F(12, 110) = 9.65$, $p < .001$, explained 51% of the variance[2].

I discuss the implications of these findings in addressing research questions 2b and 3b later in this chapter.

[2] I also ran forced-entry regressions with all 15 codes, and with 12 codes using the significant univariate predictors before Holm-Bonferroni correction ($\alpha = .05$). All led to similar conclusions with respective explained variance, .53% and .52%.

### 6.2.7 Information density analysis

Having established an understanding of the relative importance of the content present, I turned to a more global property of proof summaries. According to the Oxford English Dictionary, *summary* is defined as a noun to be 'a brief statement or account of the main points', and as an adjective as 'not including needless details or formalities; brief'. It, therefore, seems reasonable to believe that brevity would be a desirable quality of students' proof summaries. I conjectured that information density would provide a significant predictor of the Summary Task scores. Information density is defined here as the ratio of codes awarded to word-count. Three blank responses were removed to avoid division-by-zero errors, leaving 120 for the resulting analysis.

The summaries had a median length of 40 words, with a range of 10 to 159. Fifty-seven of the 120 summaries were longer than 40 words. While it is clear that many participants disregarded the word limit, I have no evidence to suggest that the *summary* aspect of the task was ignored or misunderstood. Long responses tended simply to be less succinct summaries of the given proof.

A comparison between information density and Summary Task scores yielded a significant Spearman correlation, $\rho = 0.18$, $p = .048$. While this does confirm the hypothesis, the low correlation coefficient suggested further analysis was necessary. Moreover these scores were more closely related to word-count ($\rho = .49$, $p < .001$) and number-of-codes ($\rho = .59, p < .001$, see Figure 6.6), than information density. This suggests that judges may have actively rewarded longer summaries, somewhat threatening the validity of a task fundamentally based on a request for brevity. This is consistent with the earlier finding that a majority of codes (eight of 15) correlated significantly with Summary Task scores, indicating that the inclusion of most aspects of the proof were viewed favourably by judges.

On the other hand, it is possible that this relationship between information volume and Summary Task scores is not a reflection of judges' approaches to decision-making, but evidence that students who understood less simply wrote less. To examine this possibility, I repeated the analysis using only the top half of responses as determined by Proof Comprehension Test score, thus excluding those with the weakest understanding of the proof. For these 60 summaries, Summary Task scores were not significantly related to information density ($\rho = .23, p = .075$) or word-count ($\rho = .18, p = .146$). Number-of-codes was still significantly related to Summary Task scores, $\rho = .36, p = .004$. However, this correlation coefficient is significantly lower, $Z = 1.85, p = .032$, than the equivalent coefficient generated using the full dataset.

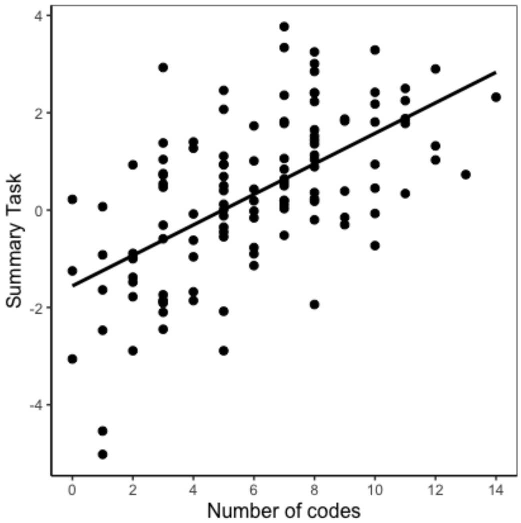


*Figure 6.6. Scatter plot comparison of the number of codes assigned to a each summary, and the respective Summary Task scores.*

These findings support the claim that students who understood less, wrote less. I return to the relationship between brevity and Summary Task scores in Chapter 9.

## 6.3 Interim discussion

This chapter presented the first study in a series of investigations aimed at understanding judges' priorities in evaluating students' proof summaries (research question 2b), as well as the reliability and validity of the resulting scores (research question 3b).

### 6.3.1 Research question 2b: What do mathematicians most value when evaluating students' proof summaries?

To investigate mathematicians' priorities when evaluating students' proof summaries, I first conducted a coding-based content analysis of the summaries. These codes were then entered into a statistical model predicting Summary Task scores. Based on the beta values from the resulting regression model, the codes *defining $f$* and *introducing $b$* were identified as most important to judges' decision-making. These two codes refer to the two major mathematical objects used in the proof. The function $f$ is the subject of the contradiction and the constructed number, $b$, is the value used to demonstrate the contradiction itself. The other two significant codes in the final model are *contradiction* and the statement $f$ *is not surjective*. Given that $f$ not being surjective is the precise statement of the contradiction, I claim that both codes address the proof method.

The implications of these findings for the validity of the task is discussed in answer to question 3b on the validity of the Summary Task scores.

### 6.3.2 Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

#### Reliability

Scale Separation and inter-rater reliability measures showed strong evidence for the reliability of the Summary Task scores in this context.

This suggests that, at least for the uncountability proof, mathematicians demonstrated a substantive degree of consensus regarding the nature of appropriate proof summaries. The generality of this finding remains an open question, to which I return in both of the next two chapters.

#### Criterion Validity

Regarding criterion validity, I have reported evidence for the validity of the Summary Task scores based on comparisons with the established Proof Comprehension Test and students' performance in the Introductory Real Analysis module from which they were recruited. The resulting correlation coefficients, while significant, were lower than in previous similar studies with comparative judgment-based assessment in other mathematical domains (Bisson et al., 2016). For the Proof Comprehension Test, the lower than expected coefficient can be partially explained by the low internal reliability of the test itself and hence is still interpreted as moderate evidence for the validity of the Summary Task scores as a measure of local proof comprehension. Again, I return to the generalisability of this conclusion in the following two chapters as I report related evidence from similar studies on two other mathematical proofs.

#### Content Validity

In addressing research question 2b, on mathematicians' priorities when evaluating students' proof summaries, I appealed to the coding-based content analysis of students' summaries. I do similarly here, comparing the codes identified as significant by the statistical modelling with the literature on proof comprehension assessment.

In particular, the statistical modelling indicated an arguably predictable pattern: that mathematicians rewarded summaries that capture the proof method and that refer to the key objects of study. Rewarding references to the proof method is consistent with at least three aspects of the proof comprehension assessment model from Mejia-Ramos et al. (2012) summarised in Table 2.2: 'Logical status of statements and proof framework', 'Summarising via high-level ideas', and 'Identifying the modular structure'. References to key mathematical objects can be interpreted as demonstrating understanding of the 'meaning of terms and statements'. While this focus on key objects may prove to be an idiosyncratic feature of the proof at hand, I conjecture that at least the focus on proof methods will generalise to other mathematical contexts.

In sum, I interpret this content analysis as evidence for the content validity of the resulting scores.

Finally, on the topic of content validity, I considered information density as a potential predictor of Summary Task scores. I found a significant correlation between information density and Summary Task scores. However, this does not appear to be robust given that this relationship was not replicated when considering only the top half of students' summaries. Further, word-count was more closely related to the Summary Task than information density, suggesting that judges may not have meaningfully engaged with the 'summary' aspect of the instructions given to students.

This analysis is consistent with the findings of Benton et al. (2018) who found when evaluating essay quality using comparative judgment that the shortest essay in an English exam received the lowest scores, but that at the top end, the relationship between length and perceived quality disappeared. While this analysis does not preclude the possibility that judges actively rewarded volume over brevity, it does temper any assertions in this direction. I return to judges' decision-making in Chapter 9, but for now, I conclude that there is limited evidence for density or volume of information as a direct influence on judges' decision-making.

The following two chapters examine similar applications of the Summary Task on two further mathematical proofs. I reserve more substantive discussion for the end of Chapter 8 after having presented the relevant quantitatively oriented data for each proof.

### Next Chapter

The next chapter presents data associated with the *primes* proof, demonstrating the infinitude of the prime integers.

# Chapter 7

# Proof summaries II: There are infinitely many primes

This chapter focuses on students' summaries of a proof of the infinitude of prime numbers. I chose this theorem because of the associated Proof Comprehension Test (Mejia-Ramos et al., 2017), used as a benchmark for the Summary Task in this chapter.

As in the previous chapter, I address research questions 2b, on mathematicians' priorities in evaluating students' work, and 3b, on the reliability and validity of the Summary Task scores. Here, I build on the work in the previous chapter by mimicking the previous design, with two important changes. First, by changing the mathematical proof, I gather evidence on the external validity of the Summary Task, in terms of its generalisability across mathematical settings. Second, this study includes a wider array of associated measures including SAT (Standardised Aptitude Test) scores and past, present and future module scores.

The data presented in this chapter were collected under the guidance of my academic host professor, during a visit to a US university in my second year of study. These data were collected as part of a department-wide research team and were a subset of a larger project focused on the teaching and learning of undergraduate mathematics. The design of the study presented here is my own conception, as are all the analyses presented in this chapter. The data presented in the following chapter on the Fibonacci Proof were also collected as part of the same project. I reserve analysis comparing data across mathematical settings for the following chapter, once all data for each proof have been presented.

## 7.1 Methods

### 7.1.1 Materials

A task booklet was created, containing the theorem and its proof (see Figure 7.1), followed by the Summary Task then the associated Proof Comprehension Test[1]. I refer to the proof in Figure 7.1 as the primes proof. As before, the Summary Task asked students to 'summarise the proof in 40 words or fewer' (Figure 6.2).

**Theorem:** The set of prime numbers is infinite.
**Proof:** Suppose the set of primes is finite. Let $p_1, p_2, p_3, \ldots, p_k$ be all those primes with $p_1 < p_2 < \cdots < p_k$. Let $n$ be one more than the product of all of them. That is, $n = (p_1 p_2 p_3 \ldots p_k) + 1$. Then $n$ is a natural number greater than 1, so $n$ has a prime divisor $q$. Since $q$ is primes, $q > 1$. Since $q$ is prime and $p_1, p_2, p_3, \ldots, p_k$ are *all* the primes, $q$ is one of the $p_i$ in the list. Thus, $q$ divides the product $p_1 p_2 p_3 \ldots p_k$. Since $q$ divides $n$, $q$ divides the difference $n - p_1 p_2 p_3 \ldots p_k$. But this difference is 1, so $q = 1$. From the contradiction $q > 1$ and $q = 1$, we conclude that the assumption that the set of primes is finite is false. Therefore, the set of primes is infinite.

*Figure 7.1. The primes proof demonstrating the infinitude of the prime integers. This proof was the basis for both the Summary Task and Proof Comprehension Test in this chapter.*

### 7.1.2 Student participants

Eighty-two undergraduate students participated in this study. All were enrolled in Introduction to Proof, a second-year module at a university in the United States of America. These students were recruited from all nine sections[2] of the module but the exact distribution was not recorded.

Participation was made voluntary by allowing students to opt out of any data analysis. However, the booklet was administered during lecture-time and those in attendance were required to complete the tasks. It was made clear that module credit was not associated with their participation in this research.

### 7.1.3 Procedure

Data collection took place in the second half of a week-seven lecture. Participants received a task booklet and were told they would have the remainder of

[1] Available at pcrg.gse.rutgers.edu/

[2] At this university, large modules are split into sections of approximately 30 students, taught by several academics but administered collectively.

the lecture (at least 30 minutes) to complete it.

Six other measures of student performance were made available by the university's centralised Office of the Registrar. These include mathematics SATs, as well as final scores for five modules: Introduction to Proof, Introductory Calculus, Further Calculus, Linear Algebra and Abstract Algebra.

None of these modules was compulsory for students, so with the exception of Introduction to Proof (from which the participants were recruited), each measure had missing data. The size of the data for each module is included in the results section.

### 7.1.4 Comparative judgment

Students' summaries were uploaded to www.nomoremarking.com. Fourteen judges were recruited using contacts from previous work. All judges were either PhD students of mathematics or held academic positions in a department of mathematics or mathematics education.

Participants were invited via email to sign up at onlinesurveys.com, where they saw the theorem, its proof, and a request for up to 100 judgments. Consenting judges were then directed to the judging platform to complete their judgments. Judges performed between 17 and 102 judgments, resulting in a total of 919. Only one participant was allowed to perform more than 100 judgments, as a result of a programming error that was fixed early in the data collection process. Each summary received between 21 and 26 comparisons, while each comparison took a median of 22.2 seconds. No compensation was offered.

These scores had a mean $M = 0.00$ ($\sigma = 2.01$). As in the previous study, the primary focus of this study is the evaluation of these scores, and their relationship with the content of the summaries themselves.

### 7.1.5 Data analysis

Reliability was first evaluated using the standard measures discussed in Chapter 3. Criterion validity was then considered, using a series of comparisons between the Summary Task scores and other measures collected. I report statistics for each of the measures, before considering their relationships with comparative judgment-based scores, both as isolated correlations and as a forced-entry regression model using all available data. Finally, I present a content analysis of the students' summaries and attempt to use the content-based codes as predictors of performance.

## 7.2 Results

### 7.2.1 Example responses

To orient the reader to the types of responses received, I provide the top three summaries in Figure 7.2.

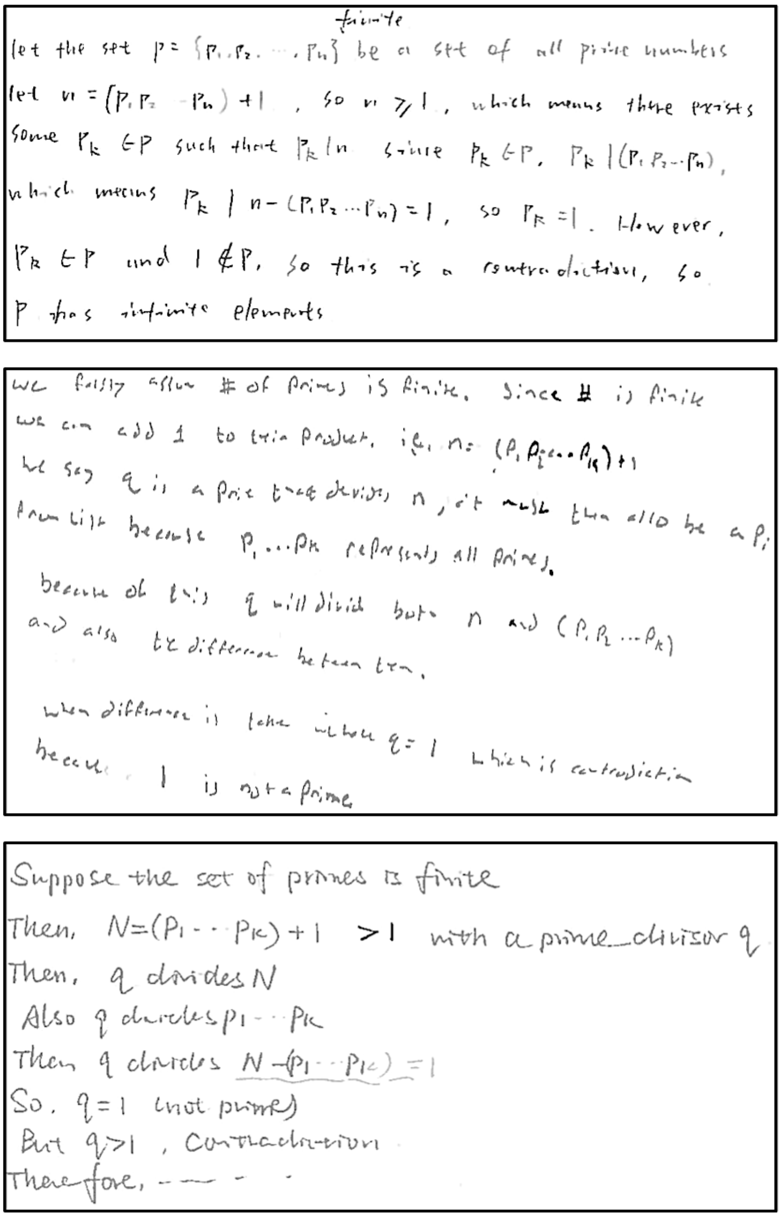
### 7.2.2 Reliability of Summary Task scores

Reliability was examined in two ways and led to the conclusion that this dataset had sufficient reliability. Internal consistency was estimated using SSR and found to be acceptable, SSR = .87, while inter-rater reliability was also acceptable, $r = .76$, based on 100 iterations.

### 7.2.3 Proof Comprehension Test

The multiple-choice Proof Comprehension Test was administered at the same time as the Summary Task, so all 82 participants completed this test. These scores ranged from 1 to 12 (out of 12) with mean, $M = 6.6$, and standard deviation, $\sigma = 2.6$.

The internal consistency of the test was measured using Cronbach's $\alpha = .66$. This was below the desirable .7 threshold, but a substantive improvement on the .53 found for the test associated with the uncountability proof discussed in the previous chapter. Unlike in the previous chapter, the mean scores were high enough to eliminate the possibility that a large volume of random guesses generated sufficient noise to interfere meaningfully with Cronbach's alpha. A principal components analysis was conducted to investigate possible causes of low reliability. The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = 0.61. Bartlett's test of sphericity, $\chi^2(66) = 134.10, p < .001$, was significant, indicating correlations between items were sufficient to justify principal component analysis. An initial analysis was run to obtain eigenvalues for each component in the data.



finite

let the set $P=\{P_1,P_2,\cdots,P_n\}$ be a set of all prime numbers
let $n=(P_1P_2 \cdots P_n)+1$, so $n \geqslant 1$, which means there exists
some $P_k \in P$ such that $P_k | n$ since $P_k \in P$, $P_k | (P_1P_2 \cdots P_n)$,
which means $P_k \mid n-(P_1P_2\cdots P_n)=1$, so $P_k=1$. However,
$P_k \in P$ and $1 \notin P$, so this is a contradiction, so
P has infinite elements

we falsly assum # of primes is finite. Since # is finite
we can add 1 to their product, i.e., $n=(P_1P_2\cdots P_k)+1$
we say q is a prie that devides n, it must then also be a $p_i$
from list because $p_1 \ldots p_k$ represents all primes.
because of this q will divid both n and $(P_1P_2 \cdots P_k)$
and also the difference between them.
when difference is take we have $q=1$ which is contradiction
because 1 is not a prime

Suppose the set of primes is finite
Then, $N=(p_1 \cdots p_k)+1 > 1$ with a prime divisor q
Then, q divides N
Also q divides $p_1 \cdots p_k$
Then q divides $\underline{N-(p_1 \cdots p_k)=1}$
So, $q=1$ (not prime)
But $q>1$, contradiction
Therefore, - — -

*Figure 7.2. The top three summaries of the primes proof, as determined by comparative judgment-based scores.*

The scree plot, shown in Figure 6.4, indicated that two components can be extracted, accounting for 34% of the variance. See Table 7.1.

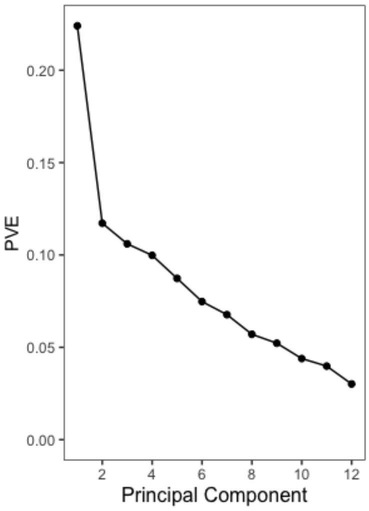


*Figure 7.3. Scree plot showing eigenvalues for the principal component analysis of the Proof Comprehension Test. Two components should be extracted.*

*Table 7.1*

*Factor loadings for the PCA of the primes Proof Comprehension Test.*

| Question | Component One | Component Two |
|---|---|---|
| 1 | .33 | .23 |
| 2 | **.42** | .33 |
| 3 | **.42** | -.52 |
| 8 | **.62** | .07 |
| 9 | **.62** | -.30 |
| 11 | **.49** | .02 |
| 14 | .06 | .16 |
| 15 | **.40** | -.58 |
| 17 | **.63** | .06 |
| 18 | **.49** | -.01 |
| 19 | **.58** | .25 |

*Note.* A bold item indicates a primary loading greater than .4. This 12-question test is a reduction of the 20-question original version presented in Mejia-Ramos et al. (2017). Original labellings were retained to facilitate comparison with the original work.

Despite the scree plot indicating a two-component structure, the loadings table showed that 11 of the 12 questions load onto Component One. Upon removal of the only question not loading onto Component One (question 14), Cronbach's alpha increased to .68, still below the standard .7 threshold. Given this minimal change in alpha, coupled with the uninformative loadings table, the criterion analysis later in this chapter is based only on the full 12-question version of the test.

### 7.2.4 Mathematics SAT scores

Mathematics SATs were available for 64 of the 82 participants in this study. These scores had a mean, $M = 709$, with standard deviation, $\sigma = 69$, and ranged from 510 to 800.

### 7.2.5 Module Scores

Students' scores were available for five modules. Here, I present descriptive statistics for each. In all cases, grades were assigned using a letter-based system from $A$ to $F$. There were converted to numerical values using the university's convention: A = 4, B+ = 3.5, B = 3, C+ = 2.5, C = 2, D = 1, F = 0.

#### Introduction to Proof

All 82 students were enrolled in Introduction to Proof at the time of data collection. Only one of the 82 participants did not sit the exam for this module. For the 81 available scores, the mean was 2.81, between grade-boundaries B and C+, with a mode grade of B+ (= 3.5). The full range was represented with 3 students receiving an F and 10 receiving the top grade, A.

#### Introduction to Calculus

This is a compulsory first-year module for students at this university, and was sat by 32 of the 82 participants before their participation in Introduction to Proof. Three students had attempted this module twice. For simplicity, I take the best of their attempts. For the 32 available scores, the mean was 3.27, with a mode of 4.

#### Further Calculus

Forty-six students completed this module in the semester following their completion of Introduction to Proof. These had a mean of 2.99 and mode grade of 2.

**Linear Algebra**

Thirty-one students completed Linear Algebra in the semester following their completion of Introduction to Proof. These scores had a mean of 2.74 and a mode of 3.

**Abstract Algebra**

Ten students completed this module in the semester following their completion of Introduction to Proof. These scores had a mean of 2.65 and a mode of 2.

### 7.2.6 Criterion validity analysis

Here, I examine criterion validity through a series of Spearman correlations comparing Summary Task scores with each of the established measures in isolation, see Table 7.2.

*Table 7.2*

*Comparing the primes Summary Task with other measures.*

| Benchmark | $N$ | $\rho$ | $p$ |
|---|---|---|---|
| Proof Comprehension Test | 82 | .23 | .034 |
| Mathematics SAT | 64 | .34 | .262 |
| Intro to Proof | 81 | .32 | .300 |
| Introductory Calculus | 32 | -.07 | .706 |
| Further Calculus | 46 | -.06 | .699 |
| Linear Algebra | 31 | .08 | .674 |
| Abstract Algebra | 10 | -.07 | .849 |

*Note.* Significance determined by the Holm Bonferroni-corrected method with initial $\alpha = .05$.

In contrast to previous chapters, summaries of the primes proof yielded no significant correlations after Holm-Bonferroni correction. Bonferroni methods are known to be conservative estimates of significance, and it is worth noting that the Proof Comprehension Test scores met a less stringent .05 threshold. This relationship is plotted in Figure 7.4. This can be interpreted as relatively weak evidence for the criterion validity, although the dataset as a whole does not support such a conclusion.

One could also make a case for excluding the Abstract Algebra data from the analysis given the small volume of related data. This would make the Holm-Bonferroni thresholds marginally more generous but does not substantively influence the analysis to follow.

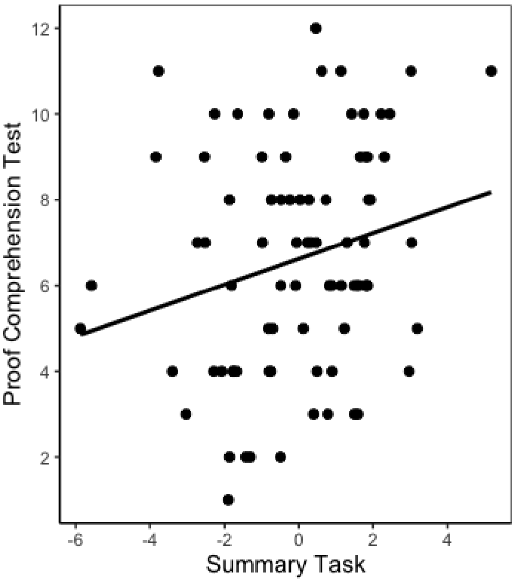


*Figure 7.4.* *Scatter plot comparison of Proof Comprehension Test and the Summary Task,* $\rho = .23, p = .034$.

To further understand the data, I consider the relationships between the Proof Comprehension Test and other measures. A series of Spearman correlations are shown in Table 7.3.

While the Proof Comprehension Test is significantly related to one of the six other measures under the Holm-Bonferroni correction, three others pass the standard $\alpha = .05$ threshold. The exact nature of the relationship between the Proof Comprehension Tests and these benchmark measures is unclear from the data presented. However, Table 7.3 shows that the Proof Comprehension Test is substantively more indicative of module scores than the Summary Task scores. Further discussion of this finding is reserved for the following chapter, when analysis in the context of the Fibonacci proof can also be considered. For now, I turn attention to a content analysis of students' summaries, attempting to understand the comparative judgment-based scores through the various aspects of the proof they elected to include.

### 7.2.7 Content analysis

In this section, I present a content analysis of the proof summaries themselves. This is intended to provide insight both into the nature of the students' written work, and into the values of the judging mathematicians. I first present the origins and implementation of the content-based coding, followed by a regression

*Table 7.3*

*Comparing the primes Proof Comprehension Test with other measures.*

| Benchmark | $N$ | $\rho$ | $p$ |
|---|---|---|---|
| Mathematics SAT | 64 | .31 | .012 |
| Intro to Proof | 81 | .44 | <.001* |
| Introductory Calculus | 32 | -.38 | .312 |
| Further Calculus | 46 | -.29 | .044 |
| Linear Algebra | 31 | .41 | .024 |
| Abstract Algebra | 10 | .22 | .548 |

*Note.* Significance determined by the Holm Bonferroni-corrected method with initial $\alpha = .05$.

analysis attempting to predict Summary Task scores using the codes produced.

### Developing a coding scheme

In generating a coding scheme for these summaries, I first examined the proof and a set of 10 summaries alone, generating a 15-code scheme. I then led a research meeting with an academic supervisor, wherein we jointly analysed a new set of 10 summaries, discussing the suitability of the existing codes for the data at hand, amending, adding and removing codes where we deemed fit. Both researchers then independently implemented the resulting 18-code scheme on all 82 summaries. As in the previous chapter, we were mindful to keep the scheme as simple as possible. We also adopted the same principle of generosity from the earlier work. The final coding scheme is presented in Table 7.4.

To examine inter-coder reliability, I considered the agreement between the two coders on the 72 summaries not coded in tandem. This resulted in a high pooled Cohen's $\kappa = .97$, with a code-by-code agreement of 96%. Instances of disagreement were discussed and found to be either coder errors or unanticipated cases that were dealt with differently by the two researchers. In these cases, consensus was reached in a final analysis meeting.

### Identifying important codes

To identify the elements of students' summaries most valued by mathematicians, I first conducted a series of Spearman correlations between each code and the Summary Task scores. A summary of this analysis features in Table 7.5.

Table 7.5 shows that none of the 18 codes was significantly related to Summary Task scores, suggesting that no particular piece of mathematical content

*Table 7.4*

*Coding scheme for the content analysis of the primes proof.*

| Code | Description | Freq |
|---|---|---|
| State Theorem | Restated the theorem. | 22% |
| FToA | Appeal to the Fundamental Theorem of Algebra. | 12% |
| Contradiction | Appealled to the structure of the proof. | 50% |
| Finite primes | Supposed the set of primes to be finite. | 55% |
| Label primes | Labelled the primes using $p_i$, or similar. | 23% |
| Ordering of primes | Explicitly stated $p_1 < p_2 < \cdots < p_k$. | 2% |
| Define n (words) | Stated $n$ is one more than the product of primes. | 56% |
| Define n (symbols) | Included the statement $n = p_1 p_2 \dots p_k + 1$. | 22% |
| $n > 1$ | Observed $n > 1$. | 7% |
| $q \mid n$ | Observed $q \mid n$. | 52% |
| Justify $q > 1$ | justified $q > 1$, knowing that $q$ is prime. | 34% |
| $q = p_i$ for some $i$ | Explicitly stated $p$ is prime. | 23% |
| $q \ \mid \Pi(p_i)$ | Stated that $q$ divides the product of the primes. | 26% |
| $q \mid 1$ | Stated that $q \mid 1$. | 56% |
| Primes not finite | Explicitly stated that the set of primes is not finite. | 5% |
| Conclude infinitude of primes | Explicitly stated that the set of primes is infiniute. | 24% |
| Deduce contradiction | Explicitly stated the contradiction: $q > 1$ and $q = 1$. | 43% |
| Deduce alternative contradiction | Deduce similar but non-identical contradiction. | 27% |

was of greater importance to judges than any other.

To further verify this, I conducted a forced-entry regression model using all content-based codes as possible predictors of comparative judgment-based scores. The choice to include all 18 predictors is arguably problematic, particularly given that there are only 82 summaries in this dataset. Nevertheless, in the absence of robust theoretical reasons to include one over another, this was deemed the only sensible approach despite the risk of over over-estimating the merit of the resulting model. Regardless, even in this case, the resulting model was not significant, $F(18, 63) = 0.51$, $p = .944$ with $R^2 = .33$.

*Table 7.5*

*Correlational analysis of summaries of the primes proof.*

| Code | $\rho$ | $p$ |
|---|---|---|
| State Theorem | .03 | .790 |
| FToA | -.07 | .527 |
| Contradiction | -.34 | .214 |
| Finite primes | .34 | .212 |
| Label primes | .08 | .455 |
| Ordering of primes | .36 | .350 |
| Define n in words | .03 | .803 |
| Define n symbolically | .04 | .698 |
| $n > 1$ | .36 | .345 |
| $q \mid n$ | -.06 | .565 |
| Justify $q > 1$ | .31 | .317 |
| $q = p_i$ for some $i$ | -.03 | .815 |
| $q \ \mid \Pi(p_i)$ | .35 | .377 |
| $q \mid 1$ | .35 | .393 |
| Primes not finite | .05 | .654 |
| Conclude infinitude of primes | .03 | .789 |
| Deduce contradiction | .01 | .908 |
| Deduce alternative contradiction | -.01 | .926 |

## 7.3 Interim discussion

This study was the second in a series of four studies on the Summary Task, each focusing on a different proof. In this chapter, I presented data associated with the primes proof, showing the infinitude of primes. In this interim discussion, I first discuss a problematic feature of many students' summaries, possibly explaining the absence of statistical relationships between Summary Task scores and the content-based coding. I then address the research questions highlighted at the beginning of the chapter. Note that the conclusions drawn here provide only interim commentary, and will be revisited at the conclusion of Chapter 8, after the presentation of two studies focused on a third proof.

### A problematic feature of the primes proof

The content analysis above highlighted a substantive problem with students' summaries of the primes proof. More than one-quarter (27%) of participants deduced an *alternative contradiction.* That is, more than one-quarter produced summaries containing a contradiction argument different from that in the given proof. There are several possible explanations for this, but first, I provide three example excerpts to aid the discussion to come:

Excerpt 1: It assumes $q$ is a prime divisor of $n$, and then proves that $q = 1$, a contradiction.

Excerpt 2: The proof introduces a new prime number and derives a contradiction by showing this number equals 1.

Excerpt 3: [The proof] uses that it has unique prime factorisation that $q$ cannot divide $n$ and is [sic] prime not in the list.

Recall that the original proof deduces a contradiction by demonstrating that, under the assumption that there are finitely many primes, we can produce a number, $q$, that is both equal to 1, and greater than 1. In Excerpt 1, the student's contradiction is based on the fact that 1 is not a prime number. Excerpt 2 does similarly. Excerpt 3 deduces yet another subtly different contradiction, by showing that $n$, the product of *all* primes is not divisible by $q$, a number known to be prime. There are 19 other such examples in the dataset, all identified as straying from the intent of the original text.

I identify three possible explanations for this observation. The first explanation is that these 22 students did not sufficiently understand the proof to produce a summary appealing to the appropriate contradiction. In this case, I would expect to find an inverse correlation between this code and their Summary Task scores. As is shown in Table 7.5, this is not the case and hence I deem this first explanation unlikely.

The second explanation is that students are aware of alternative proofs of the same theorem, either from previous mathematics instruction or popular-science content, and in writing their summaries have strayed toward their memory of a different, more familiar version of the proof. Given that students were asked to summarise the *given* proof, available as they completed the task, the possibility of this explanation is problematic for the aims of this study. It is likely that judges will have approached these 22 summaries from different positions, some deeming these alternative contradictions indicative of poor understanding and others overlooking (or not attending to) the subtle problem with these responses. This may be an explanation for the lack of correlation found throughout this study. I return to this topic in the discussion section.

A third explanation is that students either do not see the difference between their summaries and the original text, or they see the difference and view it as unimportant. In some cases, one could make the case that the students' summaries have improved the original text by streamlining the argument to reach a contradiction more directly. It appears that mathematicians have not punished those producing a different contradiction. This is also worthy of attention. The granularity of the data on this topic is such that any importance mathematicians placed on the distinction is probably lost as a result of the diverse range of alternative contradictions captured by this code. This dataset is not large enough to facilitate meaningful further investigation of these 22 summaries in particular.

### 7.3.1 Research question 2b: What do mathematicians most value when evaluating students' proof summaries?

To address research question 2b, I conducted a content analysis on the students' proof summaries, attempting to understand mathematicians' priorities when evaluating students' proof summaries. The statistical modelling of the content-based scores yielded no significant findings. From the 18-code scheme produced, no code was related to Summary Task scores in isolation, and no significant regression model could be found.

As such, the present study provides limited insight on this topic. I comment on the implications of this finding in addressing research question 3b on the validity of the Summary Task scores.

### 7.3.2 Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

#### Reliability

The Summary Task yielded reliable scores for students' summaries of the primes proof. This is further evidence that the mathematician judges agree on the nature of appropriate proof summaries.

#### Criterion validity

The data in this chapter offered limited evidence for the validity of the Summary Task as a measure of proof comprehension. Of the six measures included as indicators of convergent validity, only the Proof Comprehension Test trended toward significance with $\rho = .23$. While the $p$-value (.034) passed the .05 threshold and the scatter plot showed a modest relationship, this correlation coefficient was not significant under the Holm-Bonferroni correction.

There were two problematic features of the data, possibly leading to underestimates of the relationships between Summary Task scores and the six other measures. The first is the internal reliability of the Proof Comprehension Test, $\alpha = .66$. Without an adequately functioning benchmark measure, it is difficult to draw robust conclusions from the respective comparison. However, I have reason to trust the meaningfulness of the Proof Comprehension Test based on the rigorous development process presented in Mejia-Ramos et al. (2017) and its significant relationship with the Introduction to Proof module. Further, .66 was near the standard threshold of .7, suggesting that the internal reliability was not exceptionally problematic. The second caveat is that of data size. At least in the case of Abstract Algebra (N = 10), it is reasonable to explain the absence of a significant relationship as a product of insufficient data. Similar arguments could be also be made for Linear Algebra (N = 31) and Introductory Calculus (N = 32).

Despite these caveats, when considering the dataset as a whole, this quantitative analysis paints a compelling picture, providing no evidence for the criterion validity of the Summary Task scores, based on the proof demonstrating the infinitude of the primes.

### Content validity

As discussed in answer to research question 3b, the content analysis associated with the primes proof yielded no significant relationships in any form and provided no evidence for the content validity of the Summary Task scores.

Earlier in this chapter, I speculated that there was confusion surrounding the specifics of the contradiction deduced at the end of the proof, and that many students provided summaries with seemingly alternative logical structures to that presented in the original text. I discussed multiple explanations for the source of this confusion, but did not attend to the way the mathematicians may have responded. It is clear that mathematicians did not respond to this aspect of students' summaries in a uniform manner, evidenced by the absence of a relationship between scores and corresponding codes. However, given the theoretical importance of this aspect, the variation in the way mathematicians attended to this topic may have interfered with (or overridden) patterns in other aspects of their judging behaviour.

In the literature review in Chapter 2, I discussed the dependence of agreement amongst mathematicians on their familiarity with the mathematical domain. This was based on the idea that mathematicians tended to agree on 'typical mathematical proof', but differ on peripheral, more esoteric proofs. This gave rise to the notion that such typical proofs would form ideal tasks for comparative judgment-based assessments leading to high reliability, as found in the data here. It seems, however, that I may have chosen a proof that is *too* typical, allowing students to draw on past experiences with nearly identical arguments, leading them astray when constructing summaries of the given text. This is empirically testable by asking students to indicate their familiarity with the proof in question (and perhaps the source of that familiarity). However, this data is not available here and must be left as a topic for future research.

In the previous chapter, I conjectured that mathematicians would reward summaries that highlighted the method of proof and defined the key mathematical objects. This chapter contains no evidence to support this. This is possibly the result of the aforementioned problem of alternative contradictions, or may be a case of over-generalisation from a study in a single mathematical setting. I return to this topic in Chapter 8, after having presented similar data on a third proof.

### Next chapter

The next chapter presents two related studies, focused on a proof of evenness of every third Fibonacci number.

# Chapter 8

# Proof summaries III: Every third Fibonacci number is even

This chapter is the third on the Summary Task, this time based on the *Fibonacci proof,* showing that every third Fibonacci number is even. This is the third proof for which Mejia-Ramos et al. (2017) produced a comprehension test and as in previous chapters, this test is pivotal in the validity analysis to come.

In common with the previous two chapters, the research presented here addresses two of the research questions set out at the end of Chapter 2: 2b) What do mathematicians most value when evaluating students' proof summaries? and 3b) Do proof summaries, scored using comparative judgment, generate a reliable and valid output? By extending this research to a third mathematical context, I continue to build a picture of generality for the merit of this comparative judgment-based approach to proof comprehension assessment.

I present two related studies. The first is a small study, with data collected alongside that presented in Chapters 4 and 6. The second and more substantive study is related to the previous chapter (on the primes proof), containing a substantive overlap in student participants and the same set of measures for investigating criterion validity. The first study considers only convergent validity, while the second larger study also investigates divergent and content validity, similar to those presented in the previous two chapters.

## 8.1 Methods I

### 8.1.1 Materials

The task booklet contained a theorem and its proof, the Summary Task and the Proof Comprehension Test. The theorem and proof are shown in Figure 8.1.

**Definition:** For every natural number we define the $n^{\text{th}}$ Fibonacci number (denoted by $f_n$) as follows:

$$f_1 = 1,$$

$$f_2 = 1, \text{ and}$$

$$f_n = f_{n-1} + f_{n-2} \text{ for all } n > 2 \text{ and } n \in \mathbb{N}$$

**Theorem:** Every third Fibonacci number is even. That is, $f_{3n}$ is even for every $n \in \mathbb{N}$.

**Proof:** Since $f_3 = 1 + 1 = 2$, it is the case that the third Fibonacci number is even. Let $k$ be a natural number, and assume $f_{3k}$ is even. Since $f_{3(k+1)} = f_{3k+2} + f_{3k+1}$ and $f_{3k+2} = f_{3k+1} + f_{3k}$, then $f_{3(k+1)} = 2 \cdot f_{3k+1} + f_{3k}$. Finally, since $2 \cdot f_{3k+1}$ is even and $f_{3k}$ is even, then $f_{3(k+1)}$ is even. Thus, by the principle of mathematical induction, we conclude that for every natural number $n$, $f_{3n}$ is even. □

*Figure 8.1. A proof demonstrating the evenness of every third Fibonacci number.*

### 8.1.2 Student participants

Forty UK-based undergraduate mathematics students participated in this study. All were enrolled in the second-year module 'Mathematical Thinking', covering introductory topics in logic and formal proof. One student declined research access to their data leaving a total of 39.

Participation was made voluntary by allowing students to opt out of the research project, but all students present were required to complete the task booklet as a part of their module study. Participants were told that general feedback would be given to their lecturer based on overall performance. No module credit was awarded for any element of this study.

### 8.1.3 Procedure

Data collection took place during a week-10 lecture. As was the case in Chapter 4, participants were given 30 minutes to complete the booklet, and were asked to complete the Summary Task first before beginning on the comprehension test.

Final module scores were also made available by the course coordinator, allowing a secondary measure for criterion validity analysis. These scores stemmed from a final examination (50%), intermediate coursework (30%) and two in-class tests (20%). Thirty-five of the 39 participants consented to their module scores being included in the final analysis.

### 8.1.4 Comparative judgment

Eight judges were recruited to evaluate this set of proof summaries, all of whom had previously judged in the study presented in Chapter 7. All were current PhD students of mathematics. Judges were asked to read the relevant proof before judging and advised to keep it to hand throughout the process. Judges were compensated based on an assumed rate of 20 seconds per judgment.

Five judges completed 40 judgments each, while the other three completed 80. An error resulted in the data collection initially containing one summary that should not have been included. This error was identified during the judging process but only after the summary in question had received 14 judgments. This summary and the 14 associated judgments were deleted. In sum, 506 judgments were included in the final analysis, distributed across the 39 summaries. Each received between 20 and 24 comparisons with a median time of 20.7 seconds per judgment.

The resulting scores had a mean, $M = 0.00$ ($\sigma = 1.93$).

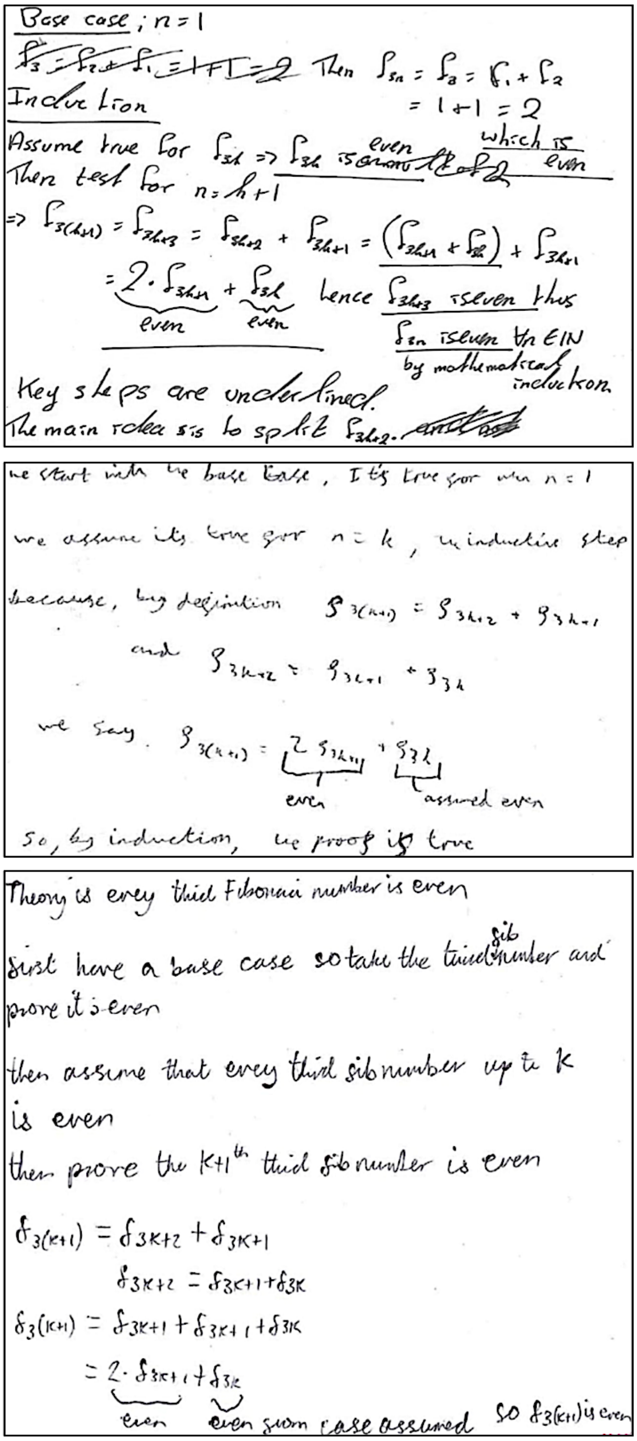
### 8.1.5 Data analysis

Reliability was first evaluated using SSR and inter-rater measures, before conducting a limited evaluation of convergent validity based on the two available measures.

## 8.2 Results I

### 8.2.1 Example responses

To orient the reader, I present the top three summaries as determined by the mathematician judges (see Figure 8.2).



Base case ; n = 1

~~$f_3 = f_2 + f_1 = 1+1 = 2$~~ Then $f_{3n} = f_3 = f_1 + f_2$
$= 1+1 = 2$

Induction

Assume true for $f_{3k}$ => $f_{3k}$ is even ~~multiple of 2~~ which is even

Then test for $n = k+1$

$\Rightarrow f_{3(k+1)} = f_{3k+3} = f_{3k+2} + f_{3k+1} = (f_{3k+1} + f_{3k}) + f_{3k+1}$

$= \underbrace{2 \cdot f_{3k+1}}_{\text{even}} + \underbrace{f_{3k}}_{\text{even}}$ hence $f_{3k+3}$ is even thus $f_{3n}$ is even $\forall n \in \mathbb{N}$ by mathematical induction

Key steps are underlined.

The main idea is is to split $f_{3k+2}$.

we start with the base case, It's true for when $n = 1$

we assume its true for $n = k$, the inductive step

because, by definition $f_{3(k+1)} = f_{3k+2} + f_{3k+1}$

and $f_{3k+2} = f_{3k+1} + f_{3k}$

we say. $f_{3(k+1)} = \underbrace{2 f_{3k+1}}_{\text{even}} + \underbrace{f_{3k}}_{\text{assumed even}}$

So, by induction, the proof is true

Theory is every third Fibonacci number is even

first have a base case so take the third fib number and prove it is even

then assume that every third fib number up to K is even

then prove the K+1th third fib number is even

$f_{3(K+1)} = f_{3K+2} + f_{3K+1}$

$f_{3K+2} = f_{3K+1} + f_{3K}$

$f_{3(K+1)} = f_{3K+1} + f_{3K+1} + f_{3K}$

$= \underbrace{2 \cdot f_{3K+1}}_{\text{even}} + \underbrace{f_{3K}}_{\text{even from case assumed}}$ So $f_{3(K+1)}$ is even

*Figure 8.2. The three proof summaries assigned the highest scores by the comparative judgment-based assessment.*

### 8.2.2 Reliability of comparative judgment-based scores

Internal consistency gave SSR = .91, while inter-rater reliability was $r = .72$, based on 100 iterations. I conclude that the scores from this instantiation of the Summary Task had acceptable reliability.

### 8.2.3 Mathematical Thinking module scores

The 35 available scores for the Mathematical Thinking module ranged from 40% to 100%, with a mean of 70.9% ($\sigma = 12.7$).

### 8.2.4 Proof Comprehension Test

Before focusing on the main body of analysis regarding the Summary Task, it is necessary to explore results from the key benchmark measure.

Students' scores on the 12-item test ranged from 2 to 12 with mean, $M = 7.4$ and standard deviation, $\sigma = 2.4$. This is higher than the two similar tests used in earlier studies.

The internal consistency, $\alpha = .58$, was again lower than expected given the results of Mejia-Ramos et al. (2017). The available data did not pass Bartlett's Sphericity Test (Field et al., 2012) so no principal component analysis was possible. Consistent with previous similar analysis, I proceed with the planned comparative analysis based on this test. However, I do so with caution, noting the necessary caveat associated with the resulting conclusions.

### 8.2.5 Criterion validity

The correlation between the Summary Task and Proof Comprehension Test was in the expected direction but was not significant, $\rho = .09, p = .601$ (see Figure 8.3). On the other hand, the Summary Task did yield a modest correlation with the final module scores, $\rho = .40, p = .022$.

## 8.3 Interim discussion

Consistent with the previous chapter, this small study provides limited evidence supporting the validity of the Summary Task scores in the context of the Fibonacci proof. The only affirmative result was the modest significant correlation with Mathematical Thinking module scores.

Given the relatively small sample size and under-performance of the Proof Comprehension Test, substantive conclusions are difficult to draw based on the

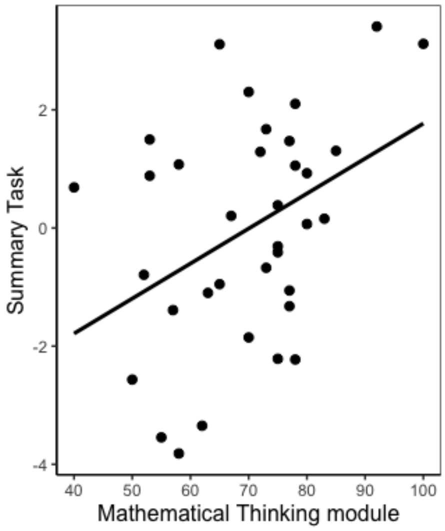


*Figure 8.3. Scatter plot comparing scores from the Summary Task and Proof Comprehension Test.*

results presented thus far. In an attempt to further understand the validity of the Summary Task scores, I now present a larger study based on the same proof, intended to provide a more robust analysis in the same mathematical setting.

## 8.4 Methods II

The method of this study is similar to that presented in the previous chapter on the primes proof. Unless stated otherwise, the methods here are as they appear in Section 7.1. As noted in Chapter 7, the data presented here were collected as part of a wider project with a large research team based at my host university during my second year of study. While the data were collected as part of a larger project, the analysis presented here is solely my own.

### 8.4.1 Materials

As in 7.1.1.

### 8.4.2 Student participants

Sixty-eight students participated in this study, 64 of whom also participated in the study on the primes proof from Chapter 7. All were enrolled in the same Introduction to Proof module.

### 8.4.3 Procedure

As in 7.1.3. Data collection took place in week 11, four weeks after the data collection for the primes proof.

### 8.4.4 Comparative judgment

Judges were recruited via email and personal communication. Due to availability and time constraints, inclusion criteria were relaxed to accept anyone with a university degree in mathematics, down from the post-graduate requirement of previous studies. Ten judges participated, completing between 44 and 100 judgments each resulting in a total of 796 across the 68 summaries. Each summary received between 22 and 27 judgments, with a median time of 11.6 seconds per judgment. This is noticeably shorter than in previous studies on the Summary Task, all having taken a median of greater than 20 seconds per judgment thus far in this thesis. However, 11.6 seconds still appears sufficient for judges to have meaningfully engaged with the task.

The resulting scores had a mean, $M = -0.01$ ($\sigma = 1.78$).

### 8.4.5 Data analysis

As in previous studies, I first analysed the reliability of the Summary Task scores. I then turned to criterion validity, comparing Summary Task scores with the Proof Comprehension Test, SAT scores and the module scores listed in Section 7.1.3. I also compared Summary Task scores from different proofs, based on the 64 participants who completed the Summary Task (and Proof Comprehension Test) for both the primes and Fibonacci proofs. Finally, I considered content validity in the same format as present in previous chapters.

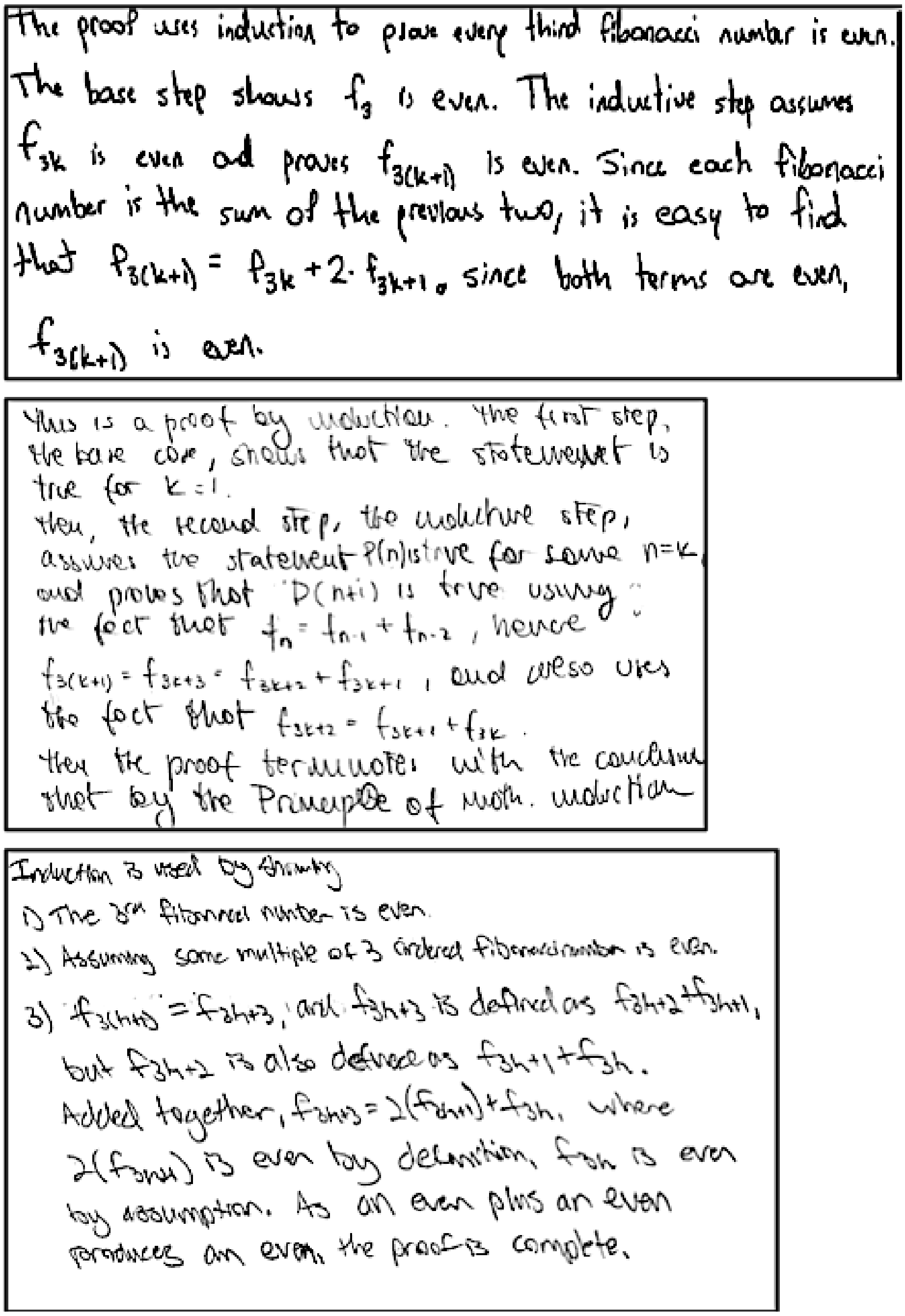
## 8.5 Results II

### 8.5.1 Example responses

Figure 8.4 shows the top three summaries from this study.

### 8.5.2 Reliability of Summary Task scores

In this study, internal consistency gave SSR = .87, with inter-rater reliability $r = .72$, based on 100 iterations. Despite the unusually short time taken by judges here, these judgments appear reliable, consistent with all other expert-based judgments presented in this thesis.



The proof uses induction to prove every third fibonacci number is even. The base step shows $f_3$ is even. The inductive step assumes $f_{3k}$ is even ad proves $f_{3(k+1)}$ is even. Since each fibonacci number is the sum of the previous two, it is easy to find that $f_{3(k+1)} = f_{3k} + 2\cdot f_{3k+1}$. since both terms are even, $f_{3(k+1)}$ is even.

This is a proof by induction. The first step, the base case, shows that the statement is true for $k=1$.
Then, the second step, the inductive step, assumes the statement $P(n)$ is true for some $n=k$, and proves that $P(n+1)$ is true using the fact that $f_n = f_{n-1} + f_{n-2}$, hence $f_{3(k+1)} = f_{3k+3} = f_{3k+2} + f_{3k+1}$, and also uses the fact that $f_{3k+2} = f_{3k+1} + f_{3k}$.
Then the proof terminates with the conclusion that by the Principle of math. induction

Induction is used by showing
1) The 3rd fibonacci number is even.
2) Assuming some multiple of 3 ordered fibonacci number is even.
3) $f_{3(h+1)} = f_{3h+3}$, and $f_{3h+3}$ is defined as $f_{3h+2} + f_{3h+1}$, but $f_{3h+2}$ is also defined as $f_{3h+1} + f_{3h}$. Added together, $f_{3h+3} = 2(f_{3h+1}) + f_{3h}$, where $2(f_{3h+1})$ is even by definition, $f_{3h}$ is even by assumption. As an even plus an even produces an even, the proof is complete.

*Figure 8.4. The top three summaries from the second iteration of the Summary Task with the Fibonacci proof.*

### 8.5.3 Proof Comprehension Test

All 68 participants completed the 12-item Proof Comprehension Test. Scores ranged from 2 to 12, with a mean $M = 6.8$. The internal consistency of the test was again investigated using Cronbach's $\alpha = .75$. This was the only time across the three proofs that one of the Proof Comprehension Tests from Mejia-Ramos et al. (2017) passed the standard threshold for internal consistency.

### 8.5.4 Mathematics SAT scores

Mathematics SAT scores were available for 52 of the 68 participants, 50 of whom also participated in the study on the primes proof. For this reason, the statistics presented here, and in the module scores, are similar to those presented in the previous chapter. Mathematics SAT scores had a mean, $M = 700$ ($\sigma = 66$) and a range of 510 to 800.

### 8.5.5 Module scores

As before, I had access to five sets of module results, each taken by a variable number of students. Grades assigned in each module have been mathematised using the university's convention: A = 4, B+ = 3.5, B = 3, C+ = 2.5, C = 2, D = 1, F = 0.

#### Introduction to Proof

All 68 students sat the exam for this module, and hence had a final module score available for analysis. These scores had a mean of 2.98, just short of a B, with a mode of 3.5.

#### Introductory Calculus

Twenty-one of the 68 participants sat this module. One student attempted it twice and their first attempt was removed from the data. For the 21 available responses, the mean was 3.33, with a mode of 4.

#### Further Calculus

Thirty-six scores were available for Further Calculus. These had a mean of 3.03 and mode of 3.5.

#### Linear Algebra

Twenty-five scores were available for Linear Algebra, with a mean of 2.86 and a mode of 3.

#### Abstract Algebra

Eleven scores were available for this module, with a mean of 2.59 and a mode of 2.

### 8.5.6 Analysis of criterion validity

#### Isolated Spearman correlations

I first present a series of Spearman correlations, investigating the independent relationships between the Summary Task and each measure, see Table 8.1.

*Table 8.1*

*Comparing the Fibonacci Summary Task with other measures.*

| Measure | $N$ | $\rho$ | $p$ |
|---|---|---|---|
| Proof Comprehension Test | 68 | .34 | .005* |
| Mathematics SAT | 52 | .20 | .351 |
| Intro to Proof | 68 | .33 | .274 |
| Introductory Calculus | 21 | .09 | .683 |
| Further Calculus | 36 | .22 | .393 |
| Linear Algebra | 25 | .08 | .702 |
| Abstract Algebra | 11 | -.08 | .823 |

*Note.* Significance determined using the Holm-Bonferroni method with initial $\alpha = .05$.

Only the Proof Comprehension Test demonstrated a significant correlation (see Figure 8.5). This is consistent with findings from the previous chapter[1]. Notably, this same relationship was not found in the smaller study presented in this chapter. I expect that this absence is a result of data size, although further research is necessary to explain the discrepancy.

Again following the analysis from the previous chapter, I consider the relationships between the Proof Comprehension Test and other measures. See Table 8.2.

[1] It is regrettable that the size of these data is not sufficient to justify a full regression analysis to predict comparative judgment-based scores with these seven measures, as was initially planned. However, the findings presented in Table 8.1 are adequate for the purposes of this analysis.

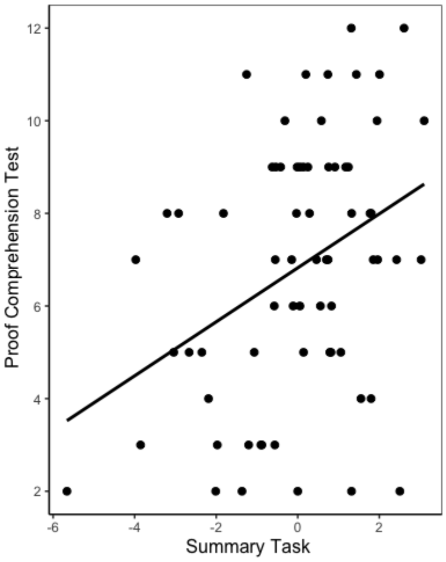


*Figure 8.5. Scatter plot comparison of performances on the and 12-question Proof Comprehension Test.*

As in the previous chapter, Table 8.2 shows a significant relationship between the Proof Comprehension Test scores and the module from which students were recruited. Again, we see that other measures, while not significant under the Holm-Bonferroni correction, meet a .05 threshold. While the specifics of these relationships are unclear, Tables 8.1 and 8.2 demonstrate that the more general measures of mathematical expertise (SAT and module scores) were more closely related to the Proof Comprehension Test than they are to Summary Task scores. I return to this topic and its implications for the validity of the Summary Task in Chapter 9.

### 8.5.7 Content analysis

As in previous chapters, this section presents an analysis of the specific content students included in their summaries of the proof demonstrating the evenness of every third Fibonacci number. This analysis leads to a deeper understanding of content validity of the Summary Task scores as a measure of proof comprehension, as well as providing further insights into the types of summaries most valued by the mathematicians in this context. At the end of this chapter, I review the findings from the previous three chapters, seeking patterns across the three proofs.

*Table 8.2*

*Comparing the Fibonacci Proof Comprehension Test with other measures.*

| Measure | $N$ | $\rho$ | $p$ |
|---|---|---|---|
| Mathematics SAT | 52 | .32 | .019 |
| Intro to Proof | 68 | .46 | <.001* |
| Introductory Calculus | 21 | .38 | .094 |
| Further Calculus | 36 | .40 | .016 |
| Linear Algebra | 25 | .33 | .303 |
| Abstract Algebra | 11 | .39 | .234 |

*Note.* Significance determined using the Holm-Bonferroni method with initial $\alpha = .05$.

### Developing a coding scheme

The first iteration of this coding scheme was developed by considering a set of 10 summaries, noting patterns and clusters of response types, which led to the generation of an initial list of 13 codes. As in Chapter 6, I intended to preserve the largest possible dataset for inter-coder reliability analysis. To this end, the initial scheme was developed using 10 of the 39 summaries discussed in the earlier half of this chapter.

I then involved the same academic supervisor who also participated in the coding of summaries of the primes proof. Together, we implemented the existing coding scheme on a further 10 summaries (also from the earlier analysis). The necessary revisions included removing or combining multiple codes, resulting in a list of 12 codes to be implemented by both researchers on the full set of 68 summaries included in this study. The reliability of this process was examined by calculating a pooled Cohen's $\kappa = .92$ with a code-by-code agreement of 89%.

While this was deemed acceptable, three codes appeared to be substantive outliers, all with isolated $\kappa$ less than .63. A pooled Cohen's $\kappa$ was calculated based on the other nine codes, yielding $\kappa = .98$ with a code-by-code agreement of 95%.

Two of these three codes regarded references to the base case of the proof: 'States the base case' and 'Computes the base case'. The descriptions of the latter code required the summary to explicitly perform some version of the calculation $f_3 = f_2 + f_1$, while the former code required only that the summary stated some version of '$P(f_3)$ is true'. The third outlying code regarded references to the definition of $k$ (the iterating variable in the induction) as a member of the integers. These three codes were discussed in a final analysis meeting with the goal of reaching full code-by-code agreement to facilitate further analysis.

The confusion regarding the pair of base case-related codes was solved by combining the two codes. A new code, labelled 'base case' was assigned to any summary that had been awarded either of the codes being replaced. After making this edit to both researchers' codebooks, the new code generated only one case of disagreement that turned out to be coder error and was easily resolved.

Regarding the problematic code referring to the iterator $k$, a discussion between researchers revealed that one coder had misapplied the code in 16 cases. Both researchers agreed to a final codebook without requiring edits to this, or any other code.

The final 11-code scheme is presented in Table 8.3, alongside the frequencies with which the codes were awarded.

*Table 8.3*

*Coding scheme for the content analysis of the Fibonacci proof.*

| Code | Description | Freq |
|---|---|---|
| Definition | Includes any re-statement of the definition of the Fibonacci sequence. | 2% |
| States theorem | Restates the theorem. | 37% |
| Assumes result | Erroneously assumes the result. E.g. 'Assume that $f(3k)$ is even for all $k$' (must be obviously distinct from assuming the inductive hypothesis). | 2% |
| Base case | Any explicit reference to the base case or verification thereof. | 49% |
| Induction | Introduces the method of proof. | 40% |
| Defines $k$ | Any explicit indication that k is natural. | 11% |
| Assume $f_{3k}$ even | Must be reasonable to believe this refers to a distinct. $k$ (i.e. not assuming the statement of the theorem). | 45% |
| Computes $f_{3(k+1)}$ | Must be either algebraic, or natural language clearly reduceable to symbolic notation. | 29% |
| Deduces $f_{3(k+1)}$ is even | Describes the reasoning *without* performing the calculation. E.g. (By assuming that $f_{3k}$ is even for some $k$, we find that...). | 46% |
| Conclusion (induction) | E.g. 'By PMI, we conclude...'. | 17% |
| Conclusion (generic) | Need not be specific. Accept all of the following: 'Hence, $f_{3n}$ is even for all $n$', 'Hence, $f_{3n}$ is even' and 'Hence, the statement holds/is true'. | 20% |

*Table 8.4*

*Regression modelling for the Fibonacci proof Summary Task scores.*

| Code | Correlation coefficients | | Regression model | | | |
|---|---|---|---|---|---|---|
| | $\rho$ | $p$ | $B$ | SE | $\beta$ | $p$ |
| Definition | -.08 | .495 | | | | |
| States theorem | .02 | .845 | | | | |
| Assumes result | .31 | .368 | | | | |
| Base case | .36 | .003* | 1.31 | 0.28 | 0.37 | <.001* |
| Induction | -.20 | .310 | | | | |
| Defines $k$ | .32 | .009 | | | | |
| Assume $f_{3k}$ even | .56 | <.001* | 0.33 | 0.36 | 0.09 | .725 |
| Computes $f_{3(k+1)}$ | .66 | <.001* | 1.98 | 0.29 | 0.62 | <.001* |
| Deduces $f_{3(k+1)}$ even | .60 | <.001* | 1.20 | 0.34 | 0.34 | <.001* |
| Conclusion (induction) | .30 | .434 | | | | |
| Conclusion (generic) | .20 | .306 | | | | |

*Note.* Codes with a significant Spearman coefficient were entered into a forced-entry regression model as predictors of Summary Task scores. Significance was determined using the Holm-Bonferroni method with initial $\alpha = .05$. The resulting four-code model, $F(3, 63) = 33.78, p < .001$, explained 68% of the variance.

### Identifying important codes

Here, I present a series of correlation-based analyses, attempting to understand the aspects of students' summaries most heavily rewarded in mathematicians' judgments. To construct a regression model predicting comparative judgment-based scores using only the most relevant codes, I first computed a series of Spearman correlations between the Summary Task scores and each content-based code. This preliminary analysis is presented in Table 8.4, alongside a forced-entry regression analysis based on the four codes identified as significant in isolation, $F(3, 63) = 33.78, p < .001$, $R^2 = 68.2$.

Noting the conservative nature of the Holm-Bonferroni correction, I also computed a forced-entry regression model including the code 'Defines $k$'. This gave a model also explaining 68% of the variance, suggesting that this code did not substantively contribute to the model's prediction of Summary Task scores, leading to the conclusion that 'defining $k$' was not predictive of Summary Task scores.

From the final column of Table 8.4, on page 150, we see three significant codes in the final model: *Base case*, *Computes* $f_{3(k+1)}$ and *Deduces* $f_{3(k+1)}$ *even.* These three codes are all related to the method of proof. This may seem unsurprising. In summarising any proof it is likely necessary to attend to each of its constituent parts. In the context of mathematical induction, this is most often referred to as the base case and the inductive step (demonstrating the inductive hypothesis '$P(k) \implies P(k+1)$'). Moreover, this is consistent with the speculation from Chapter 6 regarding the importance of the method of proof and the key mathematical objects.

On the other hand, I had expected mathematicians to reward those explicitly naming 'proof by induction', and those including a final structuring statement of the form '...hence we know by the principle of mathematical induction...'. While respectively 40% and 17% of students' summaries were awarded the 'induction' and 'conclusions (from induction)' codes (see Table 8.3), neither yielded a significant relationship with the Summary Task scores in this study.

In the final section of this chapter, I address the cumulative findings from the most recent three chapters, based on evidence from all three proofs. I present a final piece of quantitative analysis, comparing performance on the Summary Task and Proof Comprehension Tests for the Fibonacci and primes proofs.

### 8.5.8 Comparing performance from the primes and Fibonacci proofs

Given the overlapping participants from this study and that presented in the previous chapter (on the primes proof), I now compare performance from the 64 students who completed the Summary Task and the Proof Comprehension Test associated with each proof. In doing so, I gain insight into the dimensionality of proof comprehension as a potentially singular construct, as well as further understanding of the validity of the Summary Task scores.

First, I compared the Summary Task scores from the two proofs using a Pearson correlation coefficient[2], $r = .39, p = .001$ (see Figure 8.6a).

I also examined the relationship between the two Proof Comprehension Tests, and found a significant Spearman correlation $\rho = .63$, $p < .001$(see Figure 8.6b). This provides further evidence for the claim that, at least in this pairwise case, these measures of proof comprehension measure related competencies.

This provides reasonable evidence that proof is a singular construct and in particular, that understanding one proof probably predicts understanding of others. However, as was discussed by Jones et al. (2015), comparative judgment-based scores are often statistically related, regardless of the content of the two tasks. Hence, further investigation is necessary to understand the relationship

[2] I use Pearson not Spearman, here, as I am now dealing with two normally distributed, continuous variables. For completeness, I report that Spearman's coefficient and associated $p$-value differed from the Pearson values presented $\varepsilon < 10^{-3}$.

between comprehension of the two proofs.

To this end, I compared the Summary Task scores from each proof with the Proof Comprehension Test from the other. When comparing Summary Task scores for the Fibonacci proof with the Proof Comprehension Test from the primes proof, I found $\rho = .35, p = .004$( see Figure 8.6c). I interpret this as evidence that my assumption of unidimensionality was valid (at least across these two settings), and that the scores assigned to students' summaries of the Fibonacci proof are valid measures of that one-dimensional construct. On the other hand, when doing the same for the summaries of the primes proof, I found no such significant relationship with the Proof Comprehension Test from the Fibonacci proof, $\rho = .36$, $p = .204$ (see Figure 8.6d).

Consistent with the analysis from the previous chapter, this is further indication that the validity demonstrated by the primes summaries is, at best, questionable either as a measure of comprehension for the specific proof in question, or of proof comprehension more generally.

## 8.6 Discussion of research on the Summary Task

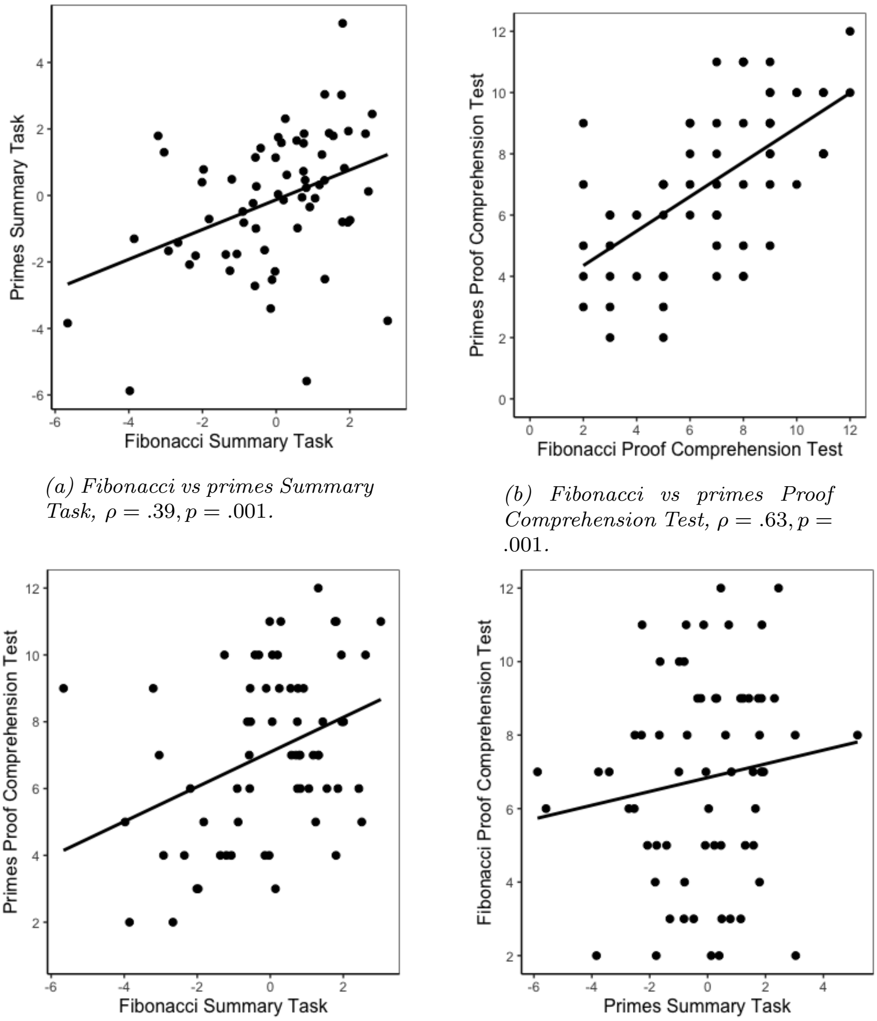
Between this and the preceding two chapters, I have presented four studies based on three proofs. This work addressed two of the research questions outlined in Chapter 2. In considering the cumulative findings across these four studies, I address each of these questions in turn.

The implications of this work are reserved for the final chapter, after presenting a final empirical study in Chapter 9.

### 8.6.1 Research question 2b: What do mathematicians most value when evaluating students' proof summaries?

In addressing this question, I first address the content-based regression modelling in Chapters 6 and 8, based on the uncountability and Fibonacci proofs and the associated content analysis of their respective summaries. Chapter 7, on the primes proof, also addressed this question but yielded no statistically significant findings. The absence of such findings is important in its own right and is addressed next.

Judges appeared to be consistent in the features they rewarded for summaries of both the uncountability and Fibonacci proofs. In broad strokes, these were references to the method of proof and the key mathematical objects used to execute the relevant method.



*(a) Fibonacci vs primes Summary Task, $\rho = .39, p = .001$.*

*(b) Fibonacci vs primes Proof Comprehension Test, $\rho = .63, p = .001$.*

*(c) Fibonacci Summary Task vs primes Proof Comprehension Test, $\rho = .35, p = .004$.*

*(d) Primes Summary Task vs Fibonacci Proof Comprehension Test, $\rho = .36, p = .204$.*

*Figure 8.6. Scatter plot comparisons of students' performance on the Summary Task and Proof Comprehension Tests associated with the primes and Fibonacci proofs.*

There was, however, variation in the codes identified as most significant. These appear to be influenced by the idiosyncratic features of the particular proofs at hand. For example, recall that the Fibonacci proof relied on the principle of mathematical induction to demonstrate the evenness of every third Fibonacci number. I found that mathematician judges most heavily rewarded summaries referencing the base case and the inductive hypothesis, but *not* those directly stating that the proof procedes by the principle of mathematical induction. On the other hand, judges of summaries of the uncountability proof rewarded those explicitly stating that we proceed via contradiction.

I conjecture that this variation is a function of the complexity of the respective proofs. For the Fibonacci proof, explicit mention of the proof method was likely deemed unnecessary by many judges, given that there is no other candidate method for proofs containing references to base cases and inductive hypotheses. On the other hand, the uncountability proof can be seen as more complex, wherein the method of proof by contradiction can easily be lost in the detailed technical work required to deduce the relevant conclusions. Each statement of the uncountability proof, viewed in isolation, could belong to any number of arguments structured by a number of different proof methods. Hence, it is this complexity that I speculate prompted judges to reward those explicitly stating the method of proof. This speculation is anecdotally supported by excerpts from the interview study, reported in the following chapter, including one judge who reported making a decision on the basis that one of the summaries had concluded the precise negation of the theorem's conclusion.

In contrast to the relative clarity of the regression modelling associated with the uncountability and Fibonacci, I now consider the primes proof. A similar analysis for summaries of the primes proof yielded no significant Spearman coefficients and no significant regression model, suggesting that the clarity of the conclusions above require further consideration.

The primes proof is also a constructive proof by contradiction, similar to the uncountability proof. Hence, based on the discussion above, one would expect that summaries referencing the construction of the relevant number, in this case $N = p_1p_2p_3 \dots p_k + 1$, should have scored better than those without such a reference. Similarly, those explicitly stating the method of proof, or deducing the appropriate conclusion, would also be expected to have scored highly. This was not the case. As was discussed in Chapter 7, the study of the primes proof appeared to suffer from unanticipated methodological issues possibly resulting from students' familiarity with the theorem and associated proofs.

This study of the primes proof should not be used to reject earlier conjectures regarding the method of proof and key mathematical objects. On the other hand, further research is necessary to verify their scope and to understand the source of the failure of the study of the primes proof. This leads to the next research question on the validity of the Summary Task scores.

### 8.6.2 Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

I address the two components of this question separately, beginning with the reliability of the scores produced.

#### Reliability

In all four studies, the reliability of the comparative judgment data was deemed acceptable. Scale Separation Reliability was at least .86 in all cases, and inter-rater reliability was always above .72. This is sufficient evidence to suggest that this robust reliability is likely to generalise beyond the three proofs investigated.

Further, insofar as these measures capture agreement amongst judges, I conclude that while mathematicians perhaps do not always agree on the nature of proof itself, there is ample agreement about what mathematicians expect their undergraduate students to say and write in such settings. I return to this topic in Chapter 10, given its implications for the literature on mathematicians' conceptions of proof.

In all studies, judges had a minimum of an undergraduate degree in mathematics. Beyond this, there was variation in the qualifications and current employment of the judging cohorts. Unlike in the work on proof conceptions, this variation was born in pragmatism, rather than a concentrated desire to understand the consequences of this variation. Regardless, the resulting findings are worthy of comment.

In particular, the study of the Fibonacci proof involved judges with a degree-level qualification in mathematics. I did not require that these judges were still actively engaged with mathematics, nor that they continued to postgraduate studies in mathematics or other disciplines as in studies presented earlier in this thesis. Despite these looser criteria, the reliability of the data was as robust, and strikingly, the resulting analysis yielded good evidence for the criterion validity of the resulting scores. Further, a content-based regression model explained 68% of the variance in the data. Beyond the more precise exploration of judge qualifications in the earlier half of this thesis, I conclude that the quality of the judgments appeared not to be sensitive to judging expertise. While it is highly likely that some level of technical expertise is necessary, further research is required to understand the requisite level and breadth.

### Validity

In evaluating the validity of the Summary Task scores, I considered criterion and content validity for each of the three proofs.

In considering criterion validity, two possible purposes could be considered target domains: localised proof comprehension and general mathematical performance. In considering localised proof comprehension, I evaluated validity through comparisons with the multiple-choice Proof Comprehension Tests from Mejia-Ramos et al. (2017). The Summary Task and Proof Comprehension Test scores showed significant relationships in two of the four studies, associated with the uncountability proof and the larger study on the Fibonacci proof. The study based on the primes proof also trended toward significance with $p < .05$ but was not statistically significant under the Holm-Bonferroni correction. The smaller study on the Fibonacci proof showed no significant relationship.

Given the methodological problems with the primes proof, and the small dataset in the smaller Fibonacci study, I place less emphasis on these unsuccessful results than on significant results from the study of the larger studies on the uncountability and Fibonacci proof. Hence, I conclude that there is substantive evidence for the criterion validity of Summary Task scores in some contexts, but that further research is necessary to clarify the scope of the applicability for the Summary Task.

In considering the validity of scores as a more general measure of mathematical expertise, Summary Task scores were compared to a series of module scores, based on traditional assessments, and mathematics SAT scores. None of these associated measures showed substantive evidence of a significant relationship, indicating that while the Summary scores demonstrated substantive validity as a measure of localised proof-specific comprehension in at least some cases, they are not reflective of general mathematical performance.

In considering content validity, I draw on my earlier answer to Research Question 2b. For two of the three proofs, content-based regression modelling yielded significant results indicating, as one might expect, that important features of proof summaries include the method of proof and the introduction of key mathematical objects. As was noted in Chapter 6, these findings are consistent with the literature on proof comprehension assessment. In particular, the proof comprehension assessment model of Mejia-Ramos et al. (2012), summarised in

Table 2.2, includes the following three aspects: ‘Logical status of statements and proof framework’, ‘Summarising via high-level ideas’, and ‘Identifying the modular structure’. These three aspects of proof comprehension are reflected in the statistical modelling associated with both the uncountability and Fibonacci proofs (but not the primes proof). Hence, I conclude that the associated content analysis of students’ summaries showed an acceptable level of evidence for the validity of the resulting scores in two of three settings.

That said, as noted earlier, data were collected in only three mathematical settings, one of which had clearly problematic features. This research features many findings providing cause for optimism about the validity of Summary Task scores in other mathematical settings. However, the generality of these findings remains an open question and further work is required to understand the nature of proofs for which the Summary Task is applicable.

Topics for such future research are reserved, alongside the implications of this work, for the final chapter of this thesis after presenting a final empirical study based on interviews with mathematician judges.

### Next chapter

Until now, the analysis of the Summary Task has been quantitatively driven, even if incorporating qualitative data in places. These analyses have established a robust understanding of the criterion and content validity of the resulting scores, with some yielding preliminary conjectures on judges’ priorities in evaluating students’ written summaries. However, the insights into judges’ decision-making are inferred from the numerical output of the comparative judgment process.

The study presented in the next chapter brings the judges to the forefront of the analysis, using a series of semi-structured interviews and a think-aloud protocol focused on understanding the motivations driving judges’ decision-making processes.

# Chapter 9

# Proof summaries IV: Interviewing judges

It is well-established that, even under assessment protocols with clear criteria, there is wide variation in the features to which assessors attend when evaluating students' work (Bloxham et al., 2016). In the previous chapter, I set out preliminary answers to research questions 2b (on mathematicians' values when evaluating students' proof summaries) and 3b (on the reliability and validity of the resulting Summary Task scores). Regarding reliability, the statistical evidence suggests that judges appear to agree on what makes a good proof summary. However, it remains unclear which features drive that agreement and the extent to which one can identify the features of students' summaries that mathematician judges value most. From the content analyses in each of the three previous chapters, I generated multiple conjectures regarding mathematicians' values. These were based on regression modelling of the relationship between content-specific features and scores generated by the comparative judgment process.

In this chapter, I present an interview-based study with mathematician judges, focused on further understanding their decision-making processes when evaluating students' proof summaries. The design of this study was based on Pollitt and Murray (1993), who evaluated the validity of their language proficiency assessment by asking judges to state their justifications for each pairwise decision. By requiring judges to 'think aloud', I provide a different perspective on the research questions addressed at the end of the previous chapter. In doing so, I am able to triangulate conclusions from differing perspectives (Chapters 6, 7 and 8).

The interviews presented in this chapter were guided by four focus questions, used to structure the data collection process:

- How do judges process the information necessary to make their decisions?
- What specific mathematical content do judges attend to?
- What features of a summary do judges attend to?
- What non-content-related features influence judges' decision-making?

## 9.1 Methods

### 9.1.1 Participants

I recruited nine judges from the same English university, referred to as J1 to J9 throughout. All judges were active researchers in mathematics or mathematics education, holding at least a Master's degree in mathematics or a related discipline.

### 9.1.2 Materials

The proof summaries used in this study all focused on the uncountability proof, presented in Chapter 6. A subset of the summaries collected for the earlier study was reused here. Twenty pairs were chosen at random from the 130 (non-blank) summaries available.

### 9.1.3 Procedure

Each interview lasted between 40 and 60 minutes, and comprised three tasks: a series of 20 judgments, a semi-structured interview and a think-aloud re-evaluation of a subset of the original judgments.

Task 1 saw participants complete a series of 20 judgments in a laboratory setting, verbalising their decision ('left' or 'right') at the same time. All nine judges saw the same 20 pairings in the same order. This is atypical for comparative judgment research but was done, as in Pollitt and Murray (1993), to facilitate comparison between judges.

Task 2 was a semi-structured interview. The list of questions guiding these interviews can be found in Appendix B.

Task 3 was a think-aloud protocol wherein judges reviewed the final 10 pairings. During this task, judges were asked to explain 'how' and 'why' they made their decisions. The interviewer also drew attention to several predetermined features of particular summaries to elicit comments regarding attitudes to explicit errors and abuses of notation. In Task 3, the interviewer emphasised that it was not important that judges replicated their earlier decision, but rather that the participant attempted to replicate or recall elements of their original decision-making process.

The order of these three tasks was chosen to optimise the ecological validity of the original decisions and to minimise the influence of the interviewer's presence. By conducting the semi-structured interview before the think-aloud task, I minimised the influence of the interviewer's responses with the biases and preconceptions inevitably communicated through the back and forth of the think-aloud task.

Both participant and interviewer were audio-recorded during Tasks 2 and 3. The decisions from Task 1 were also recorded to facilitate an elementary reliability analysis on this small set of decision data.

### 9.1.4 Data analysis

The primary purpose of this study was a thematic analysis of judges' verbal decision-making process. This was done using the six-stage process described by Braun and Clarke (2006), see Table 9.1, and an additional stage, proposed by Attride-Stirling (2001), in writing analytical summaries for each participant (see Chapter 3 for a justification of this approach).

I first transcribed all nine interviews in full. I then read and re-read these transcripts, making informal notes on potential codes and themes. This resulted in an initial list of 45 codes (see Appendix C). A more structured third reading of the data followed, in which these codes were then systematically assigned to all transcripts, attempting to collate all data relevant to each of the 45 codes.

At each stage of the analysis, codes corresponded to latent themes (Braun and Clarke, 2006) based on an assumption of a shared understanding of meaning between researcher and participant. This is consistent with an essentialist epistemology and an understanding that one 'can theorise motivations, experience, and meaning in a straightforward way, because a simple, largely unidirectional relationship is assumed between meaning and experience and language' (*ibid*, p. 91).

*Table 9.1*

*Thematic analysis, as proposed by (Braun and Clarke, 2006).*

| Phase | Description |
|---|---|
| Familiarise yourself with the data | Transcribing data (if necessary), reading and re-reading the data, noting down initial ideas. |
| Generating initial codes | Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code. |
| Searching for themes | Collating codes into potential themes, gathering all data relevant to each potential theme. |
| Reviewing themes | Checking the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic map of the analysis. |
| Defining and naming themes | Ongoing analysis to refine the specifics of each theme, and the overall story the analysis tells; generating clear definitions and names for each theme. |
| Producing the report | The final opportunity for analysis. |

After generating a list of initial codes, I constructed a series of analytical summaries for each judge following Attride-Stirling (2001), exemplified below:

> J6: *Self-identified positive marker. Looked to aggregate valid arguments to see if they 'add up to an appropriate summary'. Clear decision-making strategy in comparing responses. Valued brevity and did not require completeness or detailed summaries if overview was well phrased. If poorly phrased, referred to notational detail or accuracy to make decisions. Relied on negative marking in describing decisions on several occasions*
>
> J7: *Self-identified negative marker. Drew on previous experience with comparative judgment to clearly outline efficient methods for making decisions. Stopped reading when an egregious error was found. Skim reads both scripts to see if the judgment can be made very quickly, only reading the more detailed aspects if necessary. Was put off by poor use of notation or other mathematical fluency (e.g. describing and comparing sets as functions). Detail-oriented. Wanted to see evidence of notation. Completeness is important.*

These summaries served as a secondary tool for me to come to understand the data. The process of producing these analytical summaries informed the generation of initial themes, based on the repetition of topics naturally addressed in the set of nine summaries.

### Summarising the data

After defining and exemplifying each theme, I then provide brief numerical summaries indicative of the prevalence of each sub-theme in the data. I define prevalence as the number of judges who provided at least one related utterance, hence quantities are expressed at the participant level in the form 'sub-theme $X$ was identified in the transcripts of $N$ judges'. This intends to provide the reader with a sense of the data and should not be interpreted as a summation of the relative density. The number of utterances was deemed less important than the existence of the belief/view/approach in the data. Moreover, with only nine judges and a range of possible counting techniques, all of them limited, I elect to include these numerical summaries only to provide the reader with the informal sense that the themes identified are more substantive than outliers uttered once by one individual.

### Reliability

Given the relatively small sample size (nine judges), it is also desirable to understand how representative these judges are of the wider mathematical community. In traditional comparative judgment studies, this would be investigated using statistical reliability, using the measures of Scale Separation Reliability (SSR) and/or Bisson et al.'s (2016) split-half inter-rater reliability, as reported in previous chapters. However, both require more decisions that were collected in this study and are not applicable in this case.

As a replacement, I consider percentage-agreement between judges, and compare the decisions of the interviewed judges with the scores assigned to each of summary in an earlier study on the same dataset (Chapter 6).

### Validation of analysis

A first, informal version of this analysis was presented at a research meeting of the Mathematical Pedagogy Group at Loughborough University. This was an hour-long workshop in which I presented a series of excerpts, and prompted discussion on the appropriateness of the assigned codes. This resulted in no large-scale changes to the analysis but provided valuable feedback on how best to communicate the analysis and how to characterise relationships between various themes and sub-themes.

## 9.2 Results

I identified six themes in the data, each regarding different aspects of judges' behaviour in making their decisions. A summary with associated sub-themes is presented in Table 9.2.

*Table 9.2*

*Summary of thematic analysis.*

| Theme | Sub-themes |
|---|---|
| Reading strategies | Full reading<br>Benchmarking<br>Error-seeking |
| Approaches to assessment | Positive marking<br>Negative marking |
| Influencing features | Technical details<br>Brevity<br>Mathematical precision<br>Mathematical fluency |
| Necessary content | Construction of $b$<br>Construction of $f$<br>Proof by contradiction |
| Unnecessary content | 0s and 9s<br>(0,1) is infinite |
| Non-content-related features | Handwriting/readability<br>Arbitrary decision-making<br>Content-dependence |

In all cases, the evidence either came from self-reported behaviours and tendencies (drawn from the semi-structured interview), or from the think-aloud task, where values or justifications were inferred from the observed behaviours and tendencies.

### 9.2.1 Reading Strategies

This theme captures judges' behaviour when taking in the information present on each screen. I identified three sub-themes: *full reading*, *benchmarking* and *error-seeking*. While most judges consistently began with the same strategy for each decision, many used more than one strategy in the interview, sometimes on the same decision.

### Full reading

The full reading strategy is the most thorough approach, wherein a judge performed a detailed reading of both summaries from start to finish. This is best exemplified by J2's description of how they approached their decisions: *'So first I read the left, then the right, then try to keep them both, sort of, in my head until I decide'*. This strategy is characterised by the absence of any short-cuts and involves a detailed reading of both summaries before committing to a decision.

### Benchmarking

Judges using a benchmarking strategy reported performing a detailed reading of one summary, then assigning it some qualitative label to be used as a benchmark against which to evaluate the other summary. This strategy is more efficient than full reading, providing a judge with a mechanism for reaching a decision without processing all of the available information.

As was noted by J6, this strategy is not always useful. However, it appears to quicken the decisions for which there is an obvious gulf in quality. J6 noted that

> *'...usually, I have some kind of a decision or rating or something when I was finished reading the left one, where I'd try to then decide if the right one was better or worse than that. So there was definitely the first one was like a benchmark. And I read the right one, which I usually only had to decide if it's better or worse. And then, only in the cases where that wasn't very obvious, I went back to read the left one again'*.

While not explicit in the wording from J6, it appears that when J6 reads the second summary, they are more ready to jump to an immediate conclusion than those using a full-reading strategy. This is echoed by J8, who commented that *'. . . usually the first one you see you are more critical [of]. The second one is only interesting in so far as it is compared to the first. . . '*.

### Error-seeking

Error-seeking judges skim-read both summaries, looking for errors or omissions. In many cases, the judge identified a problematic aspect of one summary and would subsequently choose the other. This final reading strategy is arguably the most efficient of the three, and is indicative of other dispositions discussed in later themes.

For example, J7 reported beginning each comparison by checking

> *'to see if one is complete nonsense. Sometimes a student will write just at the very start and then give up and that usually means if the [other] gets any further than that's an instant [decision]...'*.

More concretely, J2 reported looking for *'whether they got their claims right'*, noting that *'sometimes there was one that says this shows the interval was countable, and another one that had some detail of the proof. [The latter] is obviously better, even if the detail isn't right'*.

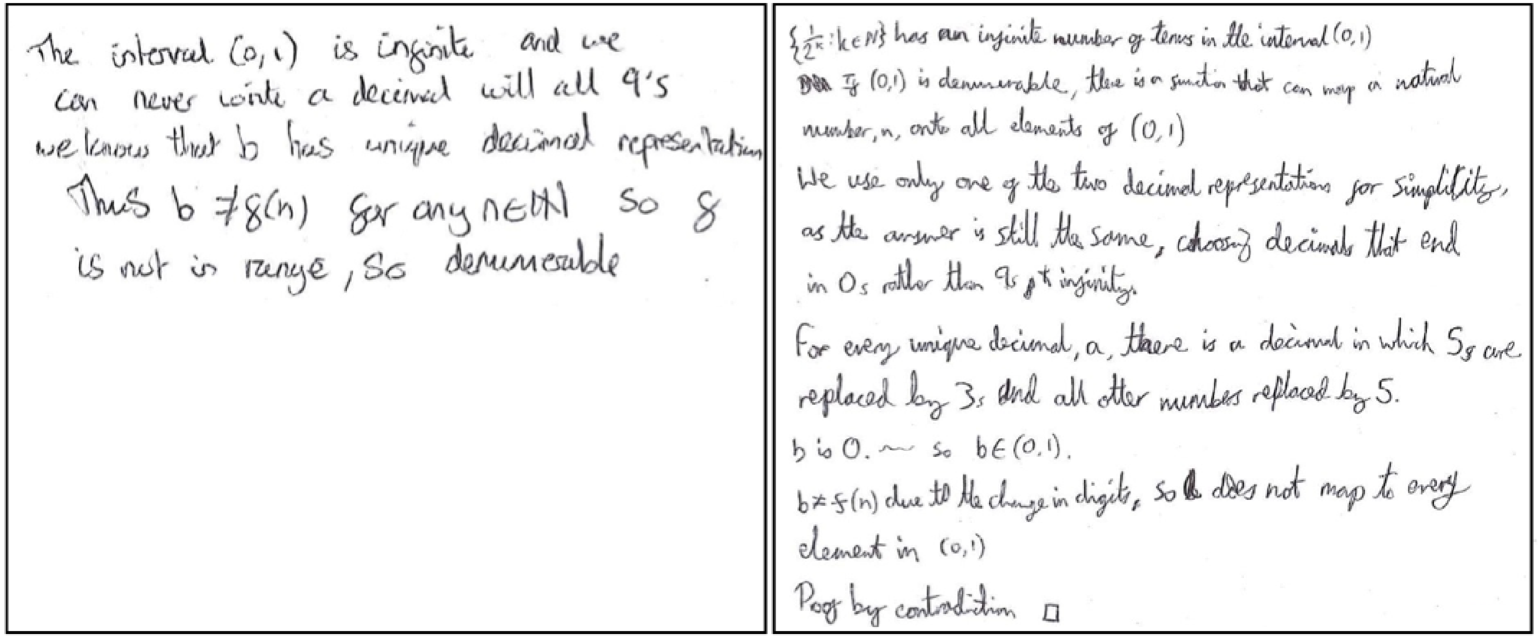
Judges using an error-seeking strategy had, in general, less reading to do. There were several decisions for which this led to a decision faster than those who either performed a detailed reading of both summaries, or read the left summary first, before skimming the second for comparison. This is illustrated by pairing 17, see Figure 9.1, on which all judges agreed that the left response is a better summary of the proof.

The error-seeker will probably find at least one objectionable element in the right response, in that the proof appears to conclude that the interval is denumerable. This is precisely the negation of the conclusion from the original proof, shown in Figure 6.1. The right-hand summary also makes no reference to proof by contradiction, which may have salvaged what is otherwise an error. This error alone was sufficient for J7 to make their decision.

At the other end of the spectrum, judges using a full reading strategy took longer to reach the error, but usually reached the same conclusion. In this particular case, those using a benchmarking strategy took as long as those using a full reading strategy. This is probably the result of the English-speaking tendency to read left-to-right. Hence, the benchmark response read in more detail is almost always the left response. If the order of these two summaries were exchanged, it seems likely that the benchmarking judge would reach a faster decision.

### Overview of reading strategies

Four judges reported reading each summary in full before making decisions. Three reported the use of a benchmarking strategy and two reported seeking errors to inform their judgments. Most judges demonstrated the use of more than one reading strategy. Based on time taken, all judges appeared to perform



The interval (0,1) is infinite and we can never write a decimal will all 9's we know that b has unique decimal representation Thus $b \neq f(n)$ for any $n \in \mathbb{N}$ so $f$ is not in range, so denumerable

$\{\frac{1}{2^k} : k \in \mathbb{N}\}$ has an infinite number of terms in the interval (0,1)

If (0,1) is denumerable, there is a function that can map a natural number, n, onto all elements of (0,1)

We use only one of the two decimal representation for simplicity, as the answer is still the same, choosing decimals that end in 0s rather than 9s at infinity.

For every unique decimal, a, there is a decimal in which 5s are replaced by 3s and all other numbers replaced by 5.

b is 0.~ so $b \in (0,1)$.

$b \neq f(n)$ due to the change in digits, so does not map to every element in (0,1)

Proof by contradiction □

*Figure 9.1. Example pairing from the think-aloud interview task.*

a full reading of at least one pair of summaries and in the think-aloud portion, all judges explicitly highlighted deficiencies to motivate their decisions in context.

### 9.2.2 Approaches to assessment

This theme refers to a spectrum of approaches used by judges to identify the most important features upon which to base their decisions. Here, I refer to a higher-level heuristic motivating the nature of the features judges chose to focus on, rather than the specific mathematical content discussed later.

#### Positive marking

At one end of this spectrum is positive marking whereby judges actively sought elements or phrases to reward. This is typified by J1 here:

> *'I was looking for good things... I read to make sense. And then, I think in the next step, I try to ... I ask myself if that's enough for me... I would say I read and try to collect the valid arguments. And then, see if they add up to what I would assume is an appropriate summary'.*

This self-reported approach to comparative judgment appears most akin to some traditional modes of assessment in which assessors anecdotally seek to give credit to the response at every opportunity. It should be noted that the above excerpt from J1 does not explicitly address their process for comparing two scripts. However, any comparison they make is at least reportedly based on the positive attributes of the two summaries. For example, in the think-aloud task, J1 decided between two summaries by explaining *'the one on the right is better because it identifies that it's a proof by contradiction'*.

The archetypal explanation of a positive marker's decision-making was characterised by the following construction: 'I choose response A because of (possibly implicitly) positive feature X'.

#### Negative marking

At the other end of the spectrum was negative marking, actively seeking errors in each summary. By definition, judges invoking an error-seeking reading strategy were negative markers.

J7 identified themselves as a negative marker:

> *'I usually go for the negative rather than the positive, it's a judgment, it's faster. When I'm comparing, it's faster to see the negative one*

> *because, this sounds really cynical, students are more likely to make mistakes. They usually make a mistake somewhere, not every student is perfect so if they're likely to make a mistake you can usually pick up the mistakes faster which means that you can see if one of them has made quite a few logical errors you know that one's not as rigorous as the other... '.*

In a similar vein, J8 noted that seeing *'incorrect statements influences the decision more than showing correct things'.* In the think-aloud task, negative marking appeared in the form 'I choose $B$ because $A$ had an undesirable attribute'. For example, J3 chose between two summaries stating: *'the left one mis-states the claim, so that already pretty much means it's weaker than the other'.*

While negative and positive marking are in some sense diametrically opposed, I describe their relationship as a spectrum for two reasons. On many occasions, judges identified both positive and negative elements of one or both response. In such cases, it became a question of weighing the influence of the various features. As a researcher, the inference of these unstated weights could only be based on the judges' final decision, and was necessarily an imprecise process. Second, judges rarely stuck to one approach exclusively throughout their judgments. Some decisions lend themselves better to different approaches and many judges demonstrated flexibility in the think-aloud task, even if having communicated a clear preference/bias when asked during the semi-structured interview.

#### Overview of approaches to assessment

Three judges self-identified primarily as positive markers and three self-identified as primarily negative markers. Two judges explicitly noted the benefits of both approaches and said their approach was dependent on context. The ninth judge made no comments clarifying their position. In the think-aloud task, all nine judges used a negative-marking construction to motivate their decision at least once. However, only two of the nine were consistent in their highlighting of exclusively negative attributes.

### 9.2.3 Influencing features

This theme refers to the non-specific mathematical features that judges attended to in making their decisions. The sub-themes here refer directly to mathematical content without referring to elements of the original proof.

I identified four features of the proof summaries that judges attended to: *technical detail, brevity, mathematical precision* and *mathematical fluency.* The first two, in particular, were divisive for the judging cohort, who demonstrated views both for and against the value of each.

### Technical detail

By technical detail, I refer to what J4 called *'the book-keeping of the proof'*. More formally, this sub-theme highlights judges' attention to the notation-heavy elements of the proof, and in particular, their desire to see that notation, or technical detail, spelt out in full rather than described in natural language. For example, J6 reported that

> *'If one of them used words and one of them used the actual notation for it, I would go for notation on the reason that if the logic is correct on the right, that's fine but, they've not actually shown that you can write something in such a way... So in this they have to introduce that function b, which is different to the standard form...'.*

Similarly, J9 explicitly reported using technical detail as a tie-breaker for pairing with little else to split them: *'If I had two [similar] proofs, one was giving a little more details, I would choose that one'.* On the other hand, some judges choose to prioritise structural elements, like J8 who was explicitly *'not looking for details of the proof, but their structure of the proof. So, the key parts of the proof [were most important]'.* Similarly, J1 said *'I had the impression that if someone started doing the summary by going too much into the details, it gave the impression that they had focused on the details, but they'd lost the context of the general picture'.*

### Brevity

Brevity refers to the conciseness of the summary. J1 acknowledged brevity directly, stating on one occasion 'this one is better than this one because at the least, it's shorter'. Similarly, J8 valued the brevity of a particular response based on its simplicity: *'It was very beautiful, it did not do any math it just said this is how they prove it. It is proved by contradiction; "assume this, assume that. The contradiction..." I like it'.*

On the other hand, others valued completeness, noting that longer summaries were likely to contain more important information regarding the proof. For example, J7 said they *'would punish for not covering all the points. So in my head, if you're summarizing something, the idea of summarising usually* *means it's shorter, but, if every single part of what is written, has to be written, then you can't really shorten it'*.

### Mathematical precision

This sub-theme refers to the accuracy of isolated statements present in a given summary. For some judges, precision was an important feature of their decision-making process, heavily punishing abuses of notation or mathematical nonsenses. J7 focused on the precision of the mathematical notation, noting that *'it really depends on [incorrect] notation, that's one thing that's a bugbear for me'*. On the other hand, others were more willing to *'cut some slack, particular with students of this level'* (J4). To this end, J2 focused on readability, saying *'it's clear what it understands... It's not correct mathematically but with undergraduate students especially I don't know... They confuse a lot of the elements with functions and so on. So yea, let's be a little bit generous'*.

### Mathematical fluency

Considerations of mathematical fluency were generally meta-level comments about the overall impression of the given summary. J5 noted that *'The one on the left is, in hindsight actually... it's probably more of a feeling of the overall thing, but it feels like it's... I don't know. There's something a bit strange about the way they've written... overused the infinity symbol'*. Similarly, they said of another summary that *'they tried to sort of give an example and I didn't really feel like the example was... I don't know. I don't think I liked it very much because it felt like they really went off-topic and they didn't seem to understand what was happening'*. J4 justified one of their decisions by reporting *'a gut feeling that they haven't understood it'*, concluding that they are *'less likely to forgive them for [other omissions]'*.

### Overview of influencing features

Regarding technical detail, four judges explicitly rewarded the inclusion of technical elements of the proof, while two judges asserted it was distracting or demonstrated that the bigger-picture elements of the proof had been missed. Similarly, four judges rewarded brevity while two others rewarded its opposite, phrased here as completeness. The dichotomy regarding mathematical precisions functioned on a different axis, with no judges explicitly rewarding mathematical imprecision. However, four judges reported punishing an absence of precision (abuses of notation) while three explicitly stated that such transgressions were overlooked or ignored. Finally, three judges reported that they punished an absence of mathematical fluency.

### 9.2.4 Necessary content

This theme captures the mathematical content judges explicitly noted as being necessary or characteristic of good summaries. Many of the excerpts reported here could have been interpreted as relating to the influencing features discussed above. They have been given their own theme to acknowledge a difference in mathematical specificity.

**Construction of b**

J5 noted that the main point *'would be the construction of a number b that is different from each one of the countably many $A_i$'*. Hence, this judge rewarded summaries including this construction. J6 and J8 made similar comments.

**Construction of f**

Four judges noted that many summaries did not introduce the function $f$, commenting that it needs to be defined for the text to have meaning. For example, J1 noted that one particular summary *'... misses that here we cannot find the set to map all of the interval one-to-one. So what kind of set are they trying to map? If you don't provide any kind of information on that, then how do you conclude that the interval is uncountable'*?

**Proof by contradiction**

Six judges commented on the importance of making explicit the notion of proof by contradiction. For example, in the think-aloud task, J9 observed that

> *'somehow the left feels nicer. A nicer summary. Because perhaps the main reason is because the person on the left told me I'm gonna proceed with proof by contradiction'*.

**Overview of necessary content**

All nine judges indicated that at least one of three sub-themes on necessary content should have been included in students' summaries. Four referred explicitly to the construction of key mathematical objects, while six noted the importance of highlighting the logical structure of the proof, or explicitly noting the contradiction itself.

### 9.2.5 Unnecessary content

At the other end of the spectrum, some judges identified elements of the proof that they deemed actively undesirable.

#### $(\mathbf{0}, \mathbf{1})$ is infinite

The uncountability proof in Figure 6.1 begins with a brief three-line argument establishing that the open unit interval is infinite, before proceeding to establish that it is uncountably infinite. J2 said of this sub-argument: *'I think I probably wouldn't have included that from the outset because, in a way, a finite set is obviously denumerable'*. To the contrary, J3 noted the potential *'pedagogical'* value of including these arguably superfluous sub-arguments. J3 went on to comment that it may well be worth including them, arguing that the *'extent students understand why it's there is a different matter. I think there were one or two proofs of summaries where that observation was pretty much the only relevant content there was. So in that case, it makes a difference'*.

#### 0s and 9s

The proof also includes a sub-argument addressing the conflict between different decimal representations resulting from infinite strings of 0s and 9s. This sub-argument was deemed unnecessary by several judges. For example, J9 said *'Obviously, many of them had this argument with the infinite string of nines, which I think is just not necessary for the summary'*.

#### Overview of unnecessary content

Four of the nine judges indicated that at least one element of the proof was deemed superfluous in at least one case. A fifth judge noted that while some content was mathematically unnecessary, they saw a benefit in its inclusion in some contexts.

### 9.2.6 Non-content-related features

This final theme captures a series of non-content-related sub-themes influencing judges' decision-making. The sub-themes document features that served to distract judges from the primary task of choosing the better summary and were noted as potential threats to the validity of the comparative judgment process.

#### Handwriting/readability

Several judges commented on the poor handwriting of several summaries, making it impossible to parse some summaries in full. Poor handwriting also served to distract some judges who noted that *'at a certain point if the handwriting is too bad, you just have to assume it's bad'* (J7).

#### Arbitrary decision-making

Other judges commented that some decisions were necessarily arbitrary for a variety of reasons. J5 reported not being able to *'choose between two equally terrible summaries'*, while J3 observed that sometimes the *'crimes committed were different but equally problematic'*, concluding that their decision was therefore meaningless.

#### Context-dependence of the task

J4 noted that their approach to judging in a real-world setting would be context-dependent.

> *'[Abuse of notation] doesn't bother me... It would in some more formal context. If they gave it to me as a piece of course work, I would be less happy than if it was in a class test or exam. I think you have the time to be a bit more careful'.*

J3 also noted the context-dependence of the task through the experience of the students:

> *'The proof hinges on a proof by contradiction on the fact they can't... From the students I would want to know whether, for example, if they remember how to construct this set, because this may be a standard procedure that they need to implement and this is why in the beginning I asked you, who are these students?'*

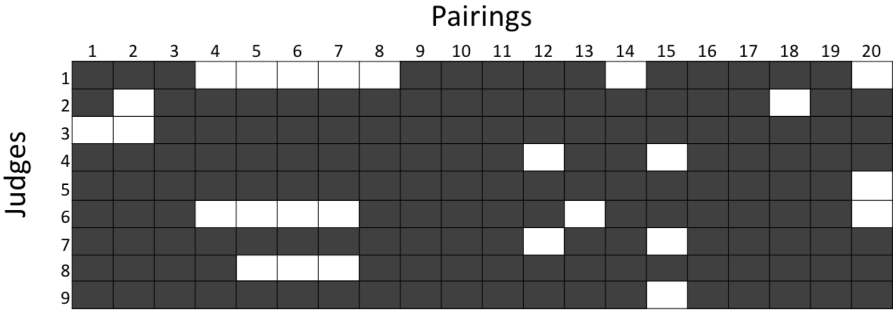
#### Overview of non-content-related features

In total, six of the nine judges raised some query regarding the validity of the task. The vast majority of these are interpreted as passing comments relevant to a small subset of the decisions, or features that caused judges to pause on occasion. Only J3 repeatedly raised objections to the validity of the task. Nevertheless, all judges, including J3, performed all judgments and appeared to engage in a meaningful way.

The following section presents two quantitative measures capturing the reliability of the judgments for each judge.

### 9.2.7 Agreement

Despite the clear variations in judges' justifications for their decisions, there was notable agreement between judges. I demonstrate this in two ways. First, percentage agreement was 85%, suggesting that in the majority of cases there was near-complete consensus among judges. See Figure 9.2, where a dark cell indicates that judge $i$ held the majority view on pairing $j$ for the nine interviewed judges.



Pairings

| Judges | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ■ | ■ | ■ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ | □ |
| 2 | ■ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ |
| 3 | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 4 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ |
| 5 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | □ |
| 6 | ■ | ■ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ | ■ | □ |
| 7 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ |
| 8 | ■ | ■ | ■ | ■ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 9 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ |

*Figure 9.2. Visualisation of agreement across pairings in the interview study. A dark cell indicates that judge i held the majority view on pairing j.*

In traditional comparative judgment studies, the reliability of the resulting scores would be examined using SSR (Pollitt, 2012a) and/or inter-rater reliability (Jones and Alcock, 2014). Both require more decisions than were collected in this study and are hence unsuitable. However, the study from Chapter 6 offers a suitable replacement, given the reliable and valid scores produced for each of the summaries in the present study. In particular, an expected outcome for each of the 20 pairings in this study can be produced by comparing the comparative judgment-based scores generated in the earlier study.

The majority decision of the interviewed judges (illustrated by dark cells in Figure 9.2) was consistent with the higher score in 19 of the 20 pairings. For the one pairing where the majority of the interviewed judges did not match the scores produced in Chapter 6, the two scores were separated by less than .02, corresponding to less than 1% of the standard deviation.

## 9.3 Discussion

This chapter featured a primarily qualitative investigation of judges' decision-making when evaluating students' proof summaries. The thematic analyses generated a fine-grained understanding of the features underpinning judges' decisions. I found wide diversity in these features, documented under six themes: *reading strategies, approaches to assessment, influencing features, necessary content, unnecessary content* and *non-content-related features.* These themes can be viewed as a taxonomy summarising the manner in which judges approach their evaluation of proof summaries in a comparative judgment context.

More surprising than the range of features motivating judges' decisions was the presence of such directly contradictory statements individual judges would make while still reaching the same conclusion about which proof summary was better. In particular, three of the influencing features identified in my analysis were associated with multiple affirmations and negations. In considering *technical detail, brevity*, and *mathematical precision*, multiple judges noted their importance or lack thereof. I found similar results when considering judges' approaches to the assessment task as a whole. In this light, one would expect those asserting technical details to be of utmost importance to disagree with those more interested in the large-scale key ideas of the proof. Similarly, for those prioritising mathematical precision, one would expect divergence from judges explicitly willing to tolerate some notational inaccuracies. By and large, this was not the case, as evidenced by the high agreement between judges.

There is a diverse range of epistemologies underlying mathematicians' (and mathematics educators') conceptions of proof in the literature (Balacheff, 2008). However, this literature provides limited insight into how this diversity manifests itself in the way mathematicians evaluate students' written work. Given the high reliability of judgments found in earlier work, it seemed plausible that this diversity would dissipate in evaluating students' written work, particularly when the assessment pertains to relatively trivial mathematics. At least on the surface, the thematic analysis presented here contradicts this possibility.

The source of agreement in judges' decisions can be explained in two ways. The thematic analysis may have captured outliers in the interview transcripts. With more rigorous attention to frequency and prevalence of particular justification, perhaps using more quantitative techniques, one may find greater consistency in the data. Another possible source of agreement may be inaccessible to the data captured by the interviews used in this study. It is plausible that the judges' explanations were the post hoc justifications most readily available in the moment. Further, a nebulous sense of 'mathematical quality' may be shared by the mathematicians but may be too difficult to articulate. Gaining access to this sense of mathematical quality is a non-trivial task and is discussed further in the final chapter.

### 9.3.1 Research question 2b: What do mathematicians most value when evaluating students' proof summaries?

Earlier answers to this question have focused on statistical modelling of content-specific features of students' proof summaries. The contribution of this chapter is the six-feature taxonomy of judges' decision-making justifications, including but not limited to mathematical content.

Chapter 7, on the uncountability proof, attempted to identify judges' content-specific priorities via regression modelling of a content-based coding scheme. This analysis identified that judges rewarded summaries referring to the two major objects of the proof, $b \in (0, 1)$ and $f : \mathbb{N} \to (0, 1)$, and those identifying the primary proof method (proof by contradiction). The thematic analysis of judges' interviews confirms this quantitative analysis in that the same content-specific themes were raised by the interviewed judges.

In the earlier study, I had also conjectured that judges would punish summaries that include the arguably unnecessary sub-arguments demonstrating the infinitude of the open interval, and clarifying the difficulty of infinite strings of 0s and 9s. In the absence of a negative correlation between Summary Task scores and the corresponding codes, this conjecture was not confirmed. The analysis in the current chapter partially explains this absence. While four judges shared my view of the superfluous nature of (at least one of) these sub-arguments, four others did not comment and another provided an argument to the contrary, suggesting that their inclusion was actively desirable. Given this division in judges' perceptions of the necessity of these arguments, the absence of a significant correlation between Summary Task scores and their corresponding codes is unsurprising.

Similarly, I conjectured in Chapter 6 that brevity would be an important feature of judges' decision-making process. However, as was reported in the information density analysis (Section 6.2.7), this was not borne out in the empirical statistics relating word-count, information density and Summary Task scores. Again, these interviews shed some light on the absence of this finding. While some judges explicitly noted the importance of brevity, others punished summaries that did not cover all elements of the proof. It seems that this distinction should generate a dichotomous split in the decision data, yielding substantive disagreement in judges' decisions. The absence of empirical evidence supporting this expectation lends further credence to the notion that judges' are attending to a more nebulous notion of mathematical quality not accessible here. I return to this discussion in the final chapter.

### 9.3.2 Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

The judgment data presented in this chapter are not sufficient to justify comment on reliability. However, the thematic analysis presented here adds further weight to the validity arguments in Chapter 8. In particular, in discussing research question 2b above, I concluded that the content-specific findings from the interview analysis corroborated the statistical modelling of Chapter 6. In isolation, these content-specific findings are not of particular importance. However, this corroboration is indicative of a more substantive conclusion regarding content validity, namely that judges' intentions are reflected in the comparative judgment-based scores. That is, when judges say they rewarded a particular feature, that feature is indeed significantly related to the resulting scores. While this may seem somewhat obvious, it is important confirmation that the statistical modelling inherent in the comparative judgment process does not mask or override the intentions of judges.

#### Next Chapter

The next chapter, which concludes this thesis, provides summative remarks on each of my research questions, before considering the relationship between the Summary and Conceptions Tasks. This final chapter concludes with the theoretical and practical implications of this work, before ending with some directions for future research.

# Chapter 10

# Final discussion and conclusions

This thesis has presented research on two comparative judgment-based tasks, examining various topics related to proof and proof comprehension. In this final chapter, I summarise this work by addressing each of the research questions set out in Chapter 2. I then highlight the practical, theoretical and methodological implications of this work, before ending with two directions for future work.

## 10.1 Research question 1: What do students and mathematicians write when explicitly asked about their conceptions of proof?

To answer research question 1, I drew on two types of evidence, presented in Chapters 4 and 5. The first and simplest form of evidence came from the content analyses of students' and mathematicians' written responses to the Conceptions Task. The second, applicable only to mathematicians, was the identification of judges' priorities through regression modelling involving content-based codes.

In the students' responses, certainty was the most frequently referenced aspect of proof, indicating a philosophical naivety consistent with their apprentice status in the mathematics community. This was consistent across both studies, with at least two-thirds of both student cohorts referring to certainty. Based on its ubiquity across both studies, I conjectured that certainty is likely to be identified as an important aspect of proof by most undergraduate students, and that such a result would be reproducible in any number of similar research settings. Other features prominent in students' responses included falsification, generality and the scope of the theorem at stake. Each of these features was a significant predictor of Conceptions Task scores in only one of the two studies, suggesting that they are, to some extent at least, idiosyncratic functions of the educational environment, rather than generalisable features that one should expect from any undergraduate student.

From the mathematicians' perspective, the most important aspect of proof was argumentation. This was reflected both in the frequency of appeals to argumentation in mathematicians' written conceptions, and in the regression modeling from both studies. Other characteristics of proof valued by mathematicians included certainty, incontrovertibility and appeals to established knowledge. These findings are consistent with the literature on proof, providing empirical backing to the theoretically driven writings of Aberdein (2009) on proof as argumentation, and Czocher and Weber (in press) on features of proof defined as a cluster concept.

These findings have implications both for the validity of the task, and for understanding proof itself. Both are discussed later in this chapter.

## 10.2 Research question 2a: What do mathematicians most value when evaluating the written proof conceptions of others?

This question was partially answered in Section 10.1, above. In evaluating proof conceptions, mathematicians most heavily rewarded responses containing reference to argumentation. This was reflected in the regression modelling in both Chapters 4 and 5. Other important features included references to established knowledge, incontrovertibility, certainty and clarity, each identified as a significant predictor in the regression modelling in one of the two relevant studies.

I view these features as philosophically consistent with the literature on proofs as arguments produced to remove doubt in the truth of the theorem (Czocher and Weber, in press) by showing the theorem to be the 'logical consequence of axioms, assumptions and/or previously established claims' (*ibid*, p. 20). References to clarity also align with Czocher and Weber's definition, given their focus on proofs as transparent justifications, comprehensible by any sufficiently knowledgeable parties. Certainty, however, is more difficult to characterise. Much of the literature argued against the primarily pre-19th century conceptualisation of mathematics as the business of certainty and truth (Marcus and McEvoy, 2016). This prompts an important question about the content validity of these judgments, given that one of the features rewarded by judges is not well aligned with the philosophical literature on the topic. I address this question in Section 10.4.

## 10.3 Research question 2b: What do mathematicians most value when evaluating students' proof summaries?

From the relevant regression modelling of students' proof summaries (see Chapters 6, 7 and 8), I inferred that judges rewarded summaries attending to the method of proof and the key mathematical objects introduced in the proof. However, such inferences were only possible regarding two of the three proofs, with the primes proof providing no supporting evidence.

The interview study (Chapter 9) took a more qualitative view of judges' decision-making. These findings were summarised by six themes: *reading strategies, approaches to assessment, influencing features, necessary content, unnecessary content* and *non-content-related features.* While the details of the themes on necessary and unnecessary content corroborated the findings from earlier regression modelling, this study also raised a series of other features influencing judges' decision-making. Beyond the range of features motivating judges' decisions, I was surprised by the presence of several features that appear to dichotomously divide the judges without creating divergence in their eventual judgments. In particular, judges were found to have treated notions of brevity, mathematical precision and technical detail from entirely differing perspectives, yet their pairwise judgments were aligned in the vast majority of cases.

This leaves a substantive source of agreement unexplained, leading to the conjecture that judges frequently attended to a more nebulous notion of mathematical quality (or fluency) not accessible via the data presented in this thesis. The consequences of this conjecture for the validity of the Summary Task scores are addressed in discussion of research question 3b, later in this chapter.

## 10.4 Research question 3a: Do written proof conceptions, scored using comparative judgment, generate a reliable and valid output?

### Reliability

The Conceptions Task scores showed acceptable reliability in all expected cases. Given the range of judges recruited, I conjecture that any suitably qualified set of judges would produce similar scores and that above a threshold of a relevant tertiary degree, the scores produced are not sensitive to the qualifications of the judges.

These findings provide insight into mathematicians' conceptions of proof. In particular, despite the diversity of conceptions present in the literature, mathematicians appear to at least agree on they want others to say about proof. This supports the conclusions of Weber and Czocher (2019) that there may be more consensus amongst the mathematical community than has previously been suggested.

### Validity

In addressing the validity of the Conceptions Task scores, I considered evidence regarding both criterion (convergent, divergent and predictive) and content validity. The resulting analysis suggested that while proof conceptions were independent of more traditional measures of proof comprehension or general mathematical performance, the Conceptions Task scores appeared to meaningfully reflect individuals' philosophical awareness when it comes to proof.

This conclusion regarding philosophical awareness was corroborated by the content validity analysis, based on regression modelling of content-based codes. In both chapters, this modelling showed argumentation as the most important feature, alongside established knowledge, incontrovertibility, clarity and certainty. I argued that all bar one of these features is consistent with the established literature, providing robust evidence for the content validity of the Conceptions Task scores as a measure of philosophical awareness.

'Certainty' was the one feature identified as significant in the regression modelling and inconsistent with the literature on proof. That said, it is a relatively recent phenomenon that proof is not commonly viewed as providing certainty (Marcus and McEvoy, 2016). Moreover, as is evidenced by mathematicians' own responses to the Conceptions Task, this view still features as an important aspect of proof for many mathematicians.

A further consideration in understanding content validity is the origin of students' conceptions. Students probably adopt their views of mathematical topics from the mathematicians by whom they are taught. Many mathematicians also featured certainty in their responses to the Conceptions Task. Hence, from the perspective that the comparative judgment-based scores are intended to reflect the collective expertise of the judging cohort, it is to be expected that conceptions containing reference to certainty be rewarded, even if this is not aligned with the philosophical literature.

The implications and practical applications of these findings are considered later. For now, I conclude that the evidence presented supports the claim that the Conceptions Task scores were reliable and valid reflections of the philosophical awareness demonstrated in the responses.

## 10.5 Research question 3b: Do proof summaries, scored using comparative judgment, generate a reliable and valid output?

### Reliability

The Summary Task scores demonstrated acceptable reliability in all cases. This suggests that there is substantive consensus amongst judges on what constitutes an appropriate proof summary, and that this consensus is not sensitive to the judging population. Given the consistently high reliability statistics across all three undergraduate-level proofs, I conjecture that such results would be expected for others.

### Validity

I considered the criterion validity of the Summary Task scores from two perspectives: as a measure of localised proof comprehension, and as a measure of a more general mathematical performance.

In addressing localised proof comprehension, I compared Summary Task scores with the Proof Comprehension Test associated with each of three proofs. The results here were mixed. As discussed in Chapter 8, significant correlation coefficients were found in two of the four studies. A third yielded a near-significant correlation and the fourth showed no significant relationship. I conclude that there is substantive evidence supporting the validity in at least some contexts, but that further work is required to understand the scope of this validity and the range of proofs for which one should expect similar results.

In considering the validity of scores as a more general measure of mathematical performance, Summary Task scores were compared to a series of other measures including module scores and SAT scores. Limited evidence was found for any relationship between the Summary Task scores and general mathematical expertise. As such, I conclude that the Summary Task captures only localised understanding of particular proofs, and is independent of students' performance on more traditional modes of assessment.

I also examined the content validity of the scores assigned to students' summaries of each of the three proofs. Similar to investigations of the Conceptions Task, this analysis was based on regression modelling of content-based coding of students' proof summaries. For two of the three proofs, there was robust evidence suggesting that judges rewarded summaries attending to the method of proof, and the key mathematical objects introduced in the proof. Given the alignment between these features and the proof comprehension assessment model of Mejia-Ramos et al. (2012), I interpret this as corroborating evidence supporting the validity of the resulting scores. In particular, Mejia-Ramos et al.'s model (Table 2.2) highlighted three aspects of proof comprehension aligned with my empirical findings: 'Logical status of statements and proof framework', 'Summarising via high-level ideas', and 'Identifying the modular structure'.

The diverse methodological approaches used to investigate the properties of the Summary Task yielded a largely coherent view: The resulting scores are reliable reflections of students' understanding of the particular proof they were asked to summarise, and that judges prioritised features of these summaries consistent with the literature on proof comprehension. However, they also raise a series of unanswered questions that serve to muddy the waters. These unanswered questions are addressed next.

## 10.6 Open questions and future work

While the summary of findings above provides evidence for the reliability and validity of the scores produced by both tasks, several questions remain open. Here, I highlight three important open research questions, identifying avenues for future research in each case.

1) Given the diversity of features influencing judges' decision-making, what forces drive the consensus amongst judges?

2) Given the significant relationship between the Proof Comprehension Test and multiple measures of general mathematical performance, what do we know about the relationships between the Summary Task, the Proof Comprehension Tests, localised proof comprehension and general mathematical performance?

3) For which proofs is the Summary Task likely to generate reliable and valid scores?

### 10.6.1 On judges' decision-making

How is it the case that a judge focused on rewarding the briefest summaries comes to consistently agrees with a judge explicitly disinterested in brevity, and who instead purports to reward completeness? Similarly, how can a judge who claims to have punished summaries including arguably superfluous sub-arguments agree with a judge who interprets the purpose of a proof summary to be providing the reader with a response sufficient to reproduce the original? And, how is it the case that a judge focused on mathematical precision and a lack of notational errors consistently agrees with a judge purporting to be disinterested in precision and detail, as long as the intention of the text is clear enough to be understood?

These questions about judges' decision-making arose from the interview study and suggest that despite the content-specific consensus indicated by the regression modelling in earlier chapters, there are other less obvious aspects of mathematical quality not yet addressed.

It is, at least theoretically, possible that content-specific features, such as references to the method of proof, are sufficient to override other sources of dissent like brevity and precision. This, however, seems unlikely for two reasons. First, such features are not sufficient to distinguish between many, if not most, pairs of summaries. And second, there are myriad other features relevant to the decision-making process that it seems implausible to have such simple factors override all others.

I conclude that there are other factors influencing judges' decision-making than those studied in this thesis. Features such as presentation have been discussed in other comparative judgment research (e.g. Jones and Alcock, 2014), and are hence left as a likely relevant side note. Instead, I focus here on features specific to the proof-related context. In particular, several judges commented on relying on an overall impression of students' understanding, implying that their decisions were based not only on mathematical content, but on their intuitive perception of how well the student understood the proof. Of course, this perception is likely partially dependent on the mathematical content, but it seems that there is an important distinction to be made here between concrete mathematical features and intuitive, holistic perceptions of the understanding conveyed by a given summary.

Given the subjective nature of comparative judgment, and the arguably unknowable complexity of human decision-making, it is not surprising that judges' motivations cannot be discretely and cleanly documented. However, further systematic investigations can take steps toward understanding the role of judges' 'gut-feelings' and further clarify the relationship between intuition and the more concrete influencing features documented in this thesis.

### Future work on judges' decision-making

I identify two directions for future work on judges' decision-making. The first direction involves judges evaluating artificially selected responses with predetermined features of interest. The second is based on tracking the eye-movements of judges' in evaluating students' responses. Both directions would provide further perspectives for triangulation with the data presented here. Although my focus is on proof conceptions and summaries, both approaches could prove productive in various types of judgment-based research.

Regarding the judging of artificially selected responses, it is possible to understand the likelihood of the earlier conjectures about judges' decision-making. For example, there is an unresolved tension between judges' focus on content-specific features (i.e. references to the method of proof and key mathematical objects) and more nebulous properties such as mathematical fluency based on judges' intuitive response to the summaries (or conceptions). To investigate this tension, several artificially generated summaries could be produced with a strong emphasis on one but not the other. For example, researchers could generate a series of summaries that address the method of proof and introduce the key mathematical objects, but also include any number of mathematical nonsenses, abuses of notation, and ambiguous language. One would expect that such summaries would be divisive for the judges, causing the reliability to fall. Further, if all such artificially generated summaries are awarded similar scores, conclusions about judges' relative priorities also become apparent. If, for example, the series of summaries described here are all scored high, one should conclude that mathematical content was more important to judges than the overall impression generated by the text. On the other hand, if (as I would expect) these summaries all receive low scores, one can conclude that judges attend more to the intuitive impression given by each summary.

An interesting extension of this further work would be in attempting to artificially generate proof summaries with the opposite characteristics. That is, is it possible to produce a proof summary that attends neither to the method of proof or the key mathematical objects, but is still consistently rewarded by the judging cohort? I suspect such a summary may not be possible to produce. However, its existence would provide compelling evidence that the content-specific features of students' summaries are only a minor factor in judges' decisions.

I also identify a second line of future research, based on tracking judges' eye-movement. From such data, it is possible to address the following three topics:

- What do judges attend to last, before making their decisions? And are these final fixation points systematically related to the nature of their decisions?
- What is the role of problematic features (i.e. abuses of mathematical notation) in judges' decision-making? Do judges fixate on such problematic features? If so, are these fixations temporally related to their decisions?
- Are more difficult judgments (defined by the difference in the resulting scores) more time-consuming? And vice versa, are more time-consuming judgments necessarily those that are more difficult?

This is analysis I intend to conduct as a Postdoctoral Research Fellow at a US university in 2020. The necessary eye-tracking data were collected alongside the interview data presented in Chapter 9.

### 10.6.2 The relationship between the Summary Task and general mathematical performance

The second open question emanates from the following three observations: 1) the Summary Task and Proof Comprehension Tests produced significantly correlated scores for the uncountability and Fibonacci Proofs (Figures 6.5 and 8.5); 2) the Proof Comprehension Tests were related to some general measures of mathematical performance (Tables 7.3 and 8.2); but 3) the Summary Task was not related to any of these more general measures (discussed in Section 8.5.6).

As was discussed in Chapter 2, traditional assessments in tertiary mathematics modules are often considered insufficient measures of proof comprehension, even in modules where the primary focus is on students' understanding of proof and mathematical reasoning (Mejia-Ramos et al., 2017; Weber, 2012). For this reason, the absence of a relationship between the Summary Task and traditional module assessments could be explained by the inadequacy of these traditional assessments.

However, the Proof Comprehension Tests, also ostensibly measures of localised proof comprehension, yielded significant relationships with module scores in several cases. This indicates that something meaningfully proof-related is measured by these traditional assessments. On the other hand, the Proof Comprehension Tests were significantly correlated with the Summary Task associated with two of the proofs investigated. While the coefficient was never greater than .35, this is evidence that they measure overlapping constructs (related to localised proof comprehension), even if there are differences yet to be accounted for.

In an attempt to understand the relationships between the Summary Task, Proof Comprehension Tests and traditional assessments, I appeal to the familiarity of the assessments to both students and assessors. It is safe to assume that all mathematics students have experience with multiple-choice tests, and that many will have established strategies for addressing difficult questions (e.g. guessing via a process by elimination). Similarly, all students have experiences with most types of assessments contributing to module scores. Again, they will likely have strategies for approaching difficult questions and garnering partial credit. This leads to the conclusion that both multiple-choice tests and traditional assessments reward students' ability to maximise their expected outcome (as well as more desirable aspects of performance including content knowledge).

On the other hand, students have likely never been asked to produce proof summaries, and are probably far less familiar with the Summary Task. This has the dual effect of minimising the effect of assessment strategy on the outcome of the assessment, while also introducing an element of flexibility into the assessment.

This conjecture accounts for the three relationships in the triad of Summary Task, comprehension test and module/SAT scores. The relationship between the Summary Task and comprehension test reflects the overlapping content domains. The relationships between the Proof Comprehension Tests and module scores are reflective of students' familiarity with the 'rules of the games' and their ability to maximise their expected score from the content knowledge they possess. Finally, the absence of a relationship between the Summary Task and the module scores is reflective of a (partial) failure of standard module assessments to adequately assess proof-specific comprehension.

**Future work on the nature of the Summary Task scores**

Without intending to be uncharitable to the traditional assessment structure of the mathematics modules involved in this study, my account of the empirical relationships is consistent with the literature noting the discrepancy between standard practice and authentic measures of students' proof comprehension (Mejia-Ramos et al., 2017; Weber, 2012). Further research is necessary to substantiate this conjecture and could follow a design briefly outlined here.

The claims above rely on students' unfamiliarity with the Summary Task. If a group of participants were either explicitly trained in what mathematicians expect from students' summaries (perhaps based on the content validity analysis in this research), or simply given sufficient exposure to the task to generate familiarity with appropriate strategies and techniques, I hypothesise that the effects discussed above would diminish. In particular, I would expect that those trained in generating proof summaries would achieve similarly on the Summary Task and standard measures of mathematical performance.

### 10.6.3 The scope of the Summary Task

I have demonstrated that the Summary Task has the potential to generate a reliable and valid measure of students' local comprehension for any number of proofs. However, as has been noted, the scope of applicable content domains remains unclear and requires further research.

To this end, I suggest that a productive data set comprises at least 10 proofs, each with a minimum of 30 associated summaries. At least one of the Uncountables, primes and Fibonacci proofs should feature in this list of 10+ proofs. Summaries can then be evaluated in one all-encompassing comparative judgment assessment, allowing summaries of different proofs to be compared side-by-side. The simultaneous judging of responses to different tasks is an established method in the comparative judgment literature (Jones and Karadeniz, 2016; Hunter and Jones, 2018). Moreover, this is a particular strength of the method, promoting comparative judgment as an ideal tool with which to conduct such research.

The 300+ summaries would require a minimum of 3000 judgments to reach an appropriate threshold for meaningful analysis of reliability and validity. Students' achievement data should also be recorded, including traditional assessments from tertiary mathematics modules and any available general measures of mathematical performance.

I intend to collect such a dataset in the fall semester of 2020. The analysis could be approached from several perspectives. Here, I identify two, both targeting questions left open by the research in this thesis.

First, this hypothetical 10-proof dataset addresses the need to better understand the scope of applicability for the Summary Task. In the absence of Proof Comprehension Tests as benchmarks for most proofs, validity analysis for the resulting scores can be based on cross-contextual analyses and content validity considerations similar to those presented here. By gathering data on a large number of proofs, it becomes easier to identify patterns, and verify existing conjectures regarding the features of proof summaries deemed most important by judges, and by extension, the validity of the resulting scores. For example, in Section 8.6.1 (page 154), I conjectured that the complexity of a given proof plays an important role in determining whether judges deem it necessary for summaries to include a reference to the method of proof at hand. This conjecture was based on observed variation across two particular proofs. With a set of proofs to examine, the veracity of this conjecture can be investigated.

Another important outcome of this analysis would be in identifying large classes/categories/types of proofs for which the Summary Task can be confidently used by practitioners and researchers. This will be accomplished both by identifying patterns in proofs for which the Summary Task is successful, and in seeking counter-examples or boundary cases wherein the Summary Task (or other comparative judgment-based approaches) are likely not to be appropriate.

Second, this hypothetical data will also provide insights into the dimensionality of proof comprehension. By comparing valid Summary Task scores from many proofs, further cross-contextual analysis can be used to understand the conjecture that students' understanding of one proof should predict their understanding of many others. Such analysis would follow a similar structure to that presented in my discussion of the primes and Fibonacci proofs (see Section 8.5.8).

## 10.7 Implications and applications

First, I consider the practical implications of this work via various applications to classroom settings. I then consider the development of new assessments based on the research presented in this thesis, before finally addressing the theoretical implications of this work.

### 10.7.1 Classroom applications of the Conceptions and Summary Task

#### The proof Conceptions Task as a tool to create awareness

I offer two suggestions for the use of the Conceptions Task in promoting productive engagement in (the philosophy of) mathematics.

The first is a formative tool, wherein students complete the task without completing the comparative judgment aspect. Consistent with the literature on beliefs and engagement (Muis, 2004), participating in an activity like the Conceptions Task can promote productive engagement with mathematics more generally, even if responses are not analysed in a structured manner.

Building on the first application, I believe the Conceptions Task could also be used as a peer assessment task wherein students are asked to evaluate the responses of others. In this way, students are asked to critically engage with a variety of responses, further promoting productive engagement with the topic. Jones and Alcock (2014) adopted a similar approach, having students judge peers' responses to a task in Introductory Real Analysis. Alongside successful findings regarding the reliability and validity of students' judgments, the authors noted the self-reported learning benefits of being asked to consider the merit of other responses.

Based on the evidence presented in this thesis regarding the relationship between conceptions and mathematical performance, I think it would be misguided to assign assessment credit based on the perceived merit of students' judgments or conceptions in most educational settings. However, this does not preclude the task from having other classroom-based merits.

#### The Summary Task as a measure of local proof comprehension

The Summary Task yielded seemingly reliable and valid scores as a measure of localised proof comprehension for multiple proofs. This suggests that the Summary Task can eventually be used as part of a varied 'assessment diet' (Iannone and Simpson, 2011) attempting to generate a holistic picture of students' understanding of proof. This will be particularly valuable given the ease with which the Summary Task can be transferred across content domains without the design burden inherent in other approaches like the resource-intensive development of the Proof Comprehension Tests of Mejia-Ramos et al. (2017).

That said, the evidence in this thesis is not sufficient to make general validity claims for an arbitrarily chosen proof from any undergraduate module. While I expect that further research will demonstrate wide-reaching applicability of the

Summary Task, practitioners should exercise caution in the absence of further research.

### 10.7.2 Designing new assessments

The above discussion focused on applications of the comparative judgment-based tasks implemented in this thesis. However, there are also numerous opportunities using this research as a base-point from which to develop new assessments. Here, I highlight two. The first focuses on an assessment of the Summary Task not requiring any further comparative judgment-based data. The second considers other proof-related applications of comparative judgment with other foci.

#### The Summary Task without comparative judgment

The comparative judgment-based Summary Task could also be used to generate new assessments that are less resource-intensive to evaluate. In particular, it is possible to use the Summary Task and associated content analysis to generate a reliable and valid assessment rubric for proof comprehension assessment. Rubric-based assessments have appeared alongside comparative judgment before (Heldsinger and Humphry, 2010). However, in this case, a rubric was generated independently of the comparative judgment process and used as a benchmark measure against which to evaluate the validity of the scores resulting from a comparative judgment-based assessment. This process is often noted as laborious to both generate and implement, and hence I propose a different role for the use of rubrics alongside comparative judgment.

Following a different strand of Heldsinger and Humphry's work, I considered the use of comparative judgment-based scores as benchmarks in their own right, against which other non-comparative judgment assessments can be generated. In their work, comparative judgment-based scores were used to exemplify responses expected at varying stages of development on a narrative writing task with primary school students. Exemplars were used as benchmarks for assessing future scripts, bypassing the comparative judgment stage. However, the authors discuss students' work in terms of Piagetian-like development stages where the writing samples are assumed to be indicative of 'writing development'. It is unclear that this assumption applies to proof summaries given the variety of responses one expects, particularly given the nature of errors and misconceptions present in lower-scoring responses. With this in mind, I propose a similar but distinct application of comparative judgment, generating a potential measure from analysis of the statements judges valued most.

One could evaluate proof summaries by awarding a score between 0 and $N$, based on reference to any $N$ aspects of the proof found to be a significant predictor of parameter estimates. In doing so, one has a presumably highly reliable rubric (recall the high pooled Cohen's $\kappa$s in each study) for scoring responses with validity based on the collective expertise of the judges in the original cohort. There are many possible iterations and variations of such an approach, either through weighting particularly important elements and/or through subtracting credit for the inclusion of unnecessary content.

### Comparative judgment as a basis for other proof comprehension-related tasks

In this thesis, the Summary Task was used to investigate the potential of comparative judgment in proof comprehension. This task was chosen for its alignment with the literature on proof comprehension assessment and the variation one would expect from such a task, which is valuable for comparative judgment-based analyses. While this task successfully demonstrated several desirable properties in my research, it was not the only available choice, and future researchers may wish to consider others.

In particular, simple proving tasks, as well as concept explanation and peer-assessment may be fruitful areas for consideration. By simple proving tasks, I refer to the familiar protocol of providing students with a theorem and requesting a valid proof. While the drawbacks discussed of accessibility and variation in responses remain (see Section 10.7.1), a straightforward proof construction task is likely familiar to students *and* judges, and may function as a reliable and valid measure of proof comprehension. The benefits of a comparative judgment-based approach to a simple proving task, over a traditional assessment approach, lie in the efficiency of marking, and the availability of reliability statistics providing feedback to the assessor regarding the merit of a particular implementation of the assessment.

Beyond new assessment approaches to familiar tasks, I also highlight concept explanation tasks and peer assessment as possible avenues for comparative judgment-based assessment in proof comprehension. By concept explanation, I draw on Jones and Karadeniz (2016) who evaluated secondary-school students' understanding of mathematical concepts/objects by asking for explanations of, for example, equations, ratios and area/volume. In a proof comprehension setting, questions may be focused on a particular proof method (i.e. explain the method of proof by induction, and provide an example to illustrate your explanation), or a pertinent mathematical object necessary for the study of a given discipline (i.e. define a group, provide an example, and prove that your example is, in fact, a group).

Finally, I consider the role of peer assessment in proof comprehension. As discussed earlier in this chapter, peer assessment has been successfully implemented in comparative judgment settings before (Jones and Alcock, 2014; Jones and Sirl, 2017). In the realm of proof comprehension, it seems that such an approach could be particularly profitable, given the reported absence of sensitivity to the judges' educational background. Peer assessment comes with the potential pedagogical benefits of critically analysing peer responses, and the practical benefits regarding efficiency.

### 10.7.3 On proof and proof comprehension

Here, I address the implications of my research for understandings of proof itself. I consider the notion of proof as a cluster concept (discussed in Chapter 2) and the empirical evidence supporting the theoretical assertions of Czocher and Weber (in press), before outlining the empirical evidence supporting a view of proof comprehension as a unidimensional construct.

#### Proof as a cluster concept

I return to the notion of proof as a cluster concept (Czocher and Weber, in press) and the empirically backed assertion of Weber and Czocher (2019) that mathematicians' agreement on proof is located in typical settings. Recall Czocher and Weber's theoretical claim that proof can profitably be seen as a probabilistic entity based on a collection of identifying features. In finding consensus between judges in my comparative judgment-based research, it seems that while it may be the case that each mathematician holds a different subset of these identifying features as most important, there is sufficient overlap or shared understanding to justify viewing proof as a cluster of overlapping but distinct conceptions of proof. Returning to Weber and Czocher's work, I suggest that my comparative judgment-based research has presented mathematicians with a series of typical cases. To make this argument, I extend the authors' original domain from proof-verification to written conceptions of proof itself. The authors found that mathematicians tended to agree on proofs using typical methods, finding substantive disagreement only in unusual settings like visual or computer proofs. In the same manner, the proof conceptions presented to mathematicians in my research were short, simple and likely largely familiar accounts of proof to most mathematicians. I would expect to find greater disagreement (and hence lower

SSR and inter-rater reliability) if the written conceptions were either longer or more nuanced than those present in my research.

**On the dimensionality of proof comprehension**

I presented a cross-contextual analysis of the relationship between scores related to two distinct proofs (Chapter 8). For both proofs, I considered the comparative judgment-based Summary Task scores and the Proof Comprehension Tests. I found high significant correlations when comparing the two comprehension tests, suggesting that success with one proof was indicative of success with the other. Similarly, comparing the Summary Task *scores* for each proof also yielded significant results. Further, the Summary Task scores for the Fibonacci proof were significantly related to the comprehension test scores for the primes proof, although the inverse, comparing Fibonacci comprehension test with primes summary scores, yielded no significant result.

This is evidence that comprehension of the two proofs is related. Further, localised proof comprehension, as captured by both the Summary Task and the multiple-choice Proof Comprehension Tests, is a singular construct. These findings are consistent with Mejia-Ramos and Weber (2016) who reported high significant correlation coefficients between any two of their three Proof Comprehension Tests.

### 10.7.4 On measuring (beliefs and) conceptions

My research offers two contributions to the literature on the measurement of conceptions. The first is specific to the measurement of students' proof conceptions, corroborating the findings of Stylianou et al. (2015) and the absence of a relationship between students' beliefs about proof and their performance on proof-related tasks.

The second contribution is methodological, providing a unique tool for quantifying any number of subjective beliefs or conceptions. In this thesis, the comparative judgment-based Conceptions Task was used to quantify the quality of written conceptions of mathematical proof. From a methodological viewpoint, there is nothing unique about proof as a target domain here. In principle, a version of the Conceptions Task could be used to evaluate (students') beliefs about myriad other topics. One particular domain of personal interest is students' beliefs about the role of empirical evidence in established knowledge in the physical sciences. Others may include the role of formal logic in mathematics, or even self-efficacy beliefs measured using closed-form questionnaires by Stylianou et al. (2015).

### 10.7.5 On comparative judgment

This thesis has provided empirical evidence supporting the use of comparative judgment in two new domains: conceptions of proof and localised proof comprehension. This contributes to the ever-growing list of content domains in which comparative judgment has proven profitable. In the study of the primes proof, I also presented a much-needed boundary case. This represents a rarely reported case in which a comparative judgment-based assessment failed to produce the desired or expected resulted. As was discussed in Chapter 2, this rarity is either a function of the file-draw problem or indicative of an assessment approach with near-unlimited applications. In each case, the study of the primes proof is a unique example to generate an understanding of domains in which comparative judgment-based tasks are not appropriate.

## 10.8 Final remarks

The research presented in this thesis offers theoretical and methodological contributions to the literature on proof, proof comprehension and comparative judgment. To the literature on proof, I offer new insights on the conceptions of proof held by students and mathematicians. To researchers of students' conceptions and beliefs, I offer a new methodological tool for assigning a quantitative value to subjective entities. To researchers of proof comprehension, I offer insights into the nature and dimensionality of proof comprehension, and a methodological tool for future investigations. To undergraduate educators, I offer a new assessment for evaluating students' local understanding of given proofs. And finally, to comparative judgment researchers, I offer a new domain of applicability, adding to the ever-growing list of content domains in which comparative judgment can add value.

My research opens several avenues for future research, and I hope it will serve as a starting point for a programme of research wherein comparative judgment can be used to better understand proof, proof comprehension and the related behaviours of both students and mathematicians.

# Appendix A

# Defining and discussing misfit

In this appendix, I first define judge (and script) misfit in comparative judgment assessment. I then discuss its use in the literature and present an argument against its usage in research on the reliability and validity of new assessments.

### Defining misfit

Formally, we label the *residual*, $\mathrm{Res}_{j,A,B}$ of judge $j$'s pairwise comparison between texts $A$ and $B$, and compute

$$\mathrm{Res}_{j,A,B} = D_{j,A,B} - P(A > B),$$

where similar to earlier, $D_{j,A,B} = 1$ when judge $j$ chooses $A$ over $B$, and 0 otherwise. We then compute the standardised residual of this comparison, dividing by the square root of the information function (Pollitt, 2012a),

$$\mathrm{StdRes} = \frac{\mathrm{Res}_{j,A,B}}{\sqrt{P(A > B)[1 - P(A > B)]}}.$$

The information function, from Pollitt (2012a), $I = P(A > B)[1 - P(A > B)]$ is a measure of the information embedded within a particular judgment. Note that $I$ has a maximum and $P = 1/2$ and minima at $P = 0$ or 1. Hence, judgments identified by the model as least certain are those with the most information. This is of particular importance in Adaptive Comparative Judgment (Pollitt, 2012b) but is not essential for my purposes. Its role in the calculation at hand is to minimise the impact of difficult decisions on the residues, and subsequently, the misfit of a given judge.,

$$\text{StdRes} = \frac{\text{Res}_{j,A,B}}{\sqrt{P(A > B)[1 - P(A > B)]}}.$$

By aggregating across the standardised residues for the set of decisions performed by a given judge, we get a measure of how well that judge's decisions 'fit' the model. Note this is a post-hoc calculation dependent on $P(A > B)$ and hence on stable estimates, $v_i$. Drawing on the Rasch literature, Pollitt (2012a) observed that there are several ways to aggregate the residues into a measure of *misfit*. The one he proposed is known as Infit (or Infit Mean Square) and is calculated by first computing the Weighted Square Residual for each decision,

$$\text{WSR}_{j,A>B} = \text{Res}^2_{j,A,B} \times I_{AB}.$$

We then sum across all judgments made by judge $j$ and divide by the information embedded in each of those judgments,

$$\text{misfit}_j = \frac{\sum_{judgments} \text{WSR}_{j,A,B}}{\sum_{judgments} I_{AB}}.$$

Pollitt also notes *Outfit* as an alternative measure, although does not spell-out its calculation. Belonging to the Rasch literature, Outfit has not been seen in the comparative judgment literature and is hence not explored further.

### Misfit in the literature

As defined by Pollitt (2012a), misfit can be interpreted as a mean Chi-square. In considering the relative quality of judges' decisions, it is standard to use the criterion of two standard deviations from the mean as a cut-off for acceptability. In this way, misfit has been used as a quality control on the judging population, used to consider excluding judges who either appear not to have engaged with the task, or have engaged in a way significantly different from their peers.

Alongside its traditional application to judges, misfit can also be applied to scripts. By analogy to Rasch analysis, we can use the symmetry between judges (or items) and scripts. By reversing the role of scripts and judges in Pollitt's misfit calculations, we can aggregate across the residues of all decisions involving a given script instead of a given judge. In this case, a script beyond two standard deviations from the mean is likely a particularly divisive text that has prompted judges to evaluate it from clearly divergent vantage points. For example, a clear, well-written summary of a given proof may be written in such poor handwriting that several judges cannot extract the exemplary mathematical content. The exploration of script misfit is not explicitly mentioned by Pollitt (2012a) but has been reported alongside judge misfit several times (Jones et al., 2015; Hunter and Jones, 2018; Heldsinger and Humphry, 2010). Only Bisson et al. (2016) reported judge misfit alone. In all cases, every judge and script remained in the final analysis, despite a non-zero number of misfitting items. These cases were justified either as an expected consequence of Pollitt's exclusion criteria (assuming a normal distribution,  2% of all judges and scripts should be beyond acceptable bounds) or in conjunction with other measures (see above) to conclude that the dataset is reliable as a whole.

The role and use of misfit in the literature has been inconsistent, raising questions about its value for education research. I position misfit as related to reliability as it is a measure of the difference between judges and can hence be viewed as a proxy for inter-rater reliability. However, its standard usage, set out by Pollitt (2012a), is one regarding quality control and is hence more closely akin to external validity. While it is reasonable to evaluate the quality of a given dataset by investigating the number of judges (and scripts) behaving unexpectedly, two problems arise when using this measure to consider excluding data. These stem from the tension between misfit as a measure of reliability or validity.

The first is a recursion problem. After excluding the misfit data, one presumably computes a new model and checks for misfit data again. While a normality assumption is now less reasonable (having excluded data from only the top end of the distribution), it is likely that new data will now appear as misfitting. This problem is solved via pre-registration, whereby the researcher makes a prior commitment to exclude misfit data based on one (or more) iterations of Pollitt's exclusion criteria. In this way, the misfit measure has become a tool to improve the quality of the data, but it has now lost its power to evaluate reliability and validity.

Similarly, it is unclear what researchers should do with texts mathematically identified as misfits, but that do not appear qualitatively unusual. In education, comparative judgment is often used on the premise that identifying the quality of scripts is difficult in isolation. A researcher can 'examine' a misfit script and qualitatively consider its place in the dataset, but this subjective approach appears to somewhat undermine the quantitatively driven method.

### Against misfit

Under its current use in the literature, misfit has been a tool to draw attention to data that potentially does not belong in the dataset. However, I claim that in appropriate cases, there are better measures available. The misfit data has two sources: the disengaged and the authentically eccentric. The disengaged judge is, for example, a paid expert with little interest in the task who performs judgments at random in order to receive maximum payment for minimal effort. The disengaged script is likely blank, incomplete, or transparently off-topic. For the uninterested judge, I argue that time data is a better measure than misfit for generating exclusion criteria. Our hypothetical judge has performed 50 judgments per minute and can be excluded on this basis alone before any further analysis is conducted. Similarly, blank, incomplete, or transparently off-topic scripts can all be removed prior to any analysis.

Best practice on dealing with blank scripts is not well-established in the comparative judgment literature. Interestingly, it is not clear that blank scripts will always filter to the bottom of a comparative judgment-based evaluating, raising questions about validity and the possibility of writing something worse than nothing. I return to this topic in later empirical work and again in the discussion of Chapter 6.

Authentically eccentric data, the other source of misfits, is more complicated to address. Here, I am thinking of the hypothetical judge who particularly values a relevant property of scripts that others deem less important, or the hypothetical script with near-illegible handwriting but excellent mathematical content. Both cases are likely to be labelled by Pollitt's criterion as misfits, but I claim that these should not be excluded for research purposes.

Consistent with the subset of the literature discussed here, I informally explored misfits in each of the empirical studies presented in this thesis. In the absence of substantive findings, and in light of the issues discussed here, these informal explorations are not reported. In the absence of Pollitt's exclusion criterion, I consider time-data as a check that each judge has meaningfully engaged with the task. Blank and incomplete scripts are addressed in accordance with the purpose of each study and are addressed separately in each empirical chapter.

# Appendix B

# Interview schedule

1) How did you find the judging process?

2) How did you make your decisions?

3) Did you have a plan or systematic approach? If so, please describe it.

4) What were you looking for, if anything, in making decisions?

5) If there was a pattern, did you usually identify the 'better' summary or the 'worse' summary when deciding which to choose?

6) Are you aware of any changes over time?

7) Did you have any strategies or approaches for identifying good/bad summaries? If so, please tell me about them.

8) Did any summaries get 'labelled' in your head to make future judgments involving that summary more efficient? If so, please identify them and explain the influence of this label on your decision-making.

9) Did the length of any summary consciously influence your decisions? In particular, the task demanded a summary of fewer than 40 words. Did this threshold influence any of your decisions? If so, how?

10) Any other comments before we open the judgments.

# Appendix C

# Initial thematic analysis code list

| | |
|---|---|
| Marking disposition | Evidence of negative marking (theoretical)<br>Evidence of negative marking (concrete)<br>Evidence of positive marking (theoretical)<br>Evidence of positive marking (concrete) |
| Changes over time | No change over time<br>Explicitly awareness change over time |
| Reading strategies | Left then right<br>Stop after error<br>Use LHS as benchmark |
| Task critique | Written text doesn't capture understanding<br>Proof summary is ambiguous<br>Educational vs content distinction unclear |
| Context dependency | Participants' content knowledge<br>Content matters<br>Task matters |
| Definition of summary | Stand-alone document<br>Must NOT reconstruct the original<br>Should provide an overview<br>Should provide the reader the ability to reproduce<br>Judge's ideal summary |
| Desirable content isolated | Logical structure (non-specific)<br>Contradiction<br>Introduce important objects<br>Technical detail<br>Mathematical fluency |

| | |
|---|---|
| Focus | Notation/detail |
| | Brevity |
| | Accuracy |
| | Structure/big-picture |
| | Fluency/overall impression |
| | Completeness |
| Undesirable content | Technical detail |
| | Defintitions |
| | Demonstrating (0,1) to be infinite |
| | Discussing 0s and 9s |
| Notation | Punish poor notation |
| | Accept poor notation |
| Arbitrary decision-making | Scripts too similar |
| | Scripts too bad |
| Observations | Responses were low quality |
| | Responses were hard to read |
| | Did not know students had been asked to summarise |
| Decision-making strategies | Generating a hierachy |
| | Looking for egregious errors |
| | Systematic R>L bias |

# References

Aberdein, A. (2009). Mathematics and argumentation. *Foundations of Science*, 14(1), 1–8.

Alcock, L. and Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *Journal of Mathematical Behavior*, 24(2), 125–134.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3), 451–462.

Atiyah, M. (1994). Responses to 'Theoretical mathematics: Toward a cultural synthesis of mathematics and theoretical physics' by A. Jaffe and F. Quinn. *Bulletin of the American Mathematical Society*, 30(2), 178–207.

Attride-Stirling, J. (2001). Thematic networks: an analytical tool for qualitative research. *Qualitative Research*, 1(3), 385–405.

Balacheff, N. (2008). The role of the researcher's epistemology in mathematics education: An essay on the case of proof. *ZDM - International Journal on Mathematics Education*, 40(3), 501–512.

Bartholomew, S. R., Ruesch, E. Y., Hartell, E., and Strimel, G. J. (2019). Identifying design values across countries through adaptive comparative judgment. *International Journal of Technology and Design Education*, 1–27. Retrieved from https://doi.org/10.1007/s10798-019-09506-8.

Benton, T. and Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters*, 26, 22–28.

Bisson, M. J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, 2(2), 141–164.

Bloxham, S., Den-Outer, B., Hudson, J., and Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment and Evaluation in Higher Education*, 41(3), 466–481.

Bradley, R. and Terry, M. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3), 324–345.

Bramley, T. (2007). Paired Comparison Methods. In Newton, P., Baird, J., Goldstein, H., Patrick, H., and Tymms, P. (Eds.), *Techniques for monitoring the comparability of examination standards*, (pp. 246–295). London, UK: Qualifications and Curriculum Authority.

Bramley, T. and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, 26(1), 43–58.

Bramley, T. (2015). Investigating the reliability of adaptive comparative judgment. *Cambridge Assessment Research Report.* Cambridge, UK: Cambridge Assessment.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.

Bundy, A., Jamnik, M., and Fugard, A. (2005). What is a proof? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363, 2377–2391.

Cohen, L., Manion, L., and Morrison, K. (2000). *Research Methods in Education* (5th ed.). London, UK: Routledge Falmer.

Conradie, J. and Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics*, 42(3), 225–235.

Cowen, C. (1991). Teaching and testing mathematics reading. *The American Mathematical Monthly*, 98(1), 50–53.

Creswell, J. W. and Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research.* (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Czocher, J. A. and Weber, K. (in press). Proof as a cluster category. To appear in *Journal for Research in Mathematics Education.*

Davies, D., Collier, C., and Howe, A. (2012). Assessing scientific and technological enquiry skills at age 11 using the e-scape system. *International Journal of Technology and Design Education*, 22(2), 247–263.

Davis, P. and Hersh, R. (1981). *The Mathematical Experience.* New York, NY: Viking Penguin Inc.

Dawkins, P. and Weber, K. (2017). Values and norms of proof for mathematicians and students. *Educational Studies in Mathematics*, 95, 123–142.

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, 24, 17–24.

De Vries, H., Elliott, M. N., Kanouse, D. E., and Teleki, S. S. (2008). Using pooled kappa to summarize interrater agreement across many items. *Field Methods*, 20(3), 272–282.

Dewey, J. (1948). *Reconstruction of Philosophy* Boston, MA: Beacon Press.

Doyle, L., Brady, A. M., and Byrne, G. (2016). An overview of mixed methods research – revisited. *Journal of Research in Nursing*, 21(8), 623–635.

Epstein, J. (2013). The Calculus Concept Inventory - measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society*, 60(8), 1018–1027.

Fawcett, H. (1938). *The Nature of Proof.* New York, NY: Bureau of Publications Teachers College, Columbia University.

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R.* Thousand Oaks, CA: Sage Publishing.

Gage, N. (1989). The paradigm wars and their aftermath. *Educational Researcher*, 18(7), 4–10.

Geist, C., Benedikt, L., and Kerkhove, B. V. (2010). Peer review and knowledge by testimony in mathematics. *Philosophy of Mathematics: Sociological Aspects and Mathematical Practice*, (pp. 155–178). London, UK: College Publications.

Hanna, G. and Janke, N. (1996). Proof and proving. In Bishop, A., editor, *International handbook of mathematics education*, (pp. 887–908). Dordrecht, NL: Kluwer Academic Publishers.

Harel, G. & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In Dubinsky, E., Schoenfeld, A., and Kaput, J. (Eds.), *Research in Collegiate Mathematics Education. III*, (pp. 234–283). Washington, D.C.: American Mathematical Society.

Harel, G. and Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In Lester, F. (Eds.), *Second Handbook of Research on Mathematics Teaching and Learning* (2nd ed.). (pp. 805–842). Greenwich, CT: Information Age Publishing.

Hathcoat, J. D. and Meixner, C. (2017). Pragmatism, factor analysis, and the conditional incompatibility thesis in mixed methods research. *Journal of Mixed Methods Research*, 11(4), 433–449.

Healy, L. and Hoyles, C. (1998). *Justifying and proving in school mathematics. Summary of the results from a survey of the proof conceptions of students in the UK.* Research report. London, UK: Institute of Education, University of London.

Healy, L. and Hoyles, C. (2000). A study of proof concepts in algebra. *Journal for Research in Mathematics Education*, 31(4), 396–428.

Heldsinger, S. and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Australian Educational Researcher*, 37(2), 1–19.

Heldsinger, S. A. and Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research*, 55(3), 219–235.

Hevey, D. (2010). Think-aloud methods. In Salkind, N. J. (Ed.), *Encyclopedia of Research Design*, (pp. 1505 – 1506). Thousand Oaks, CA: Sage Publishing.

Hodds, M., Alcock, L., and Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, 45(1), 62–101.

Holmes, S. Black, B., and Morin, C. (2018). *Marking reliability studies 2017: Rank ordering versus marking – which is more reliable?* Research Report. Coventry, UK: Ofqual.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, 17(8), 10 – 16.

Hoyles, C. and Healy, L. (2007). Curriculum change and geometrical reasoning. In Boero, P. (Ed.), *Theorems in School*, (pp. 81–115). Rotterdam, Netherlands: Sense Publishers.

Hunter, J. and Jones, I. (2018). Free-response tasks in primary mathematics: A window on students' thinking. In Hunter, J., Perger, P., and Darragh, L. (Eds.) In *Making waves, opening spaces: Proceedings of the 41st Annual Conference of the Mathematics Education Research Group of Australasia*, (pp. 400–407), Auckland, New Zealand.

Iannone, P. and Simpson, A. (2011). The summative assessment diet: How we assess in mathematics degrees. *Teaching Mathematics and its Applications*, 39(4), 186–196.

Iannone, P. and Simpson, A. (2013). Students' view of value and validity in undergraduate mathematics assessment. *Research in Mathematics Education*, 15(1), 17–33.

Iannone, P. and Simpson, A. (2015). Students' preferences in undergraduate mathematics assessment. *Studies in Higher Education*, 40(6), 1046–1067.

Inglis, M. and Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43(4), 358–390.

Inglis, M., Mejia-Ramos, J. P., Weber, K., and Alcock, L. (2013). On mathematicians' different standards when evaluating elementary proofs. *Topics in Cognitive Science*, 5(2), 270–282.

Jaffe, A. and Quinn, F. (1993). Theoretical mathematics: Toward a cultural synthesis of mathematics and theoretical physics. *Bulletin of the American Mathematical Society*, 29(1), 1–13.

James, W. (1907). *Pragmatism.* New York, NY: Dover.

Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26.

Johnson, R. B. and Onwuegbuzie, A. J. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112–133.

Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10), 1774–1787.

Jones, I., Bisson, M. J., Gilmore, C., and Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45(3), 662–680.

Jones, I. and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics*, 89(3), 337–355.

Jones, I., Wheadon, C., Humphries, S., and Inglis, M. (2016). Fifty years of A-level mathematics: have standards changed? *British Educational Research Journal*, 42(4), 543–560.

Jones, I. and Karadeniz, I. (2016). An alternative approach to assessing achievement. In Csikos, C., Rausch, A., and Szitanya, J. (Eds.) *Proceedings of the 2016 40th Conference of the International Group for the Psychology of Mathematics Education*, (pp. 113 – 120), Szeged, Hungary.

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1), 151–177.

Jones, I. and Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47, 93–101.

Jones, I. and Sirl, D. (2017). Peer assessment of mathematical understanding using comparative judgement. *Nordic Studies in Mathematics Education*, 22(4), 101-119.

Kane, M. T. (2001). Current concept in validity theory. *Journal of Educational Measurement.*, 38(4), 319–342.

Kanellos, I., Nardi, E., and Biza, I. (2018). Proof schemes combined: mapping secondary students' multi-faceted and evolving first encounters with mathematical proof. *Mathematical Thinking and Learning*, 20(4), 277–294.

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, 22(2), 135–155.

Kleiner, I. and Movshovitz-Hadar, N. (1997). Proof: A many splendored thing. *The Mathematical Intelligencer*, 19(3), 16–26.

Kline, M. (1980). *Mathematics: The Loss of Certainty.* New York, NY: Oxford University Press.

Ko, Y. and Knuth, E. (2009). Undergraduate mathematics majors' writing performance producing proofs and counterexamples about continuous functions. *Journal of Mathematical Behavior*, 28(1), 68–77.

Ko, Y. and Knuth, E. (2013). Validating proofs and counterexamples across content domains: Practices of importance for mathematics majors. *Journal of Mathematical Behavior*, 32(1), 20–35.

Küchemann, D. and Hoyles, C. (2006). Influences on students' mathematical reasoning and patterns in its development: Insights from a longitudinal study with particular reference to geometry. *International Journal of Science and Mathematics Education*, 4(4), 581–608.

Lai, Y., Weber, K., and Mejía-Ramos, J. P. (2012). Mathematicians' perspectives on features of a good pedagogical proof. *Cognition and Instruction*, 30(2), 146–169.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: Oxford University Press.

Lane-Getaz, S. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal*, 13(1), 20–47.

Larvor, B. (2012). How to think about informal proofs. *Synthese*, 187(2), 715–730.

Leech, N. L. and Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality and Quantity*, 43(2), 265–275.

Luce, R. (1959). *Individual Choice Behavior.* New York, NY: Wiley.

Mancosu, P. (2008). *The Philosophy of Mathematical Practice.* Oxford, UK: Oxford University Press.

Marcus, R. and McEvoy, M. (2016). *An Historical Introduction to the Philosophy of Mathematics: A Reader.* London, UK: Bloomsbury Publishing.

Mariotti, M. (1997). Justifying and proving in geometry: the mediation of a microworld. In Hejny, M. and Novotna, J. (Eds.)*Proceedings of the European Conference on Mathematical Education*, (pp. 21 – 26), Prague, Czech Republic.

McMahon, S. and Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 368–389.

Mejia-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., and Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, 79(1), 3–18.

Mejia-Ramos, J. P. and Inglis, M. (2011). Semantic contamination and mathematical proof: Can a non-proof prove? *Journal of Mathematical Behavior*, 30(1), 19–29.

Mejia-Ramos, J. P., Lew, K., de la Torre, J., and Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, 19(2), 130–146.

Mejia-Ramos, J. P. and Weber, K. (2016). Student performance on proof comprehension tests in transition-to-proof courses. Paper presented at *19th annual Special Interest Group of the Mathematical Association of American on Research in Undergraduate Mathematics Education.* Pittsburgh, PA. Retrieved from http://sigmaa.maa.org/rume/crume2016/Papers/RUME_19_paper_123.pdf

Menand, L. (1997). *Pragmatism: A reader* New York, NY: Vintage.

Messick, S. (1989). Validity. In Linn, R. (Ed), *Educational measurement*, (pp. 13-104). New York, NY: Macmillan Publishing.

Miller, D., Infante, N., and Weber, K. (2018). How mathematicians assign points to student proofs. *Journal of Mathematical Behavior*, 49, 24–34.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, 27(3), 249–266.

Muis, K. (2004). Personal epistemology and mathematics: A critical review and synthesis of research. *Review of Educational Research*, 74(3), 317–377.

Newhouse, C. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, 21(2), 205–220.

Onwuegbuzie, A. and Johnson, R. (2006). The validity issue in mixed research. *Research in the Schools*, 13(1), 48–63.

Peirce, C. (1878). How to make our ideas clear *Popular Science Monthly*, 12, 286-302.

Persaud, N. (2010a). Interviewing. In Salkind, N. J. (Ed.), *Encyclopedia of Research Design*, (pp. 633 – 636). Thousand Oaks, CA: Sage Publishing.

Persaud, N. (2010b). Protocol. In Salkind, N. J. (Ed.), *Encyclopedia of Research Design*, (pp. 1133 – 1135). Thousand Oaks, CA: Sage Publishing.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157–170.

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment In Education: Principles, Policy & Practice*, 19(3), 281–300.

Pollitt, A. and Murray, N. L. (1993). What raters really pay attention to. In Milanovic, M. and Saville, N. (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium*, (pp. 74-91). Cambridge, UK: Cambridge University Press.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: University of Chicago Press.

Rav, Y. (2007). A critique of a formalist-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica*, 15(3), 291–320.

Recio, A. and Godino, J. (2001). Institutional and personal meanings of proof. *Educational Studies in Mathematics*, 48(1), 83–99.

Reid, D. and Knipping, C. (2010). *Proof in Mathematics.* Wolfville, Canada: Sense Publishers.

Resch, A. and Isenberg, E. (2018). How do test scores at the ceiling affect value-added estimates? *Statistics and Public Policy*, 5(1), 1–6.

Rowland, T. (2001). Generic proofs in number theory. In Campbell, S. and Zazkis, R. (Eds.) *Learning and Teaching Number Theory: Research in Cognition and Instruction*, (pp. 157–184). Westport, CT: Ablex Publishing.

Pinto, R. M. (2010). Mixed Methods Design. In Salkind, N. J. (Ed.), *Encyclopedia of Research Design*, (pp. 813 – 819). Thousand Oaks, CA: Sage.

Seery, N., Canty, D., and Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, 22(2), 205–226.

Segal (1999). Learning about mathematical proof: Conviction and validity. *Journal of Mathematical Behavior*, 18(2), 191–210.

Selden, A. and Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, 34(1), 4–36.

Shapiro, S. (1997). *Philosophy of Mathematics: Structure and Ontology.* New York, NY: University Press.

Shapiro, S. (2000). *Thinking about mathematics: The philosophy of mathematics.* New York, NY: Oxford University Press.

Shaw, S. and Crisp, V. (2011). Tracing the evolution of validity in educational measurement: past issues and contemporary challenges. *Research Matters: A Cambridge Assessment Publication*, 11, 14–19.

Shepherd, M. D., Selden, A., and Selden, J. (2012). University students' reading of their first-year mathematics textbooks. *Mathematical Thinking and Learning*, 14(3), 226–256.

Spearman, C. (1904). General intelligence: objectively determined and measured. *American Journal of Psychology*, 14, 107-197.

Stedall, J. (2012). *The History of Mathematics: A Very Short Introduction.* New York, NY: Oxford University Press.

Stylianides, G. J., Stylianides, A. J., and Weber, K. (2017). Research on the teaching and learning of proof: Taking stock and moving forward. In Cai, J. (Ed.), *Compendium for Research in Mathematics Education*, (pp. 237–266). Reston, VA: National Council of Teachers of Mathematics.

Stylianou, D. A., Blanton, M. L., and Rotou, O. (2015). Undergraduate students' understanding of proof: Relationships between proof conceptions, beliefs, and classroom experiences with learning proof. *International Journal of Research in Undergraduate Mathematics Education*, 1(1), 91–134.

Tall, D. and Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, 12(2), 151–169.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286.

Verhavert, S. (2018). *Beyond a Mere Rank Order: the Method, the Reliability and the Efficiency of Comparative Judgment.* (Doctoral dissertation, University of Antwerp, Belgium). Retrieved from https://repository.uantwerpen.be.

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, 42(6), 428–445.

Weber, K. (2001). Student difficulties in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, 48(1), 101–119.

Weber, K. (2004). Traditional instruction in advanced mathematics courses: a case study of one professor's lecture and proofs in an introductory real analysis course. *Journal of Mathematical Behavior*, 23(2), 115–133.

Weber, K. (2010). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning*, 12(4), 306–336.

Weber, K. (2012). Mathematicians' perspectives on their pedagogical practice with respect to proof. *International Journal of Mathematical Education in Science and Technology*, 43(4), 463–482.

Weber, K. (2014). What is proof? A linguistic answer to an educational question. Paper presented at *17th annual Special Interest Group of the Mathematical Association of American on Research in Undergraduate Mathematics Education.* Denver, CO. Retrieved from http://sigmaa.maa.org/rume/crume2014/Schedule/Papers.htm.

Weber, K. (2015). Effective proof reading strategies for comprehending mathematical proofs. *International Journal of Research in Undergraduate Mathematics Education*, 1(3), 289–314.

Weber, K. and Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics*, 56(2), 209–234.

Weber, K. and Czocher, J. (2019). On mathematicians' disagreements on what constitutes a proof. *Research in Mathematics Education.* Retrieved from https://doi.org/10.1080/14794802.2019.1585936.

Weber, K., Inglis, M., and Mejia-Ramos, J. P. (2014a). How mathematicians obtain conviction: Implications for mathematics instruction and research on epistemic cognition. *Educational Psychologist*, 49(1), 36–58.

Weber, K., Mejia-R, and Amos, J. P. (2014b). Mathematics majors' beliefs about proof reading. *International Journal of Mathematical Education in Science and Technology*, 45(1), 89–103.

Williams, P. (2012). Investigating the feasibility of using digital representations of work for performance assessment in engineering. *International Journal of Technology and Design Education*, 22(2), 187–203.

Willig, C. (2013). *Introducing Qualitative Research in Psychology.* London, UK: Open University Press.

Wittgenstein, L. (1953). *Philosophical Investigations* (translated edition). New York, NY: Macmillan Publishing.

Yang, K. L. and Lin, F. L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, 67(1), 59–76.